United Nations
Educational, Scientific and
Cultural Organization

4

Interoperability and **Retrieval**

Open Access for Library Schools

United Nations
Educational, Scientific and
Cultural Organization

Interoperability and Retrieval

Module

4

Interoperability and Retrieval

Published in 2015 by the United Nations Educational, Scientific and Cultural Organization, 7, place de Fontenoy, 75352 Paris 07 SP, France

ISBN 978-92-3-100077-5

Cover design by The Commonwealth Educational Media Centre for Asia (CEMCA)
Printed in PDF

CURRICULUM DESIGN COMMITTEE

Anirban Sarma
UNESCO New Delhi, India

Anup Kumar Das
Jawaharlal Nehru University, India

Barnali Roy Choudhury
CEMCA, New Delhi

Bhanu Neupane
UNESCO, Paris, France

Bojan Macan
Ruder Boškovif Institute Library, Croatia

Dominique Babini
CLACSO, Argentina

Ina Smith
Stellenbosch University, South Africa

Iskra Panevska
UNESCO New Delhi, India

Jayalakshmi Chittoor Parameswaran
Independent Consultant, India

M Madhan
ICRISAT, India

Parthasarathi Mukhopadhyay
Kalyani University, India

Ramesh C Gaur
Jawaharlal Nehru University, India

Sanjaya Mishra
CEMCA, New Delhi, India

Shalini Urs
University of Mysore, India

Sridhar Gutam
Central Institute for Subtropical Horticulture, India

Susan Veldsman
Academy of Science of South Africa, South Africa

Uma Kanjilal
Indira Gandhi National Open University, India

Upali Amarasiri
University of Colombo, Sri Lanka

Žibutė Petrauskiene
Vilnius University Library, Lithuania

MODULE ADVISORS

Ramesh C Gaur
Jawaharlal Nehru University, India

Uma Kanjilal
Indira Gandhi National Open University, India

Project Coordinator

Sanjaya Mishra
CEMCA, New Delhi, India

MODULE PREPARATION TEAM

Writer
Parthasarathi Mukhopadhyay
Kalyani University, India

Editor
Prof. S.B. Ghosh
Formerly at Indira Gandhi National Open University, India

Chief Editor
Sanjaya Mishra
CEMCA, New Delhi

MODULE INTRODUCTION

Retrieving information is the prime concern of any information storage and retrieval system. .All the activities of information storage and retrieval systems and the components within it are organized and developed keeping in view the envisaged retrieval features of the system -- whether it is traditional resources or web enabled information resources. In the context of web-enabled information retrieval, interoperability between one system to other offers great advantages to the users for obvious reasons. The interoperability demands adherence to standards and compatibility in the resource organization and retrieval features of the information resource systems. In the context of web-enabled information system, content development is the prime activity which encompasses resource description, incorporation of retrieval features etc. for retrieving information and development of various services including push services.

This module focuses on interoperability issues, resource description and also the information retrieval in the context of open access resources. The objective is to help you understand interoperability issues, perpetual access, importance of standards, and the integration of different products in building institutional repositories and also various retrieval features that is available which can be considered for development of IR system for open access resources.

The Unit 1 of this Module deals with Resources Description for OA Resources to make you understand the basics of metadata, the elements of some important metadata formats and the need and importance of using it in the context of open access resources.

By this time, after going through other modules, you may be in a position to appreciate the importance of interoperability in general and its necessity in the context of open access resources in particular. Interoperability is required to facilitate information retrieval by the users. Various issues are involved in achieving interoperability amongst systems, different standards have been developed and various initiatives have been taken to achieve interoperability, The Unit 2 of this module on interoperability issues for Open Access provides you an insight into different issues involved in it, describes the different standards/initiatives available for interoperability and also gives you an overview of emerging trends in the field.

Retrieval of information has been a point of research and development over the ages. Many theoreticians and practitioners have developed various theories, systems and techniques to find a suitable solution to the problem of handling unstructured information that represents concepts /ideas of the authors. Though many standards have been developed in the different areas of information processing, but no uniform single standard has yet been possible which can be followed globally for developing a suitable information retrieval system, encompassing all types of information that can be followed by all. The development of web-enabled resources has added another dimension to the problem. This is a general scenario in the context of information storage and

retrieval. Retrieval of information in the context of open access resources is not an exception.

Whatever be the types of resources (form and format), the basic theories and systems remain the same. The development in ICT has provided a new opportunity to develop new methods and techniques. The Unit 3 on Retrieval of Information for OA Resources has been developed with this perspective to provide you with an insight to understand the importance of efficient retrieval of information, the fundamentals of information retrieval and also, identify the issues related to text, multimedia and multilingual retrieval systems. It is neither possible nor necessary to discuss in this space the entire theories and processes of information storage and retrieval systems, which you may already be knowing. Only those concepts related to retrieval, which are necessary to understand the topic of this unit, have been discussed. Based on these foundations, 'how' of information retrieval for OA resources have been discussed in detail. To this end, different retrieval systems and the features of different search engines have been compared. The ontological approach to retrieval of information which is a very important development in the context of web indexing has also been discussed.

At the end of this module, you are expected to be able to understand interoperability issues, perpetual access, importance of standards, and the integration of different products in building institutional repositories.

UNIT 1 RESOURCE DESCRIPTION FOR OA RESOURCES

Structure

1.0 INTRODUCTION

Metadata is a very important Component for OA resources not only for organizing and retrieval but also to inform stakeholders of OA infrastructure about the status of a resource as OA. For example - i) users need to understand what rights they have for a given knowledge object (e.g., free readership for the published version, limited reuse, etc; ii) authors want to know what rights they will retain (after publication in OA system) and whether they are compliant with a given funder policy; iii) publishers want to clearly convey what readers can and cannot do with the objects they publish; iv) research funders want to promote research output they sponsor; v) search engines, A&I databases, and other discovery services are aiming to help users in finding OA resources; and vi) libraries are seeking to help users in finding OA resources and their integration with existing library materials. These expectations of stakeholders are depending on quality description of OA resources by applying granular, comprehensive and domain-specific metadata schemas. This unit is meant for helping you in application of standard metadata schemas in organizing OA resources.

1.1 LEARNING OUTCOMES

After working through this unit, you are expected to be able to:

- Define metadata;
- Identify and describe the elements of some important metadata description formats;
- Understand policies related to metadata applications;
- Critically examine the scopes of generic and domain-specific metadata schemas for organizing OA resources;
- Explain the roles of models, crosswalks and interoperability standards in metadata applications including the scope of emerging initiatives in OA metadata landscape; and
- Explore the software-level application of metadata in organizing OA resources.

1.2 RESOURCE DESCRIPTION

Metadata, in general, is referred to as data about data, and provides basic information such as the author of a work, the date of creation, links to any related works, etc. Metadata exists for almost every conceivable object or group of objects, whether stored in electronic form or not. In the library world, one easily identifiable form of metadata is the card catalogue; the information on the card is metadata about a book. In a traditional library, where cataloguing is the work of trained professionals, complex metadata schemes such as MARC, CCF etc. are used for description of library resources.

As a library professional you know the application of metadata in the form of cataloguing. There are strong similarities between traditional library cataloguing and the description of web resources by using a set of metadata. Modern cataloguing theory and practice developed over the last 150 years or so as a tool for organizing information for retrieval in the libraries. Library catalogue typically consist of a collection of bibliographic records that describe library resources such as printed books, cartographic materials, music scores, manuscripts, etc that aim to describe the different types of resources of a library. Gradually the scope of cataloguing codes and resource description standards have expanded to include a range of newer publishing media such as sound recordings, microfilms, video recordings, films, computer files and Web resources. For such descriptions different standards and standard procedures have been developed from time to time to facilitate recording and access of the resources. Open access materials are also no exception. For example, when users retrieve journal metadata from DOAJ (Directory of Open Access Journal), one of the important elements of description is APC (i.e. Article Processing Charge). This metadata element helps contributors in selecting appropriate journal(s) for publication of research results. Another related metadata is the date from which content is available as Open Access. This

metadata elements help users in selecting appropriate resources from journals which started in close mode and subsequently available in open mode.

With the rise of Internet and the Web as global publishing media, the term metadata began to appear in the context of describing information objects on the network. Library professionals were quick to realize that they had been creating data about data, in the form of cataloguing over the last one hundred fifty years, since the time of Panizzi. However, there is inconsistent use of the term 'metadata' even within the library community. Some are using it to refer to the description of both digital and non-digital resources, and others restricting the term to the description of electronic resources. For example, definitions given by IFLA (International Federation of Library Associations and Institutions) and W3C (World Wide Consortium) are restrictive in nature. IFLA defines metadata as "The term refers to any data used to aid the identification, description and location of networked electronic resources" (IFLA, 2002). According to W3C "Metadata is the machine understandable information for the Web" (W3C, 2003). In contrast, definitions given by Getty Research Institute (GRI) and UKOLN (U.K. Office for Library and Information Networking) are fairly liberal. GRI says metadata is "Data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation" (Murtha, 2002). Similarly UKOLN says, "Metadata is normally understood to mean structured data about digital (and non-digital) resources that can be used to help and support a wide range of operations. These might include, for example, resource description and discovery, the management of information resources (including rights management) and their long-term preservation" (UKOLN, 2002). For the purpose of this unit, a liberal stand in terms of the definition and scope of the term metadata is taken. Metadata is used here to mean structured information about an information resource of any media type or format. Metadata by definition is descriptive of something, but many different use of metadata has led to the construction of a very broad typology of metadata as being descriptive, administrative and structural (Hadge, 2001):

- **Descriptive metadata** is meant to serve the purposes of discovery (i.e. how one can find a resource), identification (i.e. how a resource can be distinguished from other similar resources), selection (i.e. how to determine that a resource fills a particular need), collocation (bringing together related works), obtain (obtaining a copy of resource, or access to one) and other related functions (evaluation, linkage and usability).
- **Administrative metadata** is information intended to facilitate the management of resources such as date of creation, rights and restrictions of access and archiving, control or processing activities etc.
- **Structural metadata** is concerned with recording of relationships that holds compound digital objects together.

Metadata schemas are set of metadata elements and rules for their use that have been defined for a particular purpose. A metadata schema specifies three

independent but related aspects of metadata – semantics, content rules and syntax:

- **Semantics** refers to the metadata elements that are included in the schema by giving each of them a name and definition. A metadata schema also specifies whether each element is mandatory, optional or conditionally required and whether the element may or may not be repeated.
- **Content rules** indicate how values for metadata elements are selected and represented. For example, semantics of a metadata schema may define the element "author" but the content rules would specify which agents qualify as author (selection) and how an author's name should be recorded (representation).
- **Syntax** of a metadata schema is concerned with the encoding of metadata elements in machine-readable form. Syntax also specifies the way of transmission, transport and communication of metadata between different systems.

Based on their applications, metadata schemas can be grouped into two types – generic and domain-specific. Generic metadata schemas are intended to be generally applicable to all types of resources (e.g., Dublin Core Metadata Elements Set), whereas, domain-specific metadata schemas are primarily designed to describe items related to a particular category (e.g. VRA [Visual Resource Association] Core for visual resource collection, FGDC (Federal Geographic Data Committee) metadata schema for geospatial data etc.). All of these metadata schemas contain descriptive metadata elements, administrative metadata elements, structural metadata elements (Semantics), content rules for metadata representation and syntax for machine-readable metadata encoding. The nature of contents for different categories of metadata elements in schemas are briefly discussed below:

Descriptive metadata elements

- Bibliographic description (such as Dublin Core, MODS, MARC21, MARCXML, ONIX schemas for metadata representation);
- Content description (such as DDI, SDMX, FGDC, EAD, TEI etc.);
- Description of structure, context and source of the data; information about the methods, instruments, and techniques used in the creation or collection of the data;
- References and links to publications pertaining to the data; and
- Information on how the data have been processed prior to submission to the repository.

Administrative metadata elements

- Preservation metadata to represent lifecycle of the data, recording of events related to submission, curation and dissemination (such as PREMIS) and event history data (for linking with digital objects) ;

- Rights management metadata;
- Technical metadata (storage format etc.); and
- Representation Information (internal coding, rendering data etc).

Structural metadata elements

Structural metadata indicates relationships amongst different components of a set of associated data that are particularly important for Web aggregation. These aggregations are also called compound digital objects. These digital objects combine distributed resources with multiple media types including text, images, data and video. There are standards for the description and exchange of aggregations of Web resources such as

- FOXML (Standard in use for Fedora repository software, where compound objects are treated as a single file);
- OAI-ORE (An OAI initiative that defines compound objects distributed on the Internet through the creation of resource maps which use unique URLs for each component; It has four basic components i) Resource (an item of interest); ii) URI (a global resource identifier); iii) Representation (a DataStream accessible through URI by using a protocol like HTTP); and iv) Link (a connection between two resources);
- METS (An LoC standard that is used as a 'wrapper' for compound digital objects and very useful for import/export in repositories); and
- RDF (A W3C standard that provides a simple way to represent Web resources, in the form of subject-predicate-object expressions that relate objects to one another).

Why Metadata is important in Open Access?

The core function of a library is to deliver the right contents to users at the right time. In the context of Open Access (OA), metadata plays a crucial role to fulfill this core function. A logical question is possibly coming to your mind that why metadata is so important for disseminating OA resources. The answer is simple one. Apart for supporting all the elements necessary for discovering resources effectively, metadata in OA has additional role to inform *the status of a piece of content as open access*. If the status of a scholarly object as open access is not obvious it may lead to confusion for end users in assessing access rights and extent of permissions related to a knowledge object. Metadata in the context of OA is important for both library professionals and end users. It helps librarians in data mining, pattern identification (organization and usage), and clarity over licensing agreements, discovering of OA, and accessing open access contents within hybrid journals. On the other hand, metadata helps end user in finding and accessing OA contents, in setting priority of OA contents over paid contents (filtering of results by OA status), in knowing access and re-use permissions, and in getting help to cite OA resources.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

1) "Metadata schema deals with semantics, content rules and syntax". Elucidate.

……...

……...

……...

2) Why do you think metadata is important for dissemination of OA contents?

……...

……...

……...

1.3 OPEN ACCESS AND METADATA

The organization and dissemination of OA materials is presently passing through a complex phase. The major stakeholders of OA infrastructure like publishers, researchers, institutes, funders and end users have different concepts and expectations from OA systems and services. For example, governments (as funding agencies) want to ensure wide availability of research publications in public domain. Many governments are developing policies in this direction. (Please refer to Module 3, Unit 1 for further details). End users want to know what research is accessible to them, and to what extent they can reuse accessible contents. Another problematic zone is 'hybrid journals' in which some of the article are available freely (authors pay to make their paper freely available to readers), while the rest of the journal contents available against subscription fees. This varied environment limits – i) effective resource discovery; ii) clarity in reuse rights; and iii) possibility of adopting standards to bridge requirements of stakeholders. Till date no standardized bibliographic metadata schemas have metadata elements to specify whether a given article is openly accessible and what reuse rights are associated with it.

1.3.1 Policy Framework

An OA service (whether Gold or Green) needs to develop a policy framework for metadata in view of the importance of metadata in OA, discussed in previous sections. The policy framework for metadata needs to address issues like – i) Who can enter or edit metadata? ii) Which metadata standards are to

be followed? iii) Whether different metadata schemas are required for describing different type of documents? iv) Whether or not the repository systems allow metadata harvesting by service providers? v) Which protocols should OA system support for metadata harvesting? As per OpenDOAR (OpenDOAR, 2013) database, more than 84% repositories have not defined metadata policy (Figure 1). Analysis of ROARMAP also shows that most of the OA repositories (OAR) have no metadata policy but almost all the OARs clearly state that anyone may access the metadata.

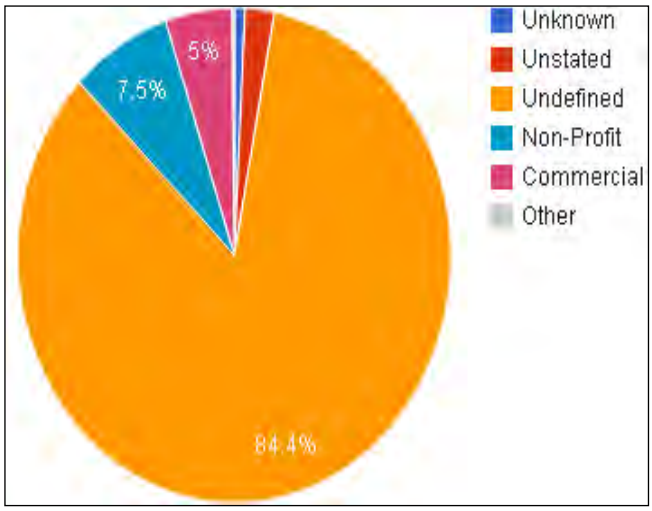

Figure 1: Metadata Policy: Aanalysis of OpenDOAR (*Source*: opendoar.org**)**

An efficient OA service must work on the basis of a standard metadata policy. Let us discuss metadata policy requirements for organizing OA resources one by one. The policy issues related to metadata are discussed on the basis of recommendations of OA experts and subsequent analysis of ROARMAP database.

Policy Issue I: Who can create or edit metadata?

OA experts' view: Many OA experts suggest (Graaf & Eijndhoven, 2008; Barton & Walker, 2002) that contributors of open contents may enter simple descriptive metadata like creator, title and keywords. In case of difficulties they may take help of intermediaries like library professionals. Some researchers and OA service providers (DINI, 2003; Pinfield, Gardner & MacColl, 2002) advocated that standardized metadata should be created and provided for exchange and harvesting services.

ROARMAP analysis: Only a few OARs (see Table 1 for an illustrative list) have suggested that metadata should be created and provided by author or eligible contributors. Library staff, if necessary, may edit or create additional metadata.

Interoperability and Retrieval

Policy Issue II: What metadata standards to be used?

OA experts' view: OAR systems differ widely in the selecting and applying metadata schema to support the ingest, management, and use of data in their collections. Most of the researchers recommended to use qualified Dublin Core as metadata standard for organizing OA resources (Graaf & Eijndhoven, 2008; Gibbons, 2004;) in general but some of the researchers are in opinion that domain-specific metadata should be employed by the OA service providers for organization of specialized contents like ETDs and learning objects.

ROARMAP analysis: It is also clear from the study that almost all the OARs use Dublin Core standards. A few repositories implemented additional or extended metadata schemas for domain specific datasets (see Table 1 for an illustrative list).

Policy Issue III: How to standardize subject access metadata elements?

OA experts' view: Expert and OA service providers (DINI, 2003; Nolan & Costanza, 2006) recommend that standard vocabularies should be adopted for populating subject access fields of metadata schema in use.

ROARMAP analysis: The analysis of the dataset shows that only a few OA service providers are using controlled vocabulary for populating subject access metadata element i.e DC.Subject metadata element for standardizing subject indexing. The other metadata elements also required use of authority list like language code (for DC.Language) etc.

Policy Issue IV: Whether metadata sets be open for harvesting?

OA experts' view: Most of the OA researchers are in favor of metadata harvesting to support developing federated search interface (Hirwade & Hirwade, 2006; Singh, Pandita & Dash, 2008; Sarkar & Mukhopadhyay, 2010). OA experts also opined that Gold and Green OA systems must be compliant with OAI/PMH standard to support metadata harvesting.

ROARMAP analysis: A detail report of the present statistics related to OAI/PMH Compliant repositories is given in Table 2.

Policy Issue V: If open for harvesting, what should be the metadata re-use policy?

OA systems need to follow a policy framework for metadata reuse to resolve issues like – i) whether harvesting requires prior permission? ii) whether link/acknowledgement is mandatory? iii) whether harvesting is open for all or restricted to non-commercial use only?

Analysis of ROARMAP shows that only a few OARs have metadata reuse policy (see table 2 for an illustrative list). Most of the OARs allow metadata harvesting in any medium without prior permission for not-for-profit purposes. In some OARs restriction is that metadata must not be re-used in any medium for commercial purposes without formal permission.

Table 1: Metadata policies in OARs I (Source: ROARMAP)

Name of the Repository	Policy related to Metadata		
	Metadata Schema Used	Bibliographic Metadata	
		Provided by	Created or Edited by
Anglia Ruskin Research Online	Simple Dublin Core		Library staff
Brandeis Institutional Repository		Eligible contributor	
Brigham Young University Library	Simple Dublin Core		
Centre for Environmental Data Archival Repository		√	
Cornell University (eCommons)			Library staff
Edith Cowan University	Unqualified Dublin Core		
Goddard Library Repository	GEMS (own)		
Griffith University		√	
Harvard University Library,		authorized submitter	
Katholieke Universiteit Leuven		√	
Kwame Nkrumah University of Science and Technology Institutional Repository (KNUSTSpace)	Qualified Dublin core		√
Loughborough University		√	
Massachusetts Institute of Technology (MIT)		Eligible contributor/ depositors	
Northeastern University Libraries Institutional Repository	METS schema & Qualified & unqualified Dublin Core (descriptive metadata)		
St John University		√	
Teesside University's Institutional Repository (TeesRep)	Dublin Core		
Trento University		√	
University of Abertay Dundee	Dublin Core	Authors/ or delegated agents	√
University of Calgary: Library and Cultural Resources		√	
University of Cambridge	Qualified Dublin Core		
University of East Anglia		√	
University of Kansas	Dublin Core Library Application Profile (DC-Lib)		
University of Melbourne Eprint Repository	Simple Dunlin Core		
University of Queensland		√	
University of Reading		√	
University of Rochester's	Dublin Core & locally defined DTDs		
University of Salford	Dublin Core		√
University of South Australia	MARCXML & DC		
University of Starling (STORRE)	Dublin Core		
University of Sydney	Qualified Dublin Core		
University of Utah's institutional repository	Dublin Core		
University of Westminister		√	
York St John University		√	

Table 2: Metadata Reuse Policy II (Source: ROARMAP)

Name of the Repository	Metadata may be re-used in any medium without prior permission	Metadata must not be re-used in any medium for
Arts and Humanities Research Council	for not-for-profit purposes	commercial purposes without formal permission
Aston University Research Archive	√	√
Canadian Cancer Society	√	√
Canadian Health Services Research Foundation	√	√
Canadian Institutes of Health Research	√	√
Centre for Environmental Data Archival Repository	√	√
Covenant University	√	√
Curtin University	√	√
Edith Cowan University	√	√
European Heads of Research Councils	√	√
European Research Advisory Board	√	√
European Research Council	√	√
European University Association	√	√
Fonds de la recherche en sante Québec	√	√
Fonds zur Foerderung der wissenschaftlichen Forschung	√	√
Goddard Library Repository	√ (Unrestricted metadata)	
Genome Canada	for not-for-profit purposes	√
Heart and Stroke Foundation of Canada	√	√
JISC (Joint Information Systems Committee)	√	√
Katholieke Universiteit Leuven	√	√
Khazar University	√	√
Kwame Nkrumah University of Science and Technology Institutional Repository (KNUSTSpace)	√	√
Leeds Metropolitan University	√	√
Loughborough University	√	√
Michael Smith Foundation for Health Research	√	√
Murdoch University	√	√
National Research Council	commercial purposes without formal permission	√
Natural Environmental Research Council	for not-for-profit purposes	√
Natural Sciences and Engineering Research Council of Canada	√	√

Northern Melbourne Institute of TAFE	√	√
Ontario Institute for Cancer Research	√	√
Queensland University of Technology	√	
St John University	√	√
Stanford University: School of Education	√	√
University of Strathclyde Institutional Repository (Strathprints)	√	√
Teesside University's Institutional Repository (TeesRep)	√	√
Trento University	√	√
Universidad Nacional de Colombia	√	
University of Bath	√	√
University of East Anglia	√	√
University of Calgary: Library and Cultural Resources	√	√
University of Edinburgh	√	√
University of Leicester	√	√
University of Lincoln	√	√
University of of Melbourne Digital Repository	√	√
University of Nottingham	√	√
University of Reading	√	√
University of Salford	√	
University of Surrey	√	√
University of Southampton Research Repository (ePrints Soton)	√	√
University of Virginia	√	√
University of Wollongong	√	
Warwick Research Archive Portal	√	√
York St John University	√	√

1.3.2 Application Framework

On the basis of metadata policies discussed in previous section, a set of recommendations may be drawn to help application of metadata standards for organizing OA resources. The list of major decisions related to OA metadata is given below:

1) Anyone may access the metadata free of charge;
2) All metadata in the repository should be based on the recognized global standard;
3) Qualified version of the Dublin Core schema as a descriptive metadata standard will be used;
4) Community/domain-specific metadata elements will be used where no suitable element or element refinement exists in generic schema like DCMES;
5) Recommends DCMES as generic metadata schema and suggests respective domain-specific schemas for special objects like ETD (UK-ETD), Learning Objects (IEEE-LOM), Journal articles (Qualified DCMES) etc. on the basis of a set of standard parameters;

6) Deposit of materials to OA system requires a minimum set of descriptive information (metadata) to be provided at the point of deposit;
7) Basic metadata will be created by authors or their delegated agents at the time of submission;
8) Library professionals will create additional metadata elements and edit basic metadata set, if required, to ensure the quality of complete metadata records;
9) Recommends following basic cataloging standards –AACR/RDA – for rendering personal and corporate names;
10) OA systems may allow metadata harvesting and supports metadata extraction through OAI-PMH standards;
11) Metadata elements must support basic retrieval tasks including advanced set of search operators;
12) Controlled vocabularies will be used to maintain consistency and to enhance the quality of records exposed to search and browse services;
13) The metadata of withdrawn items shall not be searchable;
14) Appropriate standard lists (e.g. Geographic area code), international standards (e.g. ISO date format), and authority lists (e.g. name authority) may be used to ensure quality of metadata.

Similarly, a set of recommendations may be drawn on the area of metadata reuse.

1) The metadata may be re-used in any medium without prior permission for not-for-profit purposes; and
2) The metadata must not be re-used in any medium for commercial purposes without formal permission.

1.3.3 Usage Metadata

Another important aspect of OA metadata landscape is usage metadata. There are many standards and initiatives for describing and storing usage metadata in the domain of OA such as SURE (Statistics on the Usage of Repositories), PIRUS (Publishers and Institutional Repository Usage Statistics), OA-Statistik, NEEO (Network of European Economists Online), KE-USG (Knowledge Exchange Usage Statistics Guidelines), and OpenAIRE that specify metadata formats to be used to incorporate information of usage events. The usage metadata may serve as an important value-added service for users of open contents. Apart from the contributors and users of open access resources, funding agencies are also interested in availability of integrated usage data to measure research impact and to analyze trends over time. For example, PIRUS suggests to include following metadata elements to record usage of OA resources – i) either print ISSN OR online ISSN; ii) article version, where available; iii) article DOI; iv) online publication date or date of first successful request; and v) monthly count of the number of successful full-text requests. Other optional but desirable metadata elements are - i) journal title; ii) publisher name; iii) platform name; iv) journal DOI; v) article title; and vi) article type. The item level granularity in PIRUS is achieved through two

additional metadata elements – article DOI and ORCHID as author identifier. Most of these initiatives are based on the OpenURL Context Object format. This format includes six elements: i) Referent (the item that was used, e.g. a paper deposited in a repository); ii) Referring Entity (the "atomic" entity within the referrer that contains the reference to the referent, e.g. a Google search hit); iii) Requester (the user or client requesting the referenced item, identified by its IP address); iv) Service Type (the action that is associated with the requested item, e.g. download or metadata view); v) Resolver (the service that holds or resolves to the requested item, e.g. the OAI base-URL of the repository); and vi) Referrer (the web service that provides a reference to the referent, e.g. the Google-search engine).

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

3) Do you think metadata policy is required for organizing OA resources? Explain.

...……..……..……...

……..……..……...

……..……..……...

4) What is usage metadata?

...……..……..……...

……..……..……...

……..……..……...

1.4 GENERIC METADATA SCHEMAS

A large number of standards have evolved for describing electronic resources, but the majorities are concerned with describing very specific resources. The formats like TEI (Text Encoding Initiative), FGDC (Federal Geographic Data Committee), GILS (Global Information Locator Service), OAI (Open Archive Initiative) etc. have been developed to operate within a narrowly defined subject field and generally not suitable for the description of a wider range of resources. These metadata schemas are complex in nature and thereby geared towards creation by experts and interpretation by computers.

The Dublin Core Metadata Element Set (DCMES) or Dublin-core is a small set of resource description categories which is notably different from many of the other metadata schemas due to its ease of use and interoperability. The Dublin Core Metadata Initiative (DCMI), an international community has led the

development of metadata components that enhances cross-disciplinary resource discovery. The mission of DCMI is to develop an easy and seamless mechanism for searching and indexing web resources through – i) developing metadata standards for cross-domain resource discovery; ii) defining frameworks for the interoperation of metadata sets; and iii) facilitating the development of discipline-specific metadata sets that work within the frameworks of cross-domain resource discovery and metadata interoperability. The DC element set is today a *de facto* standard for metadata on the web. The DC metadata set has 15 major elements and these metadata elements fall into three groups – i) elements related mainly to the Content of the resource; ii) elements related mainly to the Resource when viewed as Intellectual Property; and iii) elements related mainly to the Instantiation (Figure2).

"Simple Dublin Core" is DC metadata that uses no qualifiers. It applies only main 15 elements without any qualifier. On the other hand, "Qualified Dublin Core" uses additional qualifiers to increase specificity or precision of the metadata. For example, a "Date" is a DC element which may be specified to identify a particular kind of date (date of last modification, date of publication etc.). The DCMI presently admits two broad classes of qualifier – i) Element Refinement (these qualifiers make the meaning of an element specific); and ii) Encoding Schemes (these qualifiers identify schemes that aid in the interpretation of an element value; these schemes include controlled vocabularies and formal notations e.g. a term from a set of subject headings or standard expression of a date like "2013-12-25").

DC elements are flexible enough for the description of variety of resources in different subject areas. Moreover, the meanings of the elements will be understood by most users. This quality has been achieved by DC metadata by following ***Six Principles***:

- **Intrinsicality**: DC metadata is based on intrinsic data. These data refers to the property that could be identified from the intellectual content and physical form of the resource;
- **Extensibility**: It allows inclusion of extra descriptive materials for specialized requirements;
- **Syntax Independence**: It is applicable to a wide range of disciplines and application program;
- **Optionality**: All the DC elements are optional;
- **Repeatability**: All the DC elements are repeatable. For example, a resource with multiple authorship may use the "Creator" element repeatedly to accommodate all the authors; and

- **Modifiability**: Each element in the Dublin Core has a definition, which is self explanatory. Each element can be modified by an optional qualifier and in such cases the definition of the element is modified by the value of the qualifier.

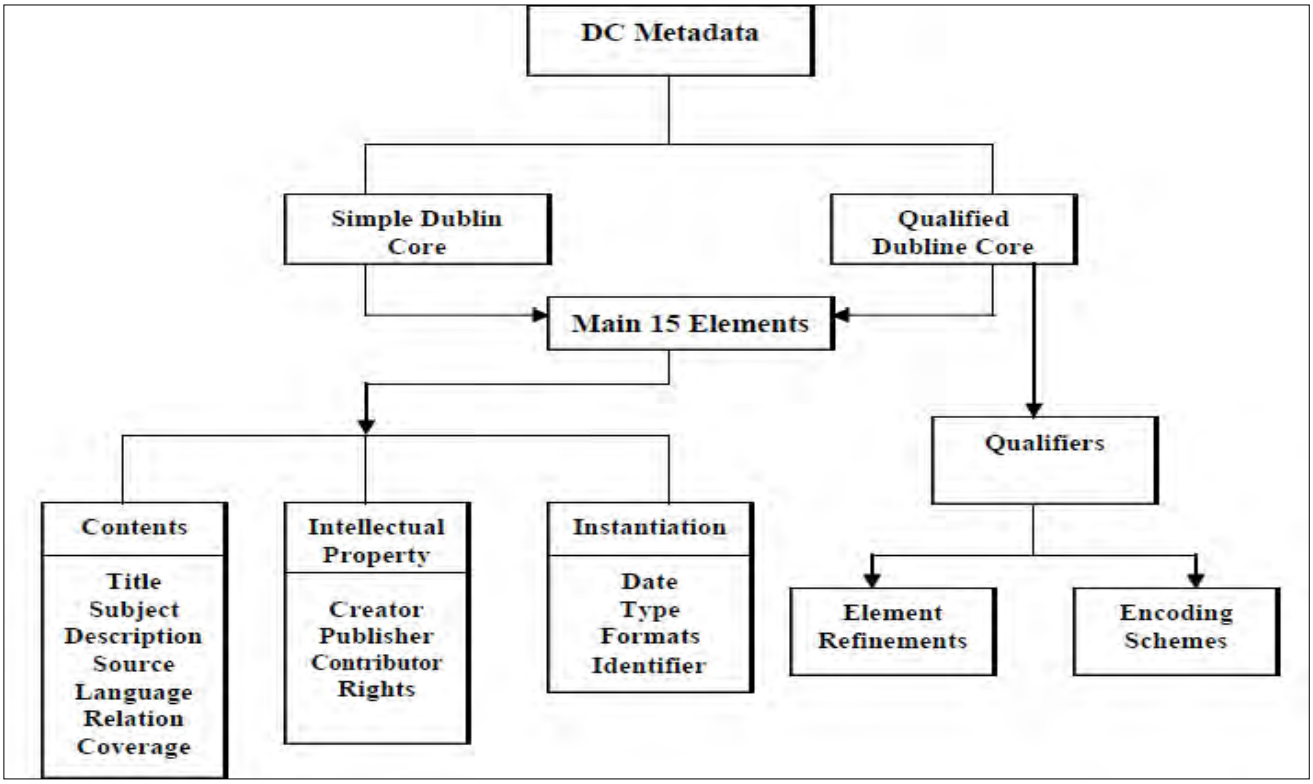

Figure 2: DCMES major Elements

The scope of major DC elements along with element refinement provisions and recommended encoding schemes are listed below for your ready reference.

Table 3: DCMES Major Elements[1]

Sl. No.	***DC Elements (Scope)***	***Element Refinement (Comments, if any)***	***Element Encoding Scheme(s) (Comments, if any)***
1	**Title** (Name given to the resource)	**Alternative** (Substitute to the formal title)	-
2	**Creator** (Entity that created the content)	-	-
3	**Subject** (Topic or Keywords)	-	LCSH, MESH, DDC. LCC. UDC
4	**Description** (Account, summary or abstract of the content)	**Table of Contents** (A list of sub-units of the content of the resource) **Abstract** (A summary of the	-

[1] http://dublincore.org/documents/dces

		contents of the resource)	
5	**Publisher** (Entity that made the resource available)	-	-
6	**Contributor** (Other entity that made a contribution)	-	-
7	**Date** (Date of an event in the life of the resource)	**Created** (Date of creation of the resource) **Valid** (Date/period of validity of a resource) **Available** (Date/period that the resource will become or did become available) **Issued** (Date of formal publication/issuance of the resource) **Modified** (Date on which the resource was changed)	**DCMI Period** (A specification of the limits of a time interval. Available at: http://dublincore.org/documents/dcmi-period/) **W3C-DTF** (W3C encoding rules for dated and times based on ISO 8601. Available at: http://www.w3.org/TR/NOTE-datetime)
8	**Type** (Nature, genre or category of the resource)	-	**DCMI Type Vocabulary** (A list of types used to categorise the nature or genre of the content of the resource. Available at: http://dublincore.org/documents/dcmi-type-vocabulary/)
9	**Format** (Physical or digital manifestation of the resource)	**Extent** (The size or duration of the resource) **Medium** (The material or physical carrier of the resource)	- **IMT** (The Internet media type of the resource. Available at: http://www.isi.edu/in-notes/iana/assignments/media-types/media-types)
10	**Identifier** (An unambiguous reference to the resource within a given context)	-	**URI** (A uniform resource identifier. Available at: http://www.ietf.org/rfc/rfc2396.txt)
11	**Source** (Reference to the resource's origin)	-	**URI** (A uniform resource identifier. Available at: http://www.ietf.org/rfc/rfc2396.txt)
12	**Language** (Language of the content of the resource)	-	**ISO 639-2** (Codes for the representation of names of languages. Available at: http://www.lcweb.loc.gov/

			standards/iso639-2/langhome.html)
13	**Relation** (Reference to a related resource)	**Is Version Of** (The described resource is a version, edition, or adaptation of the referenced resource) **Has Version** (The described resource has a version, edition or adaptation, namely the referenced resource) **Is Replaced By** (The described resource is supplanted, displaced or superseeded by the referenced resource) **Replaces** (The described resource supplants, displaces or superseds the referenced resource) **Is Required By** (The described resource is required by the referenced resource either physically or logically) **Requires** (The described resource is a physical or logical part of the referenced resource) **Is Part Of** (The described resource is a physical or logical part of the referenced resource) **Has Part** (The described resource includes the referenced resource either physically or logically) **Is Referenced By** (The described resource is referenced, cited or otherwise pointed to by the referenced resource) **References** (The described resource references, cites, or otherwise points to the referenced resource) **Is Format Of** (The described resource is the same intellectual	**URI** (A uniform resource identifier. Available at: http://www.ietf.org/rfc/rfc2396.txt)

Interoperability and Retrieval

		content of the referenced resource, but presented in another format) **Has Format** (The described resource pre-existed the referenced resource, which is essentially the same intellectual content presented in another format)	
14	**Coverage** (Extent or scope of the content of the resource)	**Spatial** (Spatial characteristics of the intellectual content of the resource e.g. place name or geographic coordinates) **Temporal** (Temporal characteristics of the intellectual contents of the resources e.g. a period label or date range)	**DCMI Point** (Identifies a point in space using its geographic coordinates. Available at: http://www.dublincore.org/documents/dcmi-point/) **ISO 3166** (Codes for the representation of names of countries. Available at: http://www.din/de/germien/nas/nabd/is03166ma/codlstp/index.html) **DCMI Box** (A specification of the limits of a time interval. Available at: http://dublincore.org/documents/dcmi-box/) **TGN** (The Getty thesaurus of geographic names. Available at: http://shiva.pub.getty.edu/tgn_browser) **DCMI Period** (A specification of the limits of a time interval. Available at: http://dublinecore.org/documents/dcmi-period/) **W3C-DTF** (Rules for encoding dates and times, based on ISO 8601. Available at: http://www.w3.org/TR/NOTE-datetime)
15	**Rights** (Information about rights held in and over the resource)	-	-

Most of the digital repository management software (e.g. Greenstone, Eprint, Dspace) include simple DCMES and qualified DCMES by default. The metadata entry interface of Greenstone is given in Figure 3. The metadata entered by submitters and/or librarians are stored in repository management software generally in XML format.

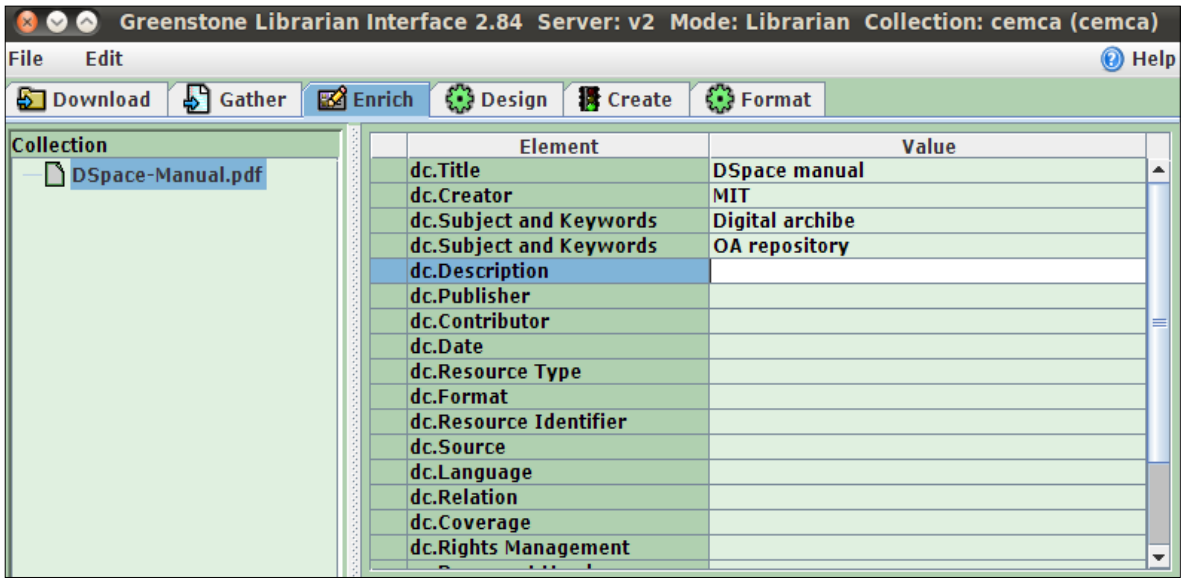

Figure 3: Interface in Metadata Entry in Greenstone

(Source: *Greenstone software*)

The metadata as entered in Figure 3 are stored inside the Greenstone in the following format as metadata.xml file.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE DirectoryMetadata SYSTEM
"http://greenstone.org/dtd/DirectoryMetadata/1.0/DirectoryMetadata.dtd">
<DirectoryMetadata>
  <FileSet>
    <FileName>DSpace-Manual\.pdf</FileName>
    <Description>
      <Metadata mode="accumulate" name="dc.Title">DSpace manual</Metadata>
      <Metadata mode="accumulate" name="dc.Creator">MIT</Metadata>
      <Metadata mode="accumulate" name="dc.Subject">Digital archibe</Metadata>
      <Metadata mode="accumulate" name="dc.Subject">OA repository</Metadata>
    </Description>
  </FileSet>
</DirectoryMetadata>
```

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

5) Discuss the intrinsic principle of DCMES.

...

...

...

...

...

6) What is Qualified DCMES? How does it differ from Simple DCMES?

...

...

...

...

...

1.5 DOMAIN-SPECIFIC METADATA SCHEMAS

DCMES consists of 15 basic elements only. These fields, being generic for any type of digital resource, do not capture any specific information about specialized contents such as maps, images, video objects, learning materials, ETDs etc. Although DC attributes such as authors, title, subject etc are definitely useful for specialized OA contents (other than journal articles) like learning objects and ETDs but at the same time DCMES does not contain attributes to describe essential attributes of specialized contents (like name of degree awarded for dissertations or learning outcome in case of learning objects). In other words, no single metadata element set will accommodate the functional requirements of all organizations or communities of practice. A generic metadata schema is not sufficient enough to describe different type of resources with all relevant elements. In OA landscape, journal articles are possibly the most visible objects and next come learning objects and ETDs. Open learning resources are increasingly available in different forms and formats (such as Moocs). An analysis of OpenDOAR shows that many repositories include ETDs as contents and some repositories are exclusively dealing with ETDs. This section therefore covers primarily learning objects and ETDs.

1.5.1 Learning Objects Domain

Learning objects are digital educational materials with pedagogical perspective. Reuse of learning materials is highly desirable and it is ensured through semantically tagging them with standard metadata. Efficient retrieval of learning materials requires applying domain-specific schema to describe educational attributes such as topic of the document, type of the document etc. In order to cope with educational concerns, various metadata standards have been developed namely IMS Metadata, SCORM, CanCore, GEM and IEEE Learning Object Metadata, AICC, ARIADNE etc. Here we will be discussing in brief three major learning object metadata schemas. Three major schemas are finally compared with three other schemas to provide you an insight of comprehensiveness of schemas.

IEEE Learning Object Metadata (IEEE - LOM)

The Learning Objects Meta data schema was published by the Institute of Electrical and Electronics Engineers (IEEE) in 2002. The IEEE Learning Object Metadata[2] aims to develop technical standards, recommended practices, and guides for learning technology. The LOM standard is mainly built on the Dublin Core and is based on the recommendations of IMS and ARIADNE project. It is a multi-part standard and contains a description of semantics, vocabulary, and extensions. LOM has a wide set of globally agreed metadata elements which are grouped into nine descriptive categories: General, Life cycle, Meta metadata, Technical, Educational, Rights, Relation, Annotation, and Classification. The LOM data model is a hierarchy of data elements, including aggregate data elements and simple data elements. It specifies a conceptual data schema that defines the structure of a metadata instance for a learning object. It is intended to reference by other standards that define the implementation descriptions of the data schema and thereby ensures reuse and exchange learning objects. The purpose of the IEEE LOM is to facilitate acquisition, search, evaluation and use of learning objects. It is intended to facilitate the sharing and exchange of learning objects by enabling the development of catalogs and inventories while taking into account the diversity of cultural and lingual contexts in which the learning objects and their metadata are reused (IEEE, 2013).

IMS

The IMS Global Learning Consortium[3] has developed and promotes the adoption of open technical specifications for interoperable learning technology. IMS is based on LOM and Dublin Core metadata. The IMS Global Learning Consortium, Inc. (IMS) project was launched by EDUCAUSE (formerly EDUCOM), a consortium of North American educational institutions and their commercial and government partners to define open technical standards for the interoperation of distributed learning applications and services (Anido et al., 2002). IMS develops and promotes open

[2] http://ltsc.ieee.org/wg12/index.html
[3] http://www imsglobal.org

specifications for facilitating online distributed learning activities such as locating and using educational contents, tracking learner progress, reporting learner performance, and exchanging student records between administrative systems (IMS, 2003). IMS is very attentive to the needs of those in the educational community generally and has the highest recognition within this community of the standards development organizations (Friesen, 2002). The IMS Content Packaging Information Model defines a standardized set of structure that can be used to exchange the learning contents. These structures provide the basis for standardized data bindings that allow the software developers and the implementers to create instructional materials that are interoperable across authoring tools, learning management systems, and run time environments. IMS has two fundamental goals: to define specific guidelines which guarantee interoperability between applications and services in e-learning; and to support guidelines application in international products and services.

SCORM (Sharable Content Object Reference Model)

SCORM was developed in 2003 by an organization called Advanced Distributed Learning (ADL). The SCORM Metadata Application Profile directly references the IEEE Learning Object Metadata (LOM) standard. It **provides specific guidance for applying metadata to learning resources.** SCORM is globally accepted as the standard for management of educational contents. It is a collection of specifications adapted from multiple sources to provide a comprehensive suite of e-learning capabilities that enable interoperability, accessibility and reusability of Web-based learning contents. The SCORM-compliant courses are reusable, accessible, interoperable and durable. It is a model that references and integrates a set of interrelated technical standards, specifications and guidelines designed to meet ADL's functional requirements, such as, accessibility, interoperability, durability and reusability for learning contents and systems.

Table 4: Comparison of LO Metadata Schemas

Attributes/Parameters	Score (1= full support; 0.5= partial support; 0= no support)					
	AICC	ARIADNE	Dublin Core (DC)	IEEE - LOM	IMS	SCORM
Absence of "dumb down" principle	1	1	0	1	1	1
No requirement of Qualified special entities	1	1	0	1	0	1
Data extraction from the entities	0	0	0	1	1	0
Form of the metadata	0	0	1	1	1	0
Flexibility of metadata schema	1	1	0	0	1	0
Vocabulary management	0	0	0	1	1	1
Multi-language support	0	0	1	1	0	0
Metadata templates	0	0	1	1	0	1
Consistency in presenting information	0.5	0.5	1	1	0	0
Application Programming Interface (API)	0	0	0	1	0	1
Multi-part standard	0	0	0	1	1	1
Learner Profile Tracking, learner progress, performance exchanging student records	0	0	0	1	1	1
Support multiple users	1	1	1	1	1	0
Independent structured data models	0	0	0	1	1	1
Creation of new metadata files	0	0	1	1	0	1
Modification of data in metadata files	0	0	1	1	0	1
Support of the XML	0	0	1	1	1	0
Total Score	**4.5**	**4.5**	**8.0**	**16.0**	**10.0**	**9.0**

1.5.2 Theses and Dissertations

This section covers three comprehensive metadata schemas in the domain of electronic theses and dissertations (ETD) namely ETD-MS, UK-ETD, and Shodhganga (mainly used in Indian universities).

ETD-MS

NDLTD is the developer of ETD-MS. The initial goal of NDLTD was to develop a standard XML DTD for encoding metadata elements for ETDs. ETDMS is based on the Dublin Core Element Set, but includes an additional element specific to metadata regarding theses and dissertations. Despite its name, ETDMS is designed to deal with metadata associated with both paper and electronic theses and dissertations. It is also designed to handle metadata in many languages, including metadata regarding a single work that has been recorded in different languages.

UK-ETD

This metadata standard is recommended by Electronic Theses Online Service (EThOS), UK. EThOS is the Electronic Theses Online System which allows individuals to find access and archive doctoral e-theses that are produced in UK Higher Education institutions. Funding from the Joint Information Systems Committee (JISC) enabled three project teams in the UK to study the issues and challenges associated with the deposit and management of theses in electronic format. It was considered important to recommend a standard set of metadata elements to describe the contents of e-theses repositories. The schema conforms to the guidelines for implementing Dublin Core in XML.

Shodhganga

The Indian ETD repository called Shodhganga (maintained by INFLIBNET, an Inter University consortium under University Grants Commission, India) originated to facilitate open access to theses amongst the academic community. The word ‘shodh’ originates from Sanskrit and means research and discovery. Ganga is the name of the largest and holiest river in India. This project was intended to provide online accessibility to Indian theses for archiving and free access. The Shodhganga metadata schema has been developed as domain-specific schema to deal with ETDs of Indian universities. Shodhganga uses the qualified Dublin core set of elements for furnishing metadata in order to provide global access of Indian research outputs. The basic DC sets consists of 15 elements and the qualified set has about 31 elements in Shodhganga. A comparison of these three schemas against a set of carefully crafted parameters may help to assess quality and comprehensiveness of these schemas.

Table 5: Comparison of Three Metadata Standards for Theses & Dissertations

Metadata elements	Scope of Metadata for ETDs	ETD -MS	UK-ETD	Shodh ganga
dc.thesis.degree	Name of the degree to which thesis/dissertation is associated. For example MPhil/PhD	Y	N	N
thesis.degree.level	For example Master's, Doctoral, Post-Doctoral etc.	Y	N	N
thesis.degree.discipline	Name of the department e.g. Bengali, English, Library and Information Science etc	Y	N	N
thesis.degree.grantor	Name of the degree awarding University/Institution	Y	N	N
dc.rights.embargotype	Whether only campus access or part/section of the thesis/dissertation can be accessed	N	Y	N
dc.rights.embargodate	Embargo period i.e. date before which ETD may not be publicly available	N	Y	N
dc.rights.embargoreason	The reason of embargo e. g. applied for patent etc.	N	Y	N
dc.relation	If any other relation with the thesis	N	N	Y
dc.relation.isReferenced By	The metadata 'jump off' page for the ETD at the institutional repository	N	Y	N
dc.relationhasVersions	Citations to previously published works related to ETD.	N	Y	N
dc.relation.references	References to other works	N	Y	N
dc.description.abstract	Abstract of the ETD	Y	N	Y
dc.description.note	ETD acceptance note of the department if any	Y	N	Y
dc.description.release	If any description of the version of the ETD	Y	N	N
dc.publisher	Name of the publisher as it appears on the title page of thesis/dissertation	Y	N	N
dc.publisher.department	Name of school, department, centre, faculty of the researcher	N	Y	N
dc.publisher.commercial	Name of the formal publisher of the thesis (If any)	N	Y	N
dc.publisher.place	Place of publication	N	N	Y
dc.publisher.university	Name of the degree awarding university	N	N	Y
dc.publisher.institution	Name of the degree awarding institution	N	Y	Y
dc.contributor	Name of the T/D supervisor(s)/guide(s)/advisors/committee member(s) etc.	Y	N	N
dc.contributor.role	Role of the person in creation the T/D e.g.	Y	N	N

	Guide/Supervisor/Advisor/ Committee member etc.			
dc.contributor.sponsor	Sponsor of the researcher/student	N	Y	N
dc.contributor.release	If any errata published by researcher	N	N	Y
dc.contributor.guide	Name of the guide, repeatable in case of co-guide	N	N	Y
dc.date	Date appears on the title page of the T/D according to ISO 8601 standard	Y	N	N
dc.date.issued	Date appears on the title page (format yyyy-mm or yyyy) according to ISO8601	N	Y	N
dc.date.registered	PhD registration date	N	N	Y
dc.date.completed	PhD completion date	N	N	Y
dc.date.awarded	Date of PhD degree award (ISO 8601 format i.e. yyyy-mmdd)	N	N	Y
dc.type.qualificationlevel	Level of the degree (e.g. Diploma, Masters, Doctoral, Postdoctoral)	N	Y	N
dc.type.qualificationname	Name of the degree e.g. MPhil, PhD, DPhil	N	Y	N
dc.format.accompanying material	If any accompanying material released with thesis	N	N	Y
dc.format.dimensons	Size of the thesis	N	N	Y
dc.format.extent	Pagination for text, time duration in case moving image, file size in bytes for electronic file	N	N	Y
dc.format.medium	File format name (auto identified by the system)	N	N	Y
dc.format	File type or in which format T/D is appeared e.g. pdf, doc, html, odt etc.	Y	N	N
dc.identifier	This element used for URL of thesis/dissertation/ ID for physical objects i.e. in the case printed T/D	Y	Y	N
dc.identifier.URI	URL of the electronic thesis/ ID for electronic objects i.e. for ET/D	N	Y	Y
dc.identifier.thesis number	If any thesis number allotted by INFLIBNET Centre	N	N	Y
dc.identifier.handle	If any handle number provided by system	N	N	Y
dc.coverage	Time period or spatial area covered in thesis/dissertation	Y	N	Y
dc.source	If the thesis harvested from the Institutional / ETD repository	N	N	Y

1.5.3 Other Domains

An illustrative list of popular domain-specific metadata schemas are given here in alphabetical order:

- **ABCD[4] - Access to Biological Collection Data**: An evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data) sponsored by Biodiversity Information Standards TDWG - the Taxonomic Databases Working Group (last modified in 2007).
- **AGLS[5]** (Australian Government Locator Service): AGLS is an Australian government metadata standard intended for the description of government resources on the Web. It uses DCMI Terms properties with a few additional metadata elements such as function and mandate.
- **AgMES[6] - Agricultural Metadata Element Set**: AgMES, developed by the Food and Agriculture Organization (FAO) of the United Nations enables description, resource discovery, interoperability and data exchange of different types of information resources in all areas relevant to food production, nutrition and rural development (last modified in 2010).
- **CanCore[7]** : CanCore is a set of guidelines for the implementation of the IEEE LOM metadata standard for describing learning resources. It is originated in Canada for managing learning objects in Canadian universities.
- **CSMD-CCLRC[8]** (Core Scientific Metadata Model): It is designed by Science and Technologies Facilities Council to support data collected within a large-scale facility's scientific workflow but the model is also designed to be generic across scientific disciplines (last modified in 2011).
- **Cataloguing Cultural Objects[9]** (CCO): A schema for cultural objects, developed by the US-based Visual Resources Association with significant input from the Getty Research Institute (last modified in 2010).
- **Categories for the Description of Works of Art[10]** (CDWA): An extensive metadata schema for cataloguing objects held by art museums developed in the US in the 1990s by the Getty Research Institute (last modified in 2010).
- **Darwin Core[11]**: A metadata schema developed Biodiversity Information

[4] http://www.dcc.ac.uk/resources/metadata-standards/abcd-access-biological-collection-data
[5] http://www.naa.gov.au/records-management/create-capture-describe/describe/AGLS/index.aspx
[6] http://www.dcc.ac.uk/resources/metadata-standards/agmes-agricultural-metadata-element-set
[7] http://cancore.athabascau.ca/en/index.html
[8] http://www.dcc.ac.uk/resources/metadata-standards/csmd-cclrc-core-scientific-metadata-model
[9] http://cco.vrafoundation.org/
[10] http://www.getty.edu/research/conducting_research/standards/cdwa/
[11] http://www.dcc.ac.uk/resources/metadata-standards/darwin-core

Standards (TWDG) to cover elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity (last modified in 2009).

- **DataCite**[12] Metadata Schema: A set of mandatory metadata elements prescribed by DataCite consortium to support persistent approach to access, identification, sharing, and re-use of digital research datasets (last modified in 2013).
- **DDI - Data Documentation Initiative**[13]: A globally recognized standard for describing data from the social, behavioral, and economics and statistics. The XML based DDI metadata specification supports the entire research data life cycle (last modified in 2009).
- **DIF - Directory Interchange Format**[14]: A domain-specific schema for Earth sciences community, intended for the description of scientific data sets. It includes elements focusing on instruments that capture data, temporal and spatial characteristics of the data (last modified in 2010).
- **e-GMS**[15]: A schema dedicated to e-governance developed in UK for describing information resources to ensure maximum consistency of metadata across public sector organizations in the UK.
- **Encoded Archival Description**[16] (EAD): A well-known schema that provides an encoding for archival descriptions. It adopts a multi-level approach to description, providing information about a collection as a whole and then breaking it down into groups, series and (if significant) individual items, grew out of work done at UC Berkeley in the mid 1990s and was influenced by TEI and ISAD(G) (last modified in 2002).
- **EXIF**[17] (Exchangeable Image File Format): A technical metadata standard that can be written to and read from a still image file itself (and formats). It was developed by JEITA (Japan Electronics and Information Technology Industries Association).
- **FGDC/CSDGM**[18] - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata: A widely-used, schema for digital geospatial data required by the US Federal Government. It is sponsored by the US Federal Geographic Data Committee (last modified in 2010).
- **FOAF**[19] (Friend of a Friend): FOAF is a RDF-enabled schema for describing people and intended to be used on the Semantic Web. It includes features for encoding names, email addresses, personal interests,

[12] http://www.dcc.ac.uk/resources/metadata-standards/datacite-metadata-schema
[13] http://www.dcc.ac.uk/resources/metadata-standards/ddi-data-documentation-initiative
[14] http://www.dcc.ac.uk/resources/metadata-standards/dif-directory-interchange-format
[15] http://www.govtalk.gov.uk/
[16] http://www.loc.gov/ead/
[17] http://www.exif.org
[18] http://www.dcc.ac.uk/resources/metadata-standards/fgdccsdgm-federal-geographic-data-committee-content-standard-digital-ge
[19] http://www.foaf-project.org/

home pages, and various online identities. In future traditional library authority files may be translated into FOAF but it needs to settle two very important issues – i) each individual has only one FOAF identity; and ii) FOAF focuses on online presence for current living persons.

- **Genome Metadata**[20]: A schema dedicated to the field of Genomics. It consists of 61 different metadata fields covering broad categories: Organism Info, Isolate Info, Host Info, Sequence Info, Phenotype Info, Project Info, and Others (last modified in 2009).
- **GEM**[21] (Gateway to Educational Materials): GEM is an RDF-enabled metadata vocabulary designed for the description of educational resources. The GEM model includes all the properties available in DCMI Terms, with a few additional education-specific elements such as educational standards and pedagogical methods.
- **GILS**[22]: Global Information Locator Service or GILS is a schema for governments, companies, or other organizations to support citizen/customer facing information services. GILS was an early metadata standard for the encoding of descriptive information for government records
- **International Virtual Observatory Alliance Technical Specifications**[23]: A schema for astronomical objects developed by the IVOA (International Virtual Observatory Alliance) to enable interoperability between and the integration of astronomical archives across the world into an international virtual observatory (last modified in 2009).
- **ISO 19115**[24]: An internationally-adopted schema for describing GIS (geographic information and services). It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data (last modified in 2009).
- **MathML**[25] (Mathematical Markup Language): MathML is a W3C Recommendation for the low-level encoding of mathematical information (mathematical data and for the content of the mathematical data) with the intention of representing this information on the Web.
- **MIDAS**[26]: It is a UK standard for describing cultural heritage assets that form the historic environment (buildings, archaeological sites, shipwrecks, areas of interest, artifacts and ecofacts).
- **MIX**[27]: It is an XML based schema for encoding the Technical Metadata

[20] http://www.dcc.ac.uk/resources/metadata-standards/genome-metadata
[21] http://www.thegateway.org/about/ documentation2/ schemas/index_html
[22] http://www.gils.net/
[23] http://www.dcc.ac.uk/resources/metadata-standards/international-virtual-observatory-alliance-technical-specifications
[24] http://www.dcc.ac.uk/resources/metadata-standards/iso-19115
[25] http://www.w3.org/Math/
[26] http://www.english-heritage.org.uk/server/show/nav.8331
[27] http://www.loc.gov/standards/mix/

for Digital Still Images standard, developed by NISO group on Metadata for Images in XML ((last modified in 2009).

- **NewsML**[28] (News Markup Language): The NewsML aims to design a complex schema for describing textual news, articles, photos, graphics, audio, and video — the components that make up or express news items.
- **OAI-ORE**[29] **(**Open Archives Initiative Object Reuse and Exchange: A W3C standard for managing rich content in aggregations of Web resources and supporting activities like authoring, deposit, exchange, visualization, reuse, and preservation (last modified in 2011).
- **OAIS**[30] (Open Archival Information System): OAIS is a "reference model" schema to support preservation of digital information. OAIS includes three subcategories – i) Submission Information Package (SIP) to support the content and metadata received from a preservation repository; ii) Archival Information Package (AIP) to support content and metadata managed by a preservation repository; iii) Dissemination Information Package (DIP) to support end user in response to a request, and may contain content spanning multiple AIPs. OAIS-compliant repository software supports a certain level of functionality and standardization of features.
- **ONIX**[31]: A schema developed by book industry to support Online Information Exchange - international standard for representing and communicating book industry product information in electronic form.
- **PBCore**[32]: Public Broadcasting Metadata Dictionary or PBCore is intended for use by television, radio and web broadcasters and hopes to describe and retrieve broadcast contents efficiently (last modified in 2011).
- **PREMIS**[33]**:** A technical metadata schema that provides a "dictionary" of core metadata elements that can be used to support the digital preservation of a resource. A key feature of the PREMIS model is the definition of Objects as made up of Representations, Files, and Bit streams. It was particularly influenced by a conceptual model called the Open Archival Information System. The Library of Congress is the official PREMIS maintenance agency (last modified in 2006).
- **SPECTRUM**[34]: A key UK standard for museum documentation (last modified in 2005).
- **SDMX**[35] - Statistical Data and Metadata Exchange: A set of common technical and statistical standards and guidelines to be used for the efficient

[28] http://www.newsml.org
[29] http://www.dcc.ac.uk/ resources/metadata-standards/oai-ore-open-archives-initiative-object-reuse-and-exchange
[30] http://public.ccsds.org/publications/ archive/650x0b1.pdf
[31] http://www.editeur.org/onix.html
[32] http://www.pbcore.org
[33] http://www.loc.gov/standards/premis
[34] http://www.collectionslink.org.uk/manage_information/spectrum
[35] http://www.dcc.ac.uk/resources/metadata-standards/sdmx-statistical-data-and-metadata-exchange

exchange and sharing of statistical data and metadata (last modified in 2012).

- **SKOS**[36] (Simple Knowledge Organization System): SKOS is a W3C standard for encoding structured vocabularies in RDF. The RDF SKOS vocabulary focuses on describing concepts, which are represented by terms, and documenting relationships between concepts.
- **SWAP**[37] (Scholarly Works Application Profile): SWAP is a DCMI-compliant application profile for the description of scholarly works, developed by UKOLN. It aims to support quality metadata encoding of knowledge objects in Green OA. SWAP is based on the FRBR conceptual model, and therefore differentiates between Works and their Manifestations.
- **Text Encoding Initiative (TEI) Header**[38]: It is a scheme for marking up electronic text. It also specifies a header portion to accommodate metadata about the object to be described. TEI headers can be used to record bibliographic information of both electronic and non-electronic sources. The TEI header can be mapped to and from MARC.
- **VRACore**[39] (Visual Resources Association Core Categories): A widely used metadata schema for describing art or cultural images, providing 17 core categories (last modified in 2007).
- **XrML**[40] (eXtensible Rights Markup Language): XrML is an XML language for the encoding of rights information. It is focused on the action of "granting" authorizations between Principals, Rights, Resources, and Conditions.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*

b) Compare your answers with those given at the end of this Module.

7) Mention metadata schema for the following domains – ETD, Image, Maps, Learning Objects, Cultural objects, and Compound digital objects.

...……..……..……...

……..……..……...

……..……..……...

……..……..……...

[36] http://www.w3.org/2004/02/skos/
[37] http://www.ukoln.ac.uk/repositories/ digirep/index/Scholarly_Works_Application_Profile
[38] http://www.tei-c.org
[39] http://www.vraweb.org/ organization/committees/datastandards/index.html
[40] http://www.xrml.org/

8) Name application area for the following domain-specific metadata schemas – AGLS, SWAP, MIDAS, ONIX, GILS and e-GMS.

……..……..……...

……..……..……...

……..……..……...

1.6 METADATA MODELING

As a library professional, you know that Paris Principles and ISBDs have served the role of bibliographic foundation for almost all the national and international cataloguing codes. But the environment within which cataloguing principles and standards operate has changed fundamentally and also substantially because of the emergence of computerized processing of bibliographic data, growth of large-scale databases, increasing use of shared cataloguing programmes, and proliferation of digital resources in Web and in libraries. Such a situation requires some general framework to assist in the understanding and further development of conventions for bibliographic description. Models for bibliographic description provide a logical base for the correlation of cataloguing rules with the data encoding structure. A model for bibliographic description endeavors to address complex bibliographic problems and provides a strong foundation to support retrieval, presentation and transfer systems in integrated environment.

1.6.1 Bibliographic Data Models

Some of the groundbreaking works towards developing bibliographic data models are discussed here to show you the application of model in resource description.

A. UKOLN's Analytical Model of Collections and their Catalogues

This model has been developed in 2000 by United Kingdom Office for Library and Information Networking (UKOLN) under the Research Support Libraries Programme (RSLP). It is applicable to physical and digital collections of all kinds, including library, art and museum materials. This model identifies 3 main entities and associated attributes — Objects (Content, Item, Collection, Location, Content-Component, Item-Component); Agents (Creator, Producer, Collector, Owner, Administrator); Indirect-Agents (Creator's Assignee, Producer's Assignee). It also prescribes two types of relationships — internal relationships (relationships between the entities in Collection Description) and external relationships (relationships between Collection Descriptions themselves). The model tries to clarify the points at which rights and conditions of access and use become operable and attempts to act as a bridge linking collections and their users.

B. IFLA Models

IFLA developed a total of three related bibliographic data models in the span of 1998 to 2010. The first one of the series is FRBR and it is followed by FRAD and FRSAD. All these models of IFLA are based upon E-R data modeling and can be applied to print resources as well as digital resources. These three data models are proposed by IFLA during 1998-2010 to upgrade standards of resource description in digital environment. FRBR (Functional Requirements for Bibliographic Data) appeared first in 1998 followed by FRAD (Functional Requirements for Authority Data) in 2009 and FRSAD (Functional Requirements for Subject Authority Data) in 2010. FRBR deals with ER modeling of bibliographic data, FRAD deals with ER modeling of authority data and FRSAD deals with subject authority data. These three ER modeling standards aim to manage bibliographic and authority data at tandem.

The FRBR model (Functional Requirements for Bibliographic Records) is a conceptual model that was developed by an IFLA group of experts from 1992 to 1997 and finally published in the year 1998. The model uses entity-relation techniques to identify entity, attributes and relationships in the bibliographic universe. It also identifies the relevance of each attribute and relationship to the generic tasks performed by users of bibliographic data. In FRBR model, the entities of bibliographic universe have been divided into three groups: i) the first group includes the products of intellectual or artistic endeavor; ii) the second group comprises those entities responsible for the intellectual or artistic content; and iii) the third group identifies entities that serve as the subjects of intellectual or artistic endeavor.

- **Group I:** The entities of this group represent the different aspects of user interests in the products of intellectual or artistic endeavor. These are: Work (a distinct intellectual or artistic creation; Expression (the intellectual or artistic realization of a work); Manifestation (the physical embodiment of an expression of a work; and Item (a single exemplar of a manifestation).
- **Group II:** The entities in the second group represent those responsible for the intellectual or artistic content, the physical production and dissemination, or the custodianship of the entities in the first group. The entities in this group include *person* (an individual) and ***corporate body*** (an organization or group of individuals and/or organizations).
- **Group III**: The entities of this group represent an additional set of entities that serve as the subjects of works. It includes concept (an abstract notion or idea), object (a material thing), event (an action or occurrence), and place (a location).

FRAD (Functional Requirements for Authority Data) is a new authority data model developed by IFLA recently. Library catalogue supports two major groups of functions – i) Finding function; and ii) Collocation function. Collocation functions require the support of authority control. The typical functions of authority control are as follows - 1) Document decisions; 2) Serve as reference tool; 3) Control forms of access points; 4) Support access to

bibliographic file; and 5) Link bibliographic and authority files. The Functional Requirements for Authority Data (FRAD) is a conceptual model and a companion document to the Functional Requirements for Bibliographic Records (FRBR) conceptual model. FRAD includes additional attributes for each of the Group 1, 2, and 3 entities, as well as a new Group 2 entity (Family). It also includes attributes intended to support the authority control process (Name, Identifier, Controlled Access Point, Rules, and Agency). In addition to expanded entities and attributes, FRAD defines a different set of user tasks for authority data than FRBR did for bibliographic data. Here, the user tasks are Find, Identify, Contextualize, and Justify. The FRAD model, together with FRBR, serves as the foundation of the content standard Resource Description and Access (RDA).

The FRSAR Group finalized in 2010 the ***FRSAD model (Functional Requirements for Subject Authority Data)***, which was published in English both as a printed book and online. This model focuses on the relationships between a work, its subjects, the way these subjects are named, and the information contained in indexing schemes about both the concepts and the appellations that refer to them.

C. XML Organic Bibliographic Information Schema (XOBIS)

XOBIS attempts to restructure bibliographic and authority data in a consistent and unified manner using Extensible Markup Language (XML). It has been developed at Lane Medical Library, Stanford University under the Medlane Project. The preliminary version (alpha version) of XOBIS appeared in September 2002. XOBIS prescribes a tripartite record element based structure in which each record consists of three required components. These are Control Data (contains metadata about record), Principal Elements (10 categories of data that provide bibliographic access and authority control to a wide variety of resources) and Relationships (element that accommodates links between any pair of principal elements). The basic structure of XOBIS may be illustrated in Figure 4:

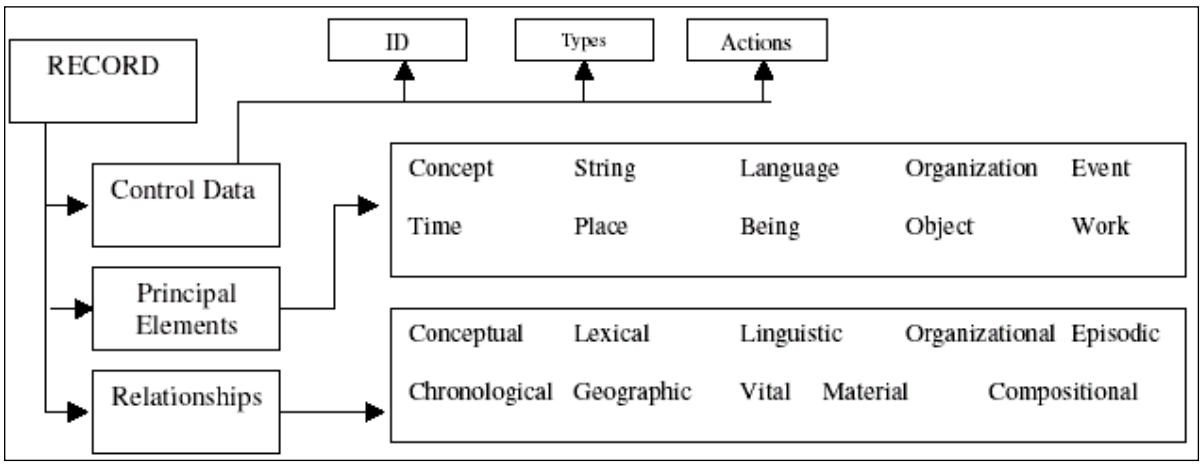

Figure 4: XOBIS Model[41]

[41] http://xobis.stanford.edu/

XOBIS is an experimental model for resource description in XML schema. The aim of XOBIS is to achieve integration of digital world and print world as far as resource description area is concerned. It is generally used to retrieve MARC records from remote library catalogs, including OCLC's World Cat, to facilitate copy cataloging and sharing of bibliographic records.

1.6.2 Applications of RDF and XML

The Network Working Group of Internet Society issued a memo in December 1999 (RFC: 2731) for encoding DC metadata in HTML. As per this memo the general syntax is

<meta name = "PREFIX.ELEMENT_NAME"
content = "ELEMENT_VALUE">

The capitalized words are replaced by the actual value at the time of description. Therefore, the authorship of this unit may be encoded as follows:

<meta name = "DC.Title" content = "Module 4: Resource description">

<meta name ="DC.Creator" content = "Mukhopadhyay, Parthasarathi">

The current activities of W3C are centered on the development and standardization of two important projects, XML and RDF. The Extensible Markup Language (XML) is a data format for structured document interchange on the web. XML permits the web authors to add tags as necessary. It is intended to make easy and straightforward use of SGML in the web. The extensible feature of XML will make the encoding of metadata easier and more flexible. But this strength of XML leads to a serious problem in standardization. Any one can create a set of tags for describing resources. It reduces the scope of harmonization of various metadata schemas. Thus, along with the XML, the web also requires a unifying architecture to accommodate different metadata schemas from various communities. The Resource Description Framework (RDF) is a W3C initiative in this direction. DC metadata (IETF RFC: 2413) and RDF are two distinct specifications but both the communities have a number of members common and have evolved side-by-side. In fact, RDF is based on Warwick Framework, a major recommendation of the Second DC Workshop at Warwick in 1996. The co-evolution of DCMES and RDF forms a natural complement within the web's greater metadata architecture. The DC has provided a semantic focus for RDF, and in turn, RDF has clarified the importance of a formal underlying data model for DC metadata. RDF is a meta-language for representing information, and serves as a key piece of the technical framework underlying Semantic Web activities. RDF defines its statements in "triples": the subject is what is being described, the predicate is an indication of what property of the subject is being described by the statement, and the object is the value of the property. A simple RDF model has three parts called RDF Triples. It says that a fact represented has three parts: a subject, a predicate (i.e. verb), and an object. The subject is what's at the start of the edge, the predicate is the type of edge (its

label), and the object is what's at the end of the edge. The subjects, predicates, and objects in RDF always indicate things: concrete things or abstract concepts. The things that names denote are called **resources** or **nodes** or **entities**. Predicates indicate relations between two things. RDF also specifies that names for subjects, predicates, and objects must be expressed in Uniform Resource Identifiers (URIs). RDF uses XML namespace for identification of metadata schema. An XML namespace is a collection of names, identified by a URI reference that are used in XML documents as element types and attribute names. As per the recommendation of DCMI, the URI of the namespace for all DCMI elements that comprise the DCMES version 1.1 is http://purl.org/dc/elements/1.1/. Therefore, within the RDF documents, it may appear as xmlns: dc = http://purl.org/dc/elements/1.1/. We already know that an expression in RDF is a "triple," consisting of a subject (the object being described e.g., the sky), a predicate (an element or field describing the object e.g., colour), and an object (the value that the predicate takes on e.g., blue). A set of RDF triples is called an RDF graph. Let's see an example (Table 6) showing the representation of the Web site of the University of Burdwan by using DCMES as schema and RDF as framework.

Table 6: RDF Modeling of DCMES

Subject (Resource)	Predicate (Attribute/property)	Object (Value of attribute)
The University of Burdwan http://www.buruniv.ac.in	dc:title	The University of Burdwan site
	dc:creator	Sarkar, B.
	dc:subject	Academic Institute
	dc:descriptipon	The University established in the year 1960 under UGC Act……..
	dc:publisher	The University of Burdwan
	dc:contributor role=content writer	Central Library, BU
	dc:date	20060101
	dc:format	text/html
	dc:identifier	http://www.buruniv.ac.in
	dc:coverage	Education and Research
	dc:rights	The University of Burdwan

The encoding of DCMES in RDF structure on the basis of above data model may be entered as:

```
<?xml version="1.0"?>

<rdf: RDF xmlns: rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"

xmlns: dc = "http://purl.org/dc/elements/1.1/">

<rdf:Description about = "http://www.buruniv.ac.in">

<dc:title>Welcome to the Home page of the University of Burdwan </dc:title>

<dc:creator> Sarkar, B.</dc:creator>

<dc:subject>Academic Institute</dc:subject>

<dc:description>The University of Burdwan , a university under UGC… </dc:description>

<dc:publisher>The University of Burdwan</dc:publisher>

<dc:contributor role="content writer"> Central Library, the University of Burdwan
</dc:contributor>

<dc:date>20060101</dc:date>

<dc:format>text/html</dc:format>

<dc:identifier>"http://www.buruniv.ac.in"</dc:identifier>

<dc:language>en</dc:language>

<dc:coverage>Education and Research</dc:coverage>

<dc:rights> The University of Burdwan </dc:rights>

</rdf:Description>

</rdf:RDF>
```

Almost all the advanced level repository management software support RDF based encoding of DCMES. For example, a deposited record in EPrint archive software stores DC metadata elements in the following RDF format (the metadata of digital resource submitted to EPrint software can also be exported in RDF format).

<rdf:RDF>

<rdf:Description rdf:about="http://lfleprints/id/eprint/1">

<bibo:abstract rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Digital asset management</bibo:abstract>

<bibo:authorList rdf:resource="http://lfleprints/id/eprint/1#authors"/><dct:creator rdf:resource="http://lfleprints/id/person/ext-7d626d4d228010d5dfc24bb7589ce50e"/><dct:date>2011</dct:date><dct:isPartOf rdf:resource="http://lfleprints/id/repository"/><dct:issuer rdf:resource="http://lfleprints/id/org/ext-d18099436db85fd6524d7fed2a19663e"/><dct:issuer rdf:resource="http://lfleprints/id/org/ext-e48cac986ee5a73a0da34f517f02104e"/><dct:subject rdf:resource="http://lfleprints/id/subject/ZA4050"/><dct:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Digital archiving </dct:title><rdf:type rdf:resource="http://purl.org/ontology/bibo/Article"/><rdf:type rdf:resource="http://purl.org/ontology/bibo/Thesis"/><rdf:type rdf:resource="http://eprints.org/ontology/EPrint"/><rdf:type rdf:resource="http://eprints.org/ontology/ThesisEPrint"/><rdfs:seeAlso rdf:resource="http://lfleprints/1/"/></rdf:Description><rdf:Description rdf:about="http://lfleprints/id/subject/ZA4050"><rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/><skos:prefLabel xml:lang="en">ZA4050 Electronic information resources</skos:prefLabel></rdf:Description><rdf:Description rdf:about="http://lfleprints/id/org/ext-d18099436db85fd6524d7fed2a19663e"><dct:isPartOf rdf:resource="http://lfleprints/id/org/ext-e48cac986ee5a73a0da34f517f02104e"/><foaf:name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Library and Information Science, Kalyani university</foaf:name><rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Organization"/></rdf:Description><rdf:Description rdf:about="http://lfleprints/id/org/ext-e48cac986ee5a73a0da34f517f02104e"><dct:hasPart rdf:resource="http://lfleprints/id/org/ext-d18099436db85fd6524d7fed2a19663e"/><foaf:name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Kalyani university</foaf:name><rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Organization"/></rdf:Description><rdf:Description rdf:about="http://lfleprints/id/person/ext-7d626d4d228010d5dfc24bb7589ce50e"><foaf:familyName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Mukhopadhyay</foaf:familyName><foaf:givenName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Parthasarathi</foaf:givenName><foaf:name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Parthasarathi Mukhopadhyay</foaf:name><rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>

</rdf:Description>

</rdf:RDF>

1.7 APPLICATION OF METADATA IN OPEN ACCESS

Application guidelines for metadata encoding are required to mitigate the detrimental effects of divergent interpretation of the metadata standards that exist in the open access landscape. There are national level initiatives that provide required guidelines in interpretation of encoding rules, use of standards in encoding and rendering of metadata elements. These guidelines may help you in different aspects of metadata application in managing OA contents - to reduce ambiguity, to boost the extent to which metadata can be harvested efficiently, and to enhance the accuracy and value of services built on metadata harvesting. This section includes two components – first one is related to guidelines developed by different initiatives and the second one shows application of the guidelines in OA content organization.

1.7.1 Guidelines and Initiatives

Most of the guidelines (as developed in US and UK) advocate to categorize metadata elements into four categories - Required, Required if Applicable, Recommended and Optional. The basic purpose of the categorization is to identify the elements necessary for a user in a shared metadata environment. Guidelines are not format-specific; rather they identify those elements commonly needed across all formats. An analysis of existing suggestions and guidelines shows the following categorization of metadata elements -

Required

- Date Created or Date Published (dc:date)
- Identifier (dc:identifier)
- Institution Name (dc:publisher)
- Title (dc:title)
- Type of Resource (dc:type)

Required if Applicable

- Creator (dc:creator)
- Extent (dc:format)
- Language of Resource (dc:language)
- Related Item (dc:relation)

Recommended

- Description (dc:description)
- Access or Use Restrictions (dc:rights)
- Format of Resource (dc:format)
- Place of Origin (dc:coverage)
- Rights Information (dc:rights)

- Subject (dc:subject)

Optional

- Citation (dc:relation)
- Collection Name
- Contributor (dc:contributor)
- Genre (dc:type)
- Keywords or Tags (dc:subject)
- Language of Metadata Record (no dc map)
- Notes (dc:description)
- Publisher (dc:publisher)

Application of metadata to describe OA resources are guided by four principles that are independent of metadata schema – i) Content Standards for Metadata (to guide what information should be recorded when describing a particular type of resource and how that information should be recorded); ii) Data Value Standards for Metadata (to help to normalize data element sets to ensure consistency between records); iii) Structural Standards for Metadata (to guide in selecting fields or elements where the data resides; and iv) Syntax Standards for Metadata (to guide in encoding for data values so that they can be processed by different systems).

Content Standards for Metadata

Content Standards improve the ability to share metadata records and the discoverability of OA resources. Consistent description of metadata records helps users to understand and analyze search results efficiently. Metadata that is formatted inconsistently (e.g. names recorded both as "Last name, First name" and "First name / Last name") impacts indexing and sorting and users suffer with confusing or incomplete results. OA content management software adopted different levels of content

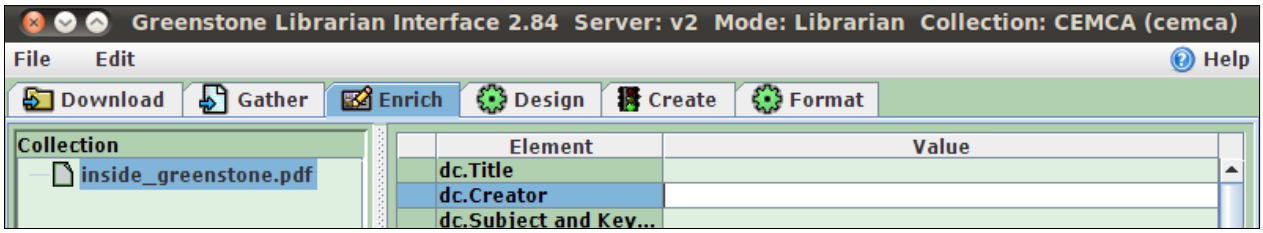

Figure 5: Content Standards for Metedata DC.Creator in Greenstone
(*Source*: Greenstone software)

standards in describing OA resources, for example, in Greenstone digital library software includes no content standards for encoding DC.Creator (Figure 5) whereas DSpace and EPrint provides scope for giving Last Name and First Name of creator separately. EPrint (Figure 6) also provides help button

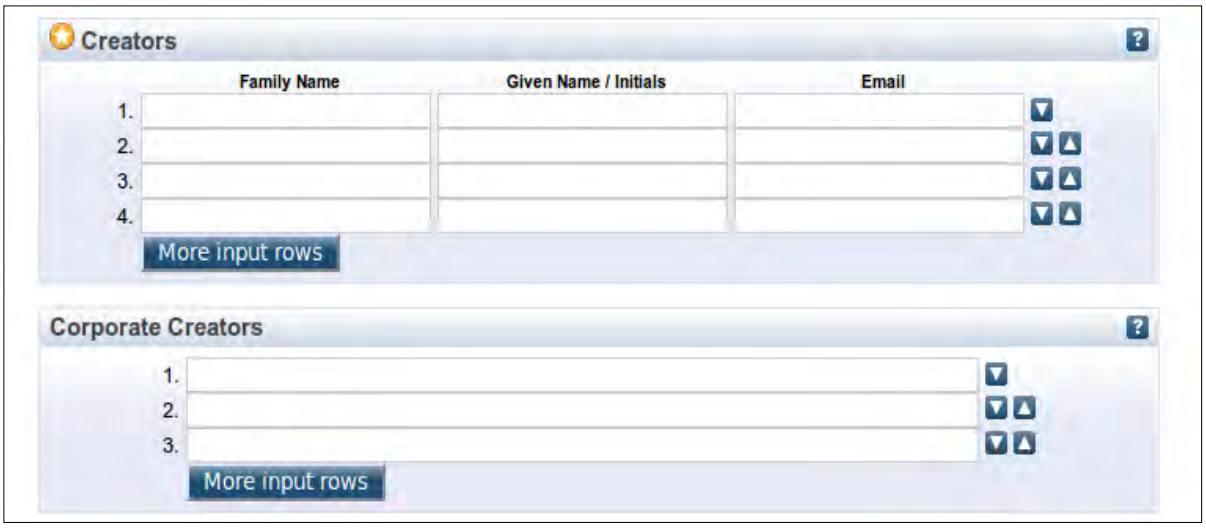

Figure 6: Content Standards for Metadata DC.Creator in Eprint
(*Source*: E-print software)

(? mark) to help submitters in encoding a particular metadata element or field. DSpace apart from maintaining contents standards provides examples and links to help file to support resource description (Figure 7).

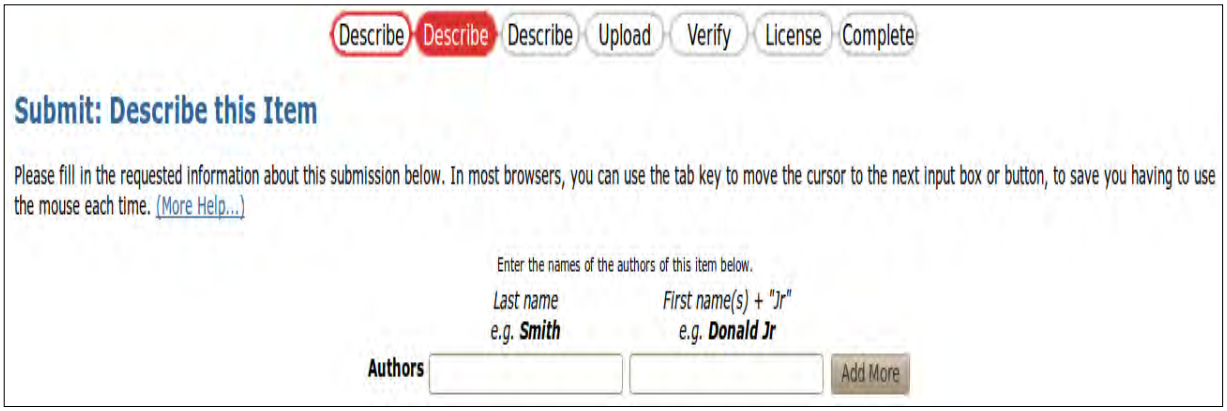

Figure 7: Content Standards for Metedata DC.Creator in DSpace
(Source: DSpace Software)

Library professionals apart, content standards provided in software may follow standards like Anglo-American Cataloguing Rules (AACR2) that covers description of different formats, and the provision of access points, Resource Description and Access (RDA) that guides content management by using FRBR principles (work/expression/manifestation/item), Cataloging Cultural Objects (CCO) that covers encoding of cultural heritage resources and Describing Archives for managing single and multi-level descriptions of archives, personal papers, and manuscripts etc.

Data Value Standards for Metadata

Standardization of data values are important for retrieval and sharing of OA contents. These standards aim to prescribe normalized list of terms to be used for certain data elements. It advocates use of controlled terms to ensure consistency and to achieve collocation of resources related to the same topic or person through the application of thesauri, controlled vocabularies, and authority files. The recommended data entry standardization tools are -

- Getty Art and Architecture Thesauri (AAT) is a structured vocabulary for terms used to describe art, architecture, decorative arts, material culture, and archival materials.
- Getty Thesaurus of Geographic Names (TGN) is a structured vocabulary for names and other information about places.
- Getty Union List of Artist Names (ULAN) is a structured vocabulary for names and other information about artists.
- Library of Congress Subject Headings (LCSH) comprises a thesaurus of subject headings, maintained by the United States Library of Congress.
- Library of Congress Name Authorities (LCNA) includes Corporate Names, Geographic Names, Conference Names, and Personal Names.
- Thesaurus of Graphic Materials I: Subject Terms (TGM-I) consists of terms and numerous cross references for the purpose of indexing topics shown or reflected in pictures.
- Thesaurus of Graphic Materials II (TGM-II) is a thesaurus of terms to describe Genre and Physical Characteristic Terms.

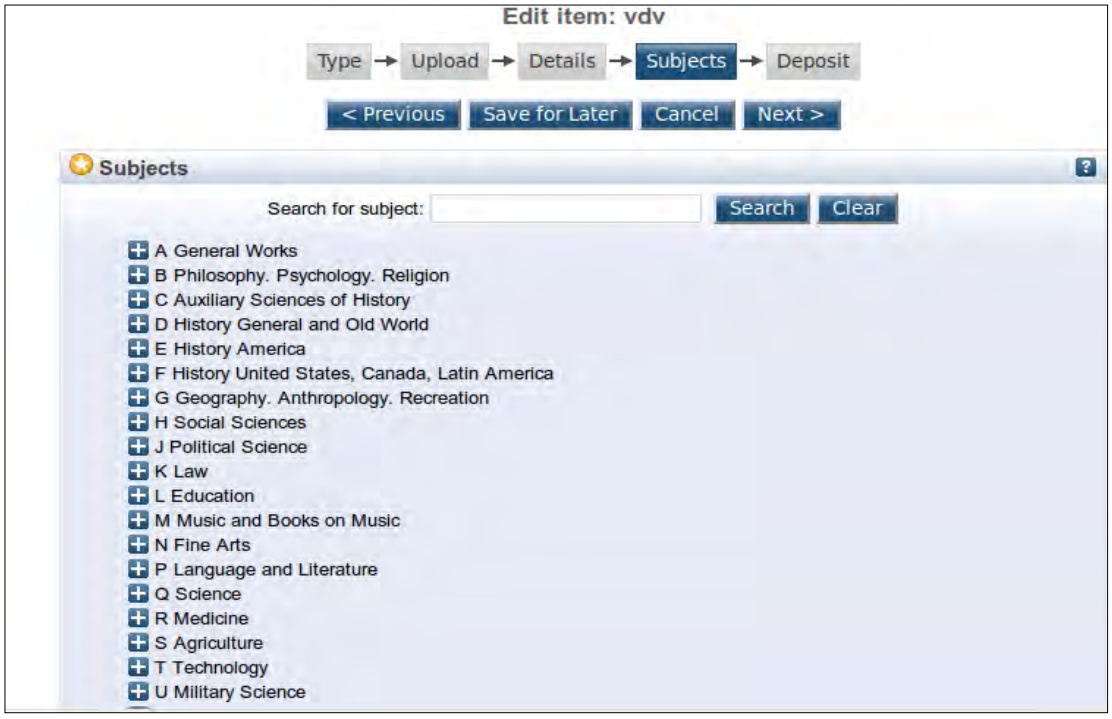

Figure 8: Content Standards for Metadata DC.Creator in e-Print
(*Source*: e-Print software)

Many OA repository software support data value standards, for example, e-Print software includes entire Library of Congress Subject Areas to support standard encoding of the field DC.Subject; DSpace includes research category list (although required to be activated through configuration file of DSpace) to help in populating DC.Subject field (see Figure8). These data standards are available to both cataloguer/indexer and searchers.

Structural Standards for Metadata

Metadata structure consists of elements for description of data. Structural standards define fields, scope of the fields and type of information that need to be stored (see Table 3 for DCMES). As a matter of rule it is always better to apply metadata structure that has a high level of granularity. The reason is simple – it is always easier to transfer metadata from granular structure to a more simple structure. In some cases Structural Standards mandate what Syntax Standards should be used (for example, W3C encoding rules for date and times[42] based on ISO 8601). Structural standards for generic and domain-specific schemas generally follow some broad principles such as - Fields/elements should be unambiguous; Fields/elements may be required; Some fields/elements may be repeatable; Some fields/elements may be mandatory; Some fields/elements may have unique value to identify record (e.g. use of DOI in DC.Identifier); and Some fields may have defined relationships with other fields, e.g. qualifiers or subfields. UK Metadata Guidelines for Open Access Repositories (2013) in its document entitled "Phase 1: Core Metadata (Version 0.9)" published in March 2013 prescribed following minimum fields/elements as structural standard for OA resources (M – Mandatory, R – Repeatable and O - Optional) (Figure 9):

Element	Inclusion
dc:title	M
dc:creator	M
dc:identifier	M
dc:source	M
dc:language	M
rioxxterms.projectid	M
rioxxterms.funder	M
dcterms:issued	M
dc:format	R
dc:publisher	R
dc:description	R
dc:subject	R
dc:rights	R
dc:coverage	O
dc:audience	O
dc:type	O
dc:contributor	O
dc:relation	O
dcterms:references	O

Figure 9: Core Metadata Inclusion Types

[42] http://www.w3.org/TR/NOTE-datetime

Interoperability and Retrieval

This standard mostly recommends simple DCMES for OA repositories with Qualified DC for two instances (dc terms: issued and dc terms: Relation). These sets of recommendation also include two new elements specific to OA resources – project ID (a unique identifier normally provided by the funder) and funder name. Most of the elements have namespace 'dc' and the two new elements have 'rioxxterms' namespace. This UK-specific Guideline is based on the Driver project, OpenAIRE Guidelines (OpenAIRE project[43]) and UKETD_DC (the metadata core set recommended by the British Library's Electronic Theses Online Service EthOS[44]). Please see section 1.5.3 for structural standards in different domains.

Syntax Standards for Metadata

These standards aim to make the metadata machine readable. Structural standards generally prescribe syntax standard(s) for fields/elements. In case structural standard does not advise syntax standard, library professionals should follow syntax that enable sharing of OA resources. Generally HTML, XML (Extensible Markup Language) and SGML (Standard Generalized Markup Language) are used as syntax standard for OA resources. UK Metadata Guidelines for Open Access Repositories (2013) recommended syntax standard for each metadata element listed in previous section. One example may be cited here for your understanding:

element:	dc:creator
status:	mandatory
scope:	The creator of a resource may be a person, organisation or service. Where there is more than one creator, use a separate dc:creator element for each one. Enter as many creators as required.
standard:	The dc:creator element should take an optional attribute called "id".
(data value)	This will hold a machine-readable unique identifier, where available, for the creator. Ideally the element will include a machine-readable id and a text string in the body of the element.
syntax:	<dc:creator id=http://"identifier-for-this-creator-entity">name-of-this-creator-entity</dc:creator>

Where the creator is a person, the recommended format is Last Name, First

Name(s) and to include an ORCID ID, if known, in its HTTP URI form, such as:

<dc:creator id=http://orcid.org/0000-0002-1395-3092>Mishra, Sanjay</dc:creator>

Note: If the creator is a person and you wish to record that person's affiliation, the affiliation should be recorded using the dc:contributor element.

[43] http://www.openaire.eu
[44] http://ethos.bl.uk/Home.do

You may consult UK Metadata Guidelines for Open Access Repositories (2013): Phase 1- Core Metadata (Version 0.9) from rioxx.net. Other related initiatives in this direction are given as below:

- **CrossMark**[45]: An initiative to support non-bibliographic metadata schema by CrossRef.
- **HowOpenIsIt?**: An initiative of PLOS, SPARC and OASPA to set criteria to measure openness (extent of rights for different stakeholders) and quality of OA resources[46].
- **Vocabularies for OA**[47] (V40A): An initiative of JISC/UKOLN to develop vocabulary control devices, category lists and authority files for OA resources.
- **RIOXX**[48]: Developing Repository Metadata Guidelines: An initiative to define a standard set of bibliographic metadata for UK Institutional Repositories.
- **ONIX-PL**[49]: An initiative to standardize license expression information necessary for OA publishing.
- **Linked Content Coalition**[50]: An initiative to develop rights management metadata for OA resources.
- **Open Discovery Initiative**[51]: A NISO initiative to develop library discovery services for non-commercial and OA resources through indexed search.
- **Incentives, Integration, and Mediation**: Sustainable Practices for Populating Repositories: An initiative of Confederation of Open Access Repositories (COAR[52]) to develop guidelines for populating OA repositories including guidance for metadata management.
- **NISO**[53] **Specification for Open Access Metadata and Indicators**: A NISO initiative to develop standard metadata set specifically meant for OA resources.
- **RSLP**[54]: A UKOLN initiative for Collection Level Descriptions (CLDs) as a tool for providing an overview of the content and coverage of OA collections.

[45] http://www.crossref.org/crossmark/
[46] http://www.plos.org/wp-content/uploads/2012/10/OAS_English_web.pdf
[47] http://www.jisc.ac.uk/whatwedo/topics/digitallibraries/pals-group/v4oa.aspx
[48] http://rioxx.net/
[49] http://www.editeur.org/21/ONIX-PL
[50] http://www.linkedcontentcoalition.org/
[51] http://www.niso.org/workrooms/odi/
[52] http://coar-repositories.org
[53] http://www.niso.org/home/
[54] http://www.ukoln.ac.uk/metadata/rslp/schema/

1.7.2 Software-level applications

Most of the repository management software (such as Greenstone, DSpace, ePrint) include predefined standard metadata schemas. For example, Greenstone includes simple DCMES, qualified DCMES, AGLS, nzgls and dls schemas (see Figure 10). Collection developer may use any one of them at the time of data entry activities. DSpace comes with only DCMES but allows customizing submission interface to include domain-specific metadata schemas. ePrint is more sophisticated in metadata handling in comparison with other OA content management software. Initiatives are also supporting software in managing metadata in standard manner. For example, UK Metadata Guidelines for Open Access Repositories, supported by UKOLN, JISC and RCUK developed a plug-in for ePrints repositories (versions 3.3.x) and a patch for DSpace repositories (version 1.8.2; version 3.x onwards) for management of content standards, data value standards, structural standards, and syntax standards of metadata. These patches are available as open source scripts and can easily be integrated with the target

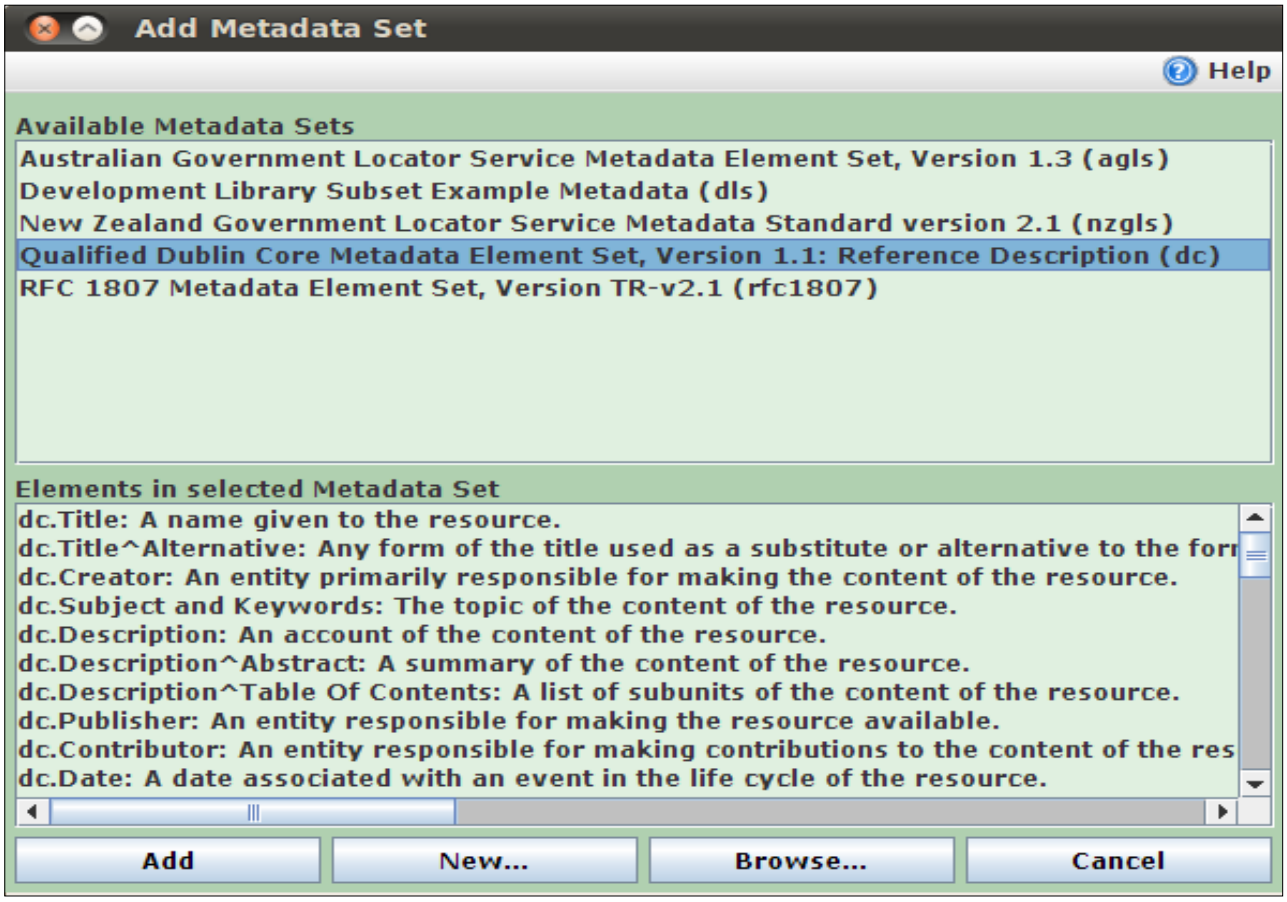

Figure 10: Available Schemas in Greenstone
(Source: Greenstone software)

DSpace has a metadata registry with all data elements of DCMES in qualified format. It allows repository manager to add, edit, refine and delete metadata element (Figure 11).

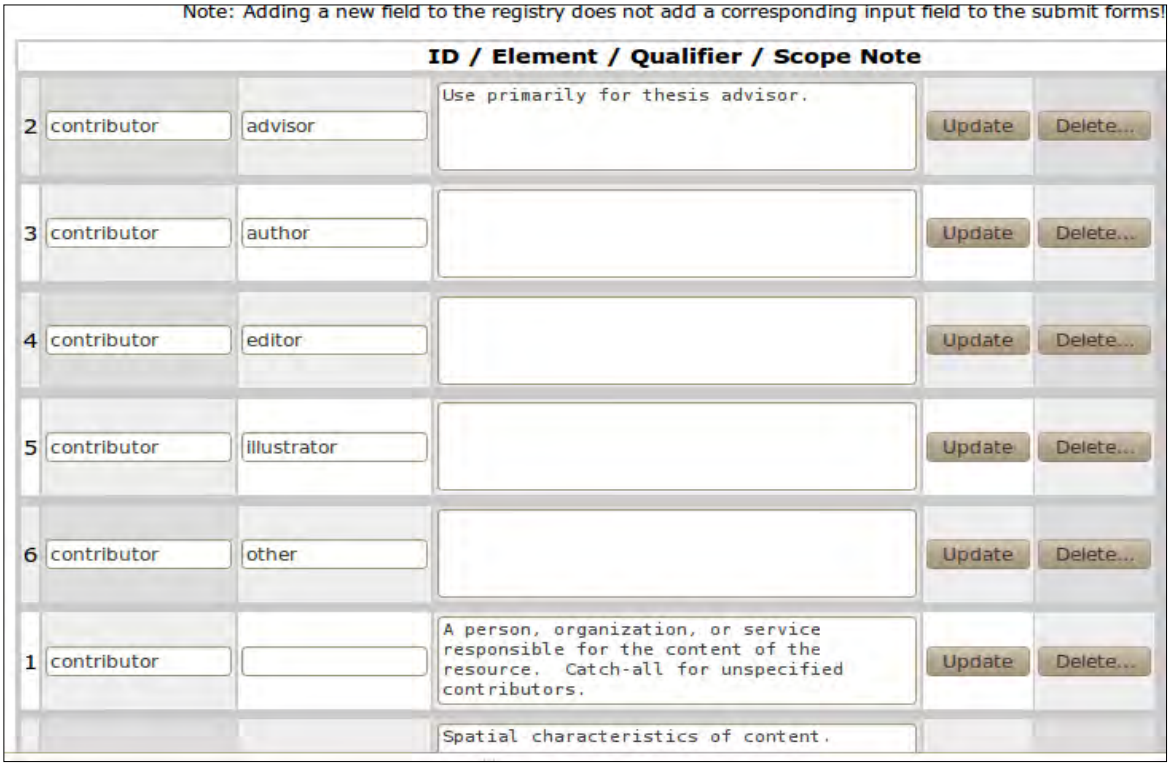

Figure 11: Metadata Registry in DSpace (Source: DSpace software)

DSpace uses a qualified version of the Dublin Core schema based on the Dublin Core Libraries Working Group Application Profile[55] (LAP). EPrint software provides six metadata sets related with OA knowledge objects, OA metadata, users of OA, search related metadata, and import related metadata and metadata for bit streams (files) (see Figure 12). As a whole the metadata management component of ePrint is a smart solution in view of different requirements of OA content management such as usage data, file format data etc.

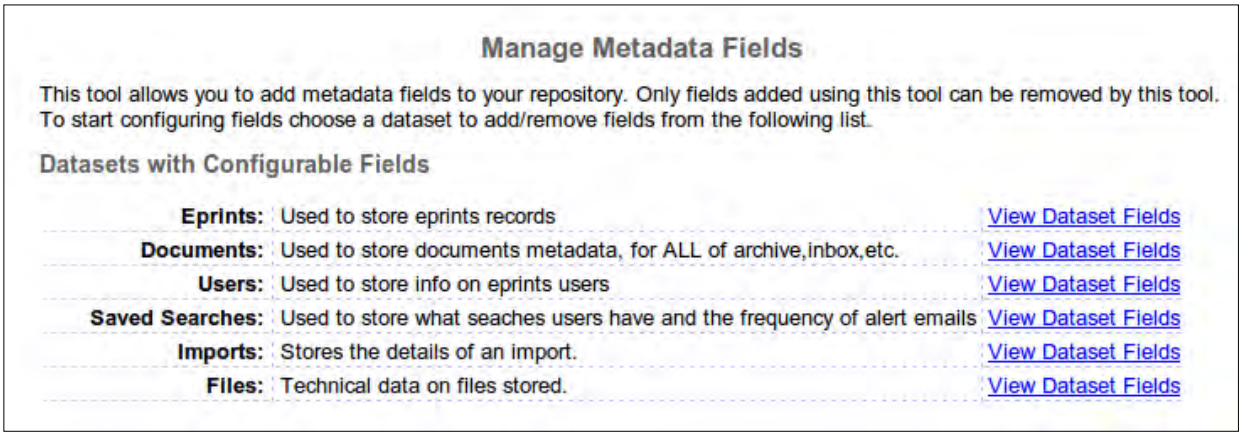

Figure12: Metadata Registry in EPrint (Source: E-print software)

For example, the bit stream of file metadata in ePrint is more comprehensive in comparison with other open source OA repository management software.

[55] http://dublincore.org/documents/library-application-profile/

copies:	["file_fieldname_copies" not defined]	Core Field
data:	["file_fieldname_data" not defined]	Core Field
datasetid:	Object dataset id	Core Field
fileid:	Unique file id	Core Field
filename:	File name	Core Field
filesize:	File size	Core Field
hash:	File checksum	Core Field
hash_type:	Checksum type	Core Field
mime_type:	Mime-Type	Core Field
mtime:	File modification time	Core Field
objectid:	Object id	Core Field
url:	["file_fieldname_url" not defined]	Core Field

1.7.3 Authority Control in Gold OA and Green OA

As a library professional you know the importance of authority files such as name authority, title authority and subject authority. These authority files are required for collocation of data values entered against DC.Creator, DC.Contributor, DC.Subject etc. In the library world VIAF (Virtual Internet Authority File) is available as a huge name authority file. It aggregates name authority data from 25 national libraries. OCLC made available VIAF as Linked Open Data (LOD). It means that this dataset can be linked dynamically with the DC.Creator metadata field in different repository software. Apart from traditional name authority files like LC Name Authority File (NAF), LC Subject Authority File (SAF), VIAF etc, there are some emerging standards for populating name fields in controlled manner such as AuthorClaim[56], LATTES[57], NARCIS[58], ArXiv[59] Author ID, Names Project[60], Researcher ID[61], ORCID[62] etc. The details of all these standards for controlled data value standards are discussed at length in Unit 2 (section 2.3.5). In case of subject authority, most of the OA repository management software is applying standard controlled vocabulary

[56] http://authorclaim.org
[57] http://lattes.cnpq.br/
[58] http://www.narcis.nl
[59] http://www.arxiv.org
[60] http://names.mimas.ac.uk/
[61] http://www.researcherid.com
[62] http://www.orcid.org/

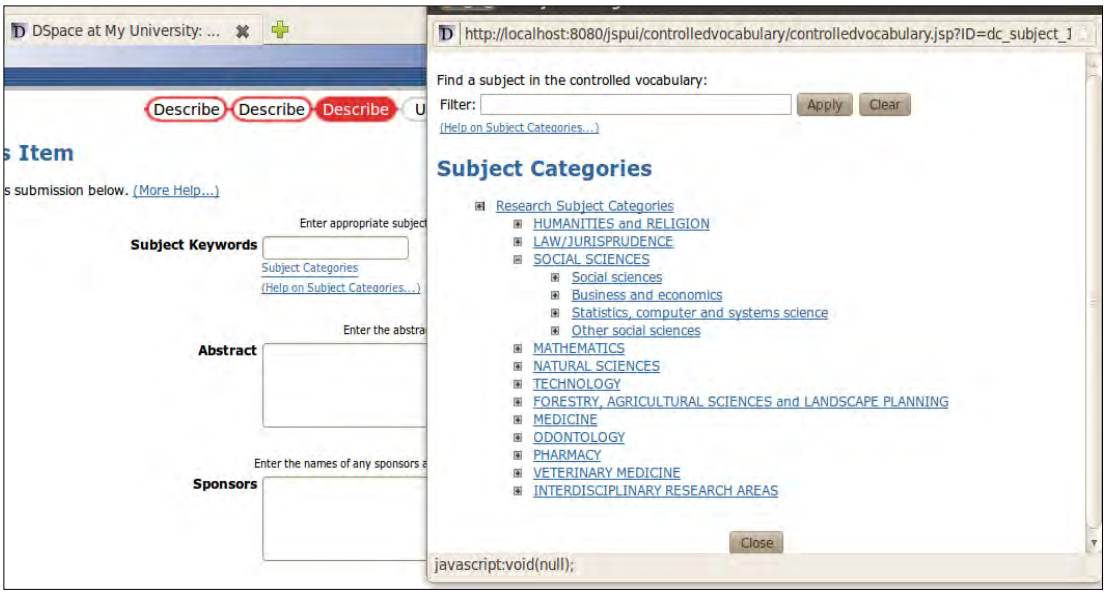

Figure 13: Subject Authority Control in DSpace (Source: DSpace software)

Devices, such as ePrint is using LC subject categories, DSpace is using research subject categories by default (Figure 13) but allows inclusion of any standard subject category list such as Dewey Decimal Classification (see Unit 3 section 3.5), if formatted in SKOS (Simple Knowledge Organization System – a W3C standard).

1.8 METADATA: CROSSWALKS AND INTEROPERABILITY STANDARDS

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 functionality. There are two approaches to interoperability—cross-system search and metadata harvesting. The Z39.50 protocol is commonly used for cross-collection search. The Z39.50 client (called origin) maps search syntaxes to a common set of search attributes for extracting information from Z39.50 server (called target). Open Archives Initiative[63] is a protocol for metadata harvesting, which allows all partners to translate their native metadata to a common core set of elements and expose those for harvesting. A search service then gathers the metadata sets into a central index to allow cross-repository searching regardless of the metadata formats used by participating repositories. Metadata crosswalks facilitate the interoperability and exchange of metadata. A crosswalk is a mapping of the elements, semantics and syntax from one metadata schema to those of another. It allows metadata created by one community to be used by another group that employs a different metadata standard. The Library of Congress' Network Development and MARC Standards Office is developing a framework for working with MARC data in a XML environment. This framework is intended to be flexible and extensible to allow users to work with MARC data in ways

[63] http://www.openarchives.org

specific to their needs. The framework will contain many components such as schemas, style sheets, and software tools developed and maintained by the Library of Congress. MARC-XML could potentially be used for representing a complete MARC record in XML, as an extension schema to METS (Metadata Encoding and Transmission Standard), to represent metadata for OAI harvesting, for original resource description in XML syntax and for metadata in XML that may be packaged with an electronic resource. A crosswalk mapping of Dublin Core, MARC 21 and Z 39.50 attributes is illustrated here to make it clear to you.

Table 7: Crosswalk of DCMES and MARC 21 Bibliographic data format

Sl. No	Z39.50 USE Attributes		Dublin Core Elements	MARC21 Fields
	Name	Value		
1	DC-Title	1097	Title	245 $a
2	DC-Contributor	1098	Creator	100, 110, 111, 700, 710, 711 and 720
3	DC-Subject	1099	Subject	600, 610, 611, 630, 650, 653
4	DC-Description	1100	Description	500 –599 excluding 506, 530, 540, 546
5	DC-Publisher	1101	Publisher	260 $a and 260 $b
6	DC-OtherContributor	1106	Contributor	--
7	DC-Date	1102	Date.issued	260 $c
8	DC-ResourceType	1103	Type	655
9	DC-Format	1107	Format	856 $q
10	DC-ResourceIdentifier	1104	Identifier	856 $u
11	DC-SourceIdentifier	1108	Source	786 $o $t
12	DC-Language	1105	Language	008/35-37, 041, 546
13	DC-Relation	1109	Relation	530, 760-787 $o $t
14	DC-Coverage	1110	Coverage	651, 752
15	DC-RightsManagement	1111	Rights	506, 540

Resource description area is dominated by NISO standards, ISO standards and standards developed by Library of Congress. The major standards developed by NISO and ISO are -

- Z39.91 (Stage Collection Description Specification)
- Z39.92 (Stage Information Retrieval Service Description Specification)
- Z39.85 (The Dublin Core Metadata Element Set)
- Z39.86 (Specifications for the Digital Talking Book)
- Z39.87 (Data Dictionary - Technical Metadata for Digital Still Images)
- ISO 15836:2003(Information and documentation - The Dublin Core metadata element set)

- ISO 17933:2000 (GEDI -- Generic Electronic Document Interchange)

In the area of metadata interoperability ISO 2709 and ANSI/NISO Z 39.2 (Information Interchange Format) standards played a historic role and till date ISO-2709 is considered as a mandatory standard in the library world for import and export of cataloguing data.

The standards developed by Library of Congress are also considered as important milestones in resource description area such as

- **MODS** (Metadata Object Description Standard[64]) - XML markup for selected metadata from existing MARC 21 records as well as original resource description (developed by Library of Congress)
- **MADS** (Metadata Authority Description Standard[65]) - XML markup for selected authority data from MARC21 records as well as original authority data (developed by Library of Congress)
- **METS** (Metadata Encoding andTransmission Standard[66]) - Structure for encoding descriptive, administrative, and structural metadata (developed by Library of Congress)

METS (Metadata Encoding and Transmission Standard) is a standard for encoding metadata within prescribed format. It contains descriptive and administrative elements of its own, but a major purpose of METS is to provide structure to accommodate other metadata schemas for exchange or delivery. It ensures that different digital objects can be described and linked together in a record to develop structure of description for complex digital resources (for example different components of a digitized book such as bibliographic data, images, transcribed text, maps etc may be described by using generic and domain-specific metadata schemas and finally packed in METS format). METS has its origin in a large scale digitization project Making of America II (MOA2), sponsored by the US Digital Library Federation. It is now maintained by the US Library of Congress (METS version 1.1 was released in 2001; the current version is 1.5, released in 2005). A MODS (Metadata Object Description Standard), on the other hand, is a metadata schema for encoding information about library resources (particularly books). It is based on a subset of the MARC 21 standard (MARC 21 Bibliographic Format) and it expresses MARC-compatible metadata format in XML and by using language-based element names. The first (draft) version 1.0 was released in 2002. The current version is 3.2, published in 2006. It is maintained by the US Library of Congress and is often used with METS as a descriptive metadata structure standard which means that METS acts as metadata wrappers and MODS acts as metadata schema for storage or exchange of digital objects. MADS (Metadata Authority Description Standard) is a companion schema to MODS. It is based on MARC 21 Authority format. It aims to support expressing authority format in XML. It deals with headings and cross references ('see' and 'see also' in the library community), including name authority (personal

[64] http://www.loc.gov/standards/mods
[65] http://www.loc.gov/standards/mads
[66] http://www.loc.gov/mets

names, corporate names, name/title entries), title authority (title entries, uniform titles), and subject authority (subject, genres, and geographic places). MODS and MADS are often used in harmony to describe bibliographic and authority datasets in XML and METS provides a metadata wrapper to store, deliver and sharing of resource description datasets.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

9) What is the role of Data Value Standards in OA metadata encoding?

...……..……..……...

……..……..……...

……..……..……...

10) Explain relation of MODS, MADS and METS.

...……..……..……...

……..……..……...

……..……..……...

1.9 LET US SUM UP

This unit starts with introducing you with the definition, types and importance of metadata and ends with metadata interoperability. The journey to these two ends covers many important issues related to application of metadata in organizing OA resources such as, importance of metadata in OA infrastructure, rights management metadata, policy framework for OA metadata, elements of usage metadata, and application of generic and domain specific metadata schemas. It shows you application of different models in OA metadata including RDF/XML framework, IFLA models and XOBIS. This unit also provides a comprehensive list of metadata schemas in different areas of human activities that are important for OA movement *per se* such as ETDs, learning objects and multimedia. The software-level application of metadata in organizing OA resources may help you in solving practical problems related to metadata encoding. Further, the section on crosswalk and interoperability aims to provide you an insight on export, import and sharing of metadata in greater OA infrastructure.

UNIT 2 INTEROPERABILITY ISSUES FOR OPEN ACCESS

Structure

2.0 INTRODUCTION

Due to development and availability of repositories in different domains, it is becoming difficult for end users to search those repositories comprehensively that provide scholarly materials freely as they need to move from one repository to another that calls for the need to learn retrieval techniques and search operators in use in different repository software systems. This situation calls for the development of a single window search service covering all the repositories in a given domain of knowledge. The above scenario is tried to be solved by metadata integration. Similarity, other areas of scholarship like unique identifiers for resources and contributors, exchange of complex digital resources, sharing of data related to usage of open access resources etc require exchange frameworks and standards and to achieve this potential, we need interoperability.

As the situation stands now, the interoperability landscape is presently chaotic and complex as initial dust settlement phase is going on now. This unit attempts to provide you an overview on i) needs, requirements, types and technical issues related to interoperability in open access contents

dissemination; ii) present interoperability initiatives; iii) metadata interoperability and harvesting and iv) trends and future possibilities in interoperability.

2.1 LEARNING OUTCOMES

After working through this unit, you are expected to be able to:

- Assess the need of interoperability in developing open access infrastructure at global scale;
- Understand different areas of interoperability and related interoperability standards;
- Critically examine the technical issues and initiatives related to achieving interoperability;
- Apply metadata harvesting software to develop single-window search interface; and
- Realize the trends and future course of development in interoperability.

2.2 INTEROPERABILITY

Open access resources are very important elements in creating information support system for creating global research and development infrastructure and availability of resources. Open access repositories, the green path of open-access, are playing a significant role in creating world-wide e-Research framework but the real value of repositories lies in their ability to be integrated with existing resources for providing a single-window search interface for end users. For example, OpenDOAR lists a total of 156 repositories on *Physics* and these repositories are different in their coverage, software usage, nature of contents and most importantly in retrieval techniques and tools.

To have access to these repositories what is needed is to develop a mechanism in the form of interoperability to facilitate search with single window search system. Interoperability means the ability of multiple systems (with different hardware and software platform, data structure, and user interface) to exchange data with minimal loss of content functionality. In bibliographic domain, interoperability is supported by Crosswalk. A crosswalk is a mapping of the elements, semantics and syntax from one metadata schema to those of another. It allows metadata created by one community to be used by another group that employs a different metadata standard. Interoperability and crosswalk ensures exchange of bibliographic data and contents amongst heterogeneous open contents systems across the globe. Open contents retrieval systems can achieve interoperability by following guidelines for setting up repositories, and by applying relevant protocols and interoperability standards.

Integration of open access resources available from repositories distributed across the globe is the need of the time for the success of open access philosophy. Interoperability is the magic wand that makes this integration

possible. In other words, interoperability helps to achieve the goal of open access movement – to increase access, visibility, and impact of publicly funded research activities. Interoperability helps end users to locate required information resources from a unified search interface without knowing location of objects and repository specific retrieval techniques. The success of green path of open access i.e. dissemination of open contents through institutional and subject based repositories directly depends on interoperability. The gold path of open access i.e. open access journals may also be benefited through integration of usage data, citation data, article-level metrics etc on the basis of interoperability.

2.2.1 Types of Interoperability

The IEEE Glossary defines interoperability as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" (Geraci, 1991). Interoperability, in broader sense, is the ability for systems (including information systems) to communicate with each other and pass information back and forth in a usable format. In open access information systems, interoperability may help us in contents aggregation, data mining, and on-the-fly integration of related resources from different locations in real time, improvement of existing information services and introduction of new information services. Interoperability may fundamentally be grouped into two categories – i) Syntactic interoperability; and ii) Semantic interoperability.

In syntactic interoperability, two different systems communicate and exchange data on the basis of standard data formats (e.g. MARC 21 or Dublin Core), standard exchange format (e.g. ISO-2709 or MARC-XML), text-encoding standards (e.g. ASCII, ISCII or Unicode) and communication protocol (e.g. Z 39.50 or OAI/PMH). Semantic interoperability, on the other hand, supports automatic interpretation of information elements on the basis of common information exchange reference model (e.g. integration of two different thesaurus or classification schemes; conversion of bibliographic data available in CCF format into MARC formats on the basis of Crosswalks). However, in the open access domain, COAR (Confederation of Open Access Repositories) identified following major areas of interoperability:

Metadata level interoperability: It refers to integration of metadata from different open access resources into a single-window service on the basis of metadata harvesting protocols and standards like OAI/PMH version 2.0 protocol. This helps to develop subject-specific portals and specialized search engines such as OAIster and BASE (Bielefeld Academic Search Engine).

Content level interoperability: This refers to the facilities of multiple-deposit process where authors submit document in one place and automatically contents are transferred from one system to another. This cross-system contents transfer is supported by protocols like SWORD (Simple Web-service Offering Repository Deposit) for multiple deposit and OA-RJ (Open Access Repository Junction) for managing multi-authored and multi-institutional open knowledge objects. Multiple deposit means simultaneous submission into multiple repositories – author's own institutional repository (IR), co-authors'

IRs, subject specific repositories, and funder repositories. CRIS-OAR (Current Research Information and Open Access Repositories), on the other hand, aims to support integration of research administration and open access repositories at the institutional level.

Network level interoperability: This supports development of national and regional repository networks on the basis of metadata harvesting. But global *de facto* standard for metadata harvesting OAI/PMH version 2.0 supports only unqualified Dublin Core metadata. Network level interoperability initiatives aims to layer some essential additional fields (may vary from network to network) on top of OAI/PMH. The DRIVER (Digital Repository Infrastructure Vision for European Research) project of European repository community first applied this model of interoperability which was later followed by OpenAIRE (Open Access Infrastructure Research for Europe) project.

Statistics and usage data level interoperability: Interoperability in usage statistics is emerging as an important area in open access domain. It allows measuring impact of individual open knowledge objects (e.g. research articles) and supports aggregation and exchange of usage information from different repositories and information systems (like CiteSeer). Many protocols and standards are being developed in the area of cross-repository usage statistics like SURE (Statistics on the Usage of Repositories) and PIRUS (Publishers and Institutional Repository Usage Statistics).

Identifier level interoperability: As a library professional, you are aware of the importance of authority data to support collocation of library documents. The same concept is also required for effective organization of open access resources. Like name authority, title authority and subject authority, we need consistency in identification and naming of authors, items, location of items, institutions, funding agencies, grants etc in organizing open access resources. Different standards and systems for unique author identification (e.g. ORCID and AuthorClaim), object identification (e.g. DOI, Handle system, PersID) and dataset identification (e.g. DataCite) are emerging standards and services to support this area of open access interoperability.

Object level interoperability: Open access resources are increasingly becoming multimedia objects. These include different media types (text, audio, video, streaming video etc) and are called compound digital objects. These resources require standards of interoperability for exchange of web resource aggregations. OAI-ORE (Open Archive Initiative – Object Reuse and Exchange) is considered as the *de facto* global interoperability standard in this area.

Semantic level of interoperability: This refers to meaningful exchange of data at machine-level. A standard such as the Resource Description Framework (RDF) is applied to achieve semantic interoperability in digital domain. RDF, as a greater metadata architecture, helps to express digital objects relationships in a machine understandable way. RDF-enabled open access information systems allows machines to create sophisticated services

through integrating knowledge objects distributed across repositories and other systems.

2.2.2 Technical Issues

Open knowledge objects are distributed globally in different open access journals, open access repositories and open datasets. Several digital asset management software including repository management and journal management software, both from the open source domain and commercial domain, appeared in the last ten years or so. Unfortunately, these software developed independently from each other with less emphasis on technologies that support system-level sharing and exchange of digital assets. In view of the panoramic and distributed nature of open access resources, heterogeneity is expected to be the norm. Interoperability is crucial to accommodate intra-system and inter-system data exchange and thereby requires a model to identify essential concepts, axioms and relationships which are independent of specific standards, technologies or implementations. The DL.org project[67] has identified six major areas of interoperability (independent of software, systems and standards) on the basis of The DELOS Digital Library Reference Model.

Architecture

The Reference model identified two major technical components to achieve architecture level interoperability – i) component profile and ii) application framework. The first one prescribes that each architectural component must be associated with a profile to describe functionality of the software component. A comprehensive component profile increases possibility of re-using the component by different software systems for different context. This facility also allows other systems to select and integrate software component into its workflow. The application framework prescribes that seamless exchange of information requires standardization of component roles, component-to-component interaction pattern, and component interaction interfaces. There are two major issues with architectural level interoperability[68] – content storage (components related to storage of digital knowledge objects) and content access (components deal with access to digital knowledge objects including parts and relations).

Contents

Contents are key resources for digital knowledge management system including open access information systems. The content management workflow (i.e. selecting, digitizing, describing, and digitally curating content resources) is labor-intensive, time-consuming and expensive. Therefore content level interoperability is an important issue in open access domain. The Reference model[69] prescribes standardization in following five sectors – i) Information object format (refers to data types to describe the structural properties of digital object); ii) Information object attributes (metadata that

[67] http://www.dlorg.eu/
[68] https://workinggroups.wiki.dlorg.eu/index.php/Architecture_Working_Group
[69] https://workinggroups.wiki.dlorg.eu/index.php/Content_Working_Group

describes resources must be comprehensive, structured and granular); iii) Information object context (metadata elements that records the relations with other entities like people, places, moments, time and semantics); iv) Information object provenance (metadata elements that records the process causing the object to be in its current state); v) Information object identifier (standards that uniquely identify and universally refer to the same information object).

Functionality

The technical issue related with the functionality[70] refers to all the processing aspects that can occur on resources and activities that can be observed by stakeholders of open digital content management. The Reference model prescribes – i) precise description of functions of each software modules; ii) recording of complementary and mutually dependent functions; iii) re-using of software modules that implement the desired functionality; iv) detailing of functionality profile of a digital assets management system, a digital asset management software and a digital asset management software module along with the associated interfaces.

Policy

The model refers to policy interoperability and policy classification. The policy level[71] interoperability helps to achieve integration with third-party service providers, such as data archives and cloud providers. It prescribes standards for – i) encoding of policies for machine discovery (languages of representation); ii) policy management (policy are appraisal and enforcement); iii) evolution of policies over time; and iv) relation between policy and quality.

Quality

This refers to the three most important elements of digital asset management system - quality of contents, quality of services and quality of policies. It aims to investigate interoperability issues that prevent software of the domain from working together from the perspective of quality. Finally, it aims to develop a quality framework[72] to support exchange of knowledge objects to achieve the goal of unified resource discovery.

Users

This refers to the Actor of digital asset management system and deals with issues such as user modeling, user profiling, user context, and user management. Till date there is no generally accepted user model that can be used in every software that supports green and gold path of open access. The Reference model identified two areas[73] of user level interoperability – i) interoperability of user profile from system to system; and ii) interoperability of usage pattern across the systems.

[70] https://workinggroups.wiki.dlorg.eu/index.php/Functionality_Working_Group
[71] https://workinggroups.wiki.dlorg.eu/index.php/Policy_Working_Group
[72] https://workinggroups.wiki.dlorg.eu/index.php/Quality_Working_Group
[73] https://workinggroups.wiki.dlorg.eu/index.php/User_Working_Group

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

1) What are the major areas of interoperability in open access domain?

...……..……..……...

……..……..……...

……..……..……...

2) Identify the factors that are responsible for architecture-level interoperability.

...……..……..……...

……..……..……...

……..……..……...

2.3 INTEROPERABILITY INITIATIVES

Open access resources are increasing steadily right from the first decade of 21st century. Recently BASE (an exclusive search engine for open contents) is reported to achieve indexing of 52 million open access resources. In this distributed, growing and complex open-access information environment interoperability holds the key for effective dissemination of open knowledge objects. We have already discussed different types of interoperability in previous section. There are seven levels of interoperability in the domain of open access resources as prescribed by COAR and DL.org. The semantic interoperability in open access domain is still in research bed. Therefore, in this section, we are going to study different interoperability initiatives under six major heads.

2.3.1 Metadata-level Interoperability Initiatives

Metadata interoperability is possibly the most visible initiative in the open access domain. Almost all open access repositories support metadata harvesting. It means sharing of metadata across an array of open access repositories. OAI/PMH is presently the only standard available in this direction. Open Archives Initiative – Protocol for Metadata Harvesting (OAI/PMH) is a low barrier and low-cost mechanism for harvesting metadata records from 'data providers' to 'service providers'. It works on the basis of Six Verbs (see section 4.2.4.2). OAI/PMH has its root in open access movement initiated by the establishment of eprint archives (arXiv, CogPrints, NACA (NASA), RePEc, NDLTD, NCSTRL) and developed by Open Archive Initiative. The present release of the protocol is OAI/PMH Version 2.0. There are many global open access services working on the basis of OAI/PMH

protocol such as OAIster[74], Europeana[75] and Connecting-Africa[76]. Almost all the open source repository management software like DSpace, Eprint, Fedora and Greenstone are compliant with OAI/PMH Version 2.0 and allows service providers to harvest metadata of deposited items in these software.

2.3.2 Content-level Interoperability Initiatives

Cross-system contents transfer aims – i) to manage multi-deposit; ii) to handle multi-authored and multi-institutional knowledge objects; and iii) to integrate digital knowledge archive and research administration. There are three major interoperability initiatives in this direction - SWORD, OA-RJ and CRIS-OAR respectively.

SWORD (Simple Web-service Offering Repository Deposit)

SWORD is a lightweight protocol to facilitate multiple deposits. It helps authors/submitters to deposit knowledge object into multiple repositories in one go. SWORD was first developed in 2007 under the sponsorship of JISC, UKOLN. It is based on AtomPub standard to achieve interoperability. This protocol facilitates transfer of metadata, and metadata plus digital objects (including compound digital objects). It supports content transfer for different combinations like Publisher to Repository, User's machine to Repository, Repository to Repository, Conference management system to Repository. It also supports repository bulk ingest and collaborative authoring.

URL: http://swordapp.org

Present version: Version 2.0

Implementation: DSpace, Fedora, EPrints (recent versions only)

Documentation: http://swordapp.org/the-sword-course

OA-RJ (Open Access Repository Junction)

OA-RJ is a protocol to support automatic deposition of multi-authored and multi-institutional knowledge objects into multiple repositories (both Institution-specific and Subject-specific repositories). OA-RJ aims to reduce the problems related to simultaneous submission into multiple repositories – author's own institutional repository (IR), co-authors' IRs, subject specific repositories, and funder repositories. It uses the ORI (Organization and Repository Identification) to achieve interoperability in workflows. OA-RJ helps submitters to refer and redirect to appropriate repositories through the use of API.

URL: http://edina.ac.uk/projects/oa-rj/index.html

Present version: Version 1.0

Sponsor: EDINA, JISC

Documentation: http://edina.ac.uk/projects/oa-rj/about.html

[74] http://oaister.worldcat.org
[75] http://www.europeana.eu/portal
[76] http://www.connecting-africa.net

CRIS-OAR (Current Research Information and Open Access Repositories)

The aim of this interoperability initiative is to define a metadata exchange format for integrating research information system and open access institutional repository with the help of an associated common vocabulary system. It transfers metadata of publications automatically from research information system to institutional repository with option (from authors) to integrate full-text resources.

URL: http://www.knowledge-exchange.info/

Present version: Version 1.0

Sponsor: Knowledge Exchange

Documentation: https://infoshare.dtv.dk/twiki/bin/view/KeCrisOar/WebHome

2.3.3 Network-level Interoperability Initiatives

This group of interoperability initiatives is dedicated to develop coordinated network of digital repositories at the national and regional levels. These initiatives aim to achieve high degree of interoperability and enhanced services to end users of open access resources. There are three major initiatives in this direction namely DRIVER, OpenAIRE and UK RepositoryNet+.

DRIVER (Digital Repository Infrastructure Vision for European Research)

DRIVER aims to create an infrastructure for open-access repositories in Europe. It provides a set of best practice guidelines (known as DRIVER guidelines) for content provider to build pan-European research infrastructure. The guidelines include – i) local data management policies; ii) OAI/PMH applications; iii) value-added services for repositories; iv) essential standards and processes for standardization; and v) development of D-NET v 1.0 toolkit to setup national repositories.

URL: http://www.driver-community.eu/

Present version: Version 2.0 of DRIVER Guidelines

Sponsor: DRIVER Consortium, EC

Documentation: http://www.driver-support.eu/documents

Implementation: Belgium DRIVER Portal, RCAAP – Portuguese portal, Recolecta – Spanish portal

OpenAIRE (Open Access Infrastructure Research for Europe)

OpenAIRE initiative is based on DRIVER guidelines. It provides guidelines and standards to integrate OA repositories and OA journals by following FP7 OA policy and ERC Guidelines for OA. This initiative also prescribes the metadata format to manage usage data (including events and statistics) and

also describes appropriate transfer protocols for the purpose (with emphasis on usage data of scholarly literature related to EC funded projects).

URL: http://www.openaire.eu/

Present version: Version 2.2, 2010

Sponsor: OpenAIRE Consortium, European Commission

Documentation: http://www.openaire.eu/en/support/guides/repository-managers

Implementation: ORBi – Open Repository and Bibliography (http://orbi.ulg.ac.be/)

KOPS - IR of Universität Konstanz (http://kops.ub.uni-konstanz.de/)

UK RepositoryNet+

UK RepositoryNet+, also known as RepNet, is an initiative to support open access interoperability through socio-technical infrastructure to support deposition, curation and content exposure. RepNet concentrates on four areas – Deposit (application of OA-RJ and ORI); Policies (brings together SHERPA RoMEO and JULIET); Reporting (use of IRUS-UK to create COUNTER-compliant repository usage at item level); and Innovation.

URL: http://www.repositorynet.ac.uk/

Present version: Version 2.0 of DRIVER Guidelines

Sponsor: DRIVER Consortium, EC

Documentation: http://www.driver-support.eu/documents

DINI Certificate for Document and Publication Services

Although this initiative is not a technical specification for OA network level interoperability but it provides a comprehensive socio-legal guideline in maintaining OA repositories. The German Initiative for Network Information (DINI) maintains DINI certificate (DINI releases new version in every three years in German, Spanish and English languages) that specifies minimum essential elements for sustainable maintenance of open access repositories in terms of technical, organizational, and legal aspects. DINI is a national certificate and provides DINI seal to repositories to assure trustworthiness and quality of the services. Many German OA initiatives like EconStar, pedocs of German Institute for International Educational Research, edoc of Humboldt University are DINI certified OA services. DINI Certificate supports many international interoperability initiatives such as OAI-PMH, Dublin Core and fully compliant with DRIVER.

URL: www.dini.de/dini-zertifikat/english

Present version: Year 2010 Version (Year 2013 Version is due)

Sponsor: DINI

Documentation: http://nbn-resolving.de/urn:nbn:de:kobv:11-100182800

2.3.4 Statistics and usage data-level Interoperability Initiatives

Citation is an important part of scholarly communication process. With the advent of ICT-enabled scholarly communication, different other parameters like number of hits, number of downloads, ranking by popularity are considered as parameters for measuring quality of research output. In OA (both Green OA and Gold OA) log entries store usage events. Analysis of log entries may be utilized for assessing the usage of OA objects. Many value-added services may also be generated from usage statistics like creation of a network of related resources, linking researchers working in the same area and development of recommender system. Obviously, usage statistics based services can be much more effective through integration of usage data from different OA journals and OA repositories. The usage statistics service is considered as an important value-added service for open contents management systems. Apart from the contributors and users of open access resources, funding agencies are also interested in availability of integrated usage data to measure research impact and to analyze trends over time. The major challenge is to develop a techno-organizational model for the recording, reporting and consolidation of usage of open contents available from OA journal publishers, toll publishers, aggregators, institutional repositories and subject repositories. In open access repositories multiple-deposit is a common feature. For example, open contents may be written by multiple authors from different institutions and thereby may be deposited in multiple repositories including publisher portals. In such cases availability of complete usage data for a specific open content is simply beyond the scope of a single repository. Many guideline and best practices in OA advocated for the provision of usage data from repositories to end users (such as DRIVER, OpenAIRE, RepNet etc). OpenAIRE specifies the metadata format that can be used to incorporate information of usage events and describes appropriate transfer protocols. It also prescribes a model for harvesting usage statistics from OAI/PMH compliant repositories (see Figure 14).

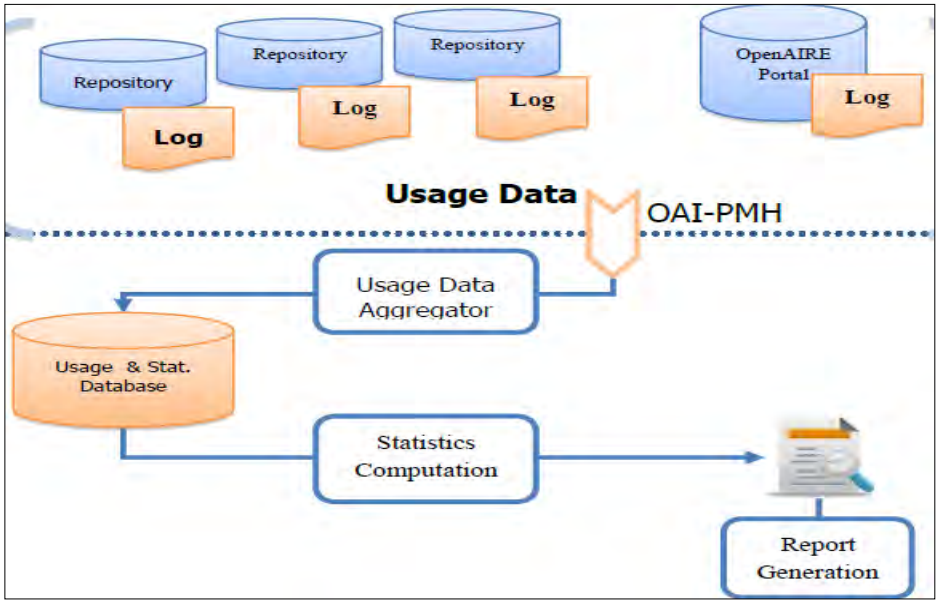

Figure 14: Model for Harvesting Usage Data (*Source*: OpenAIRE usage statistics service model)

Some of the well-known interoperability initiatives in this direction are:

COUNTER (Counting Online Usage of Networked Electronic Resources)

You have already understood from the above discussion that standardization is required for comparing, analyzing and aggregating usage data from distributed repository services. Uniformity is required primarily at two levels – i) standards for storing usage data in a uniform format; and ii) standards for transfer of usage data across repositories. The project COUNTER is the first such initiative in this direction. It may be considered as the mother project for standardization of usage data and statistics. Most of the large-scale national repository initiatives already defined standards for COUNTER compliant usage statistics (for example PIRUS in UK, OA-Statistics in Germany, SURFSure in Netherlands and NEEO in Belgium). COUNTER allows four categories of non-textual resource- image, video, audio and other. COUNTER is a code of practice for managing usage data for digital resource repositories.

URL: http://www.projectcounter.org/index.html

Present version: Stable

Sponsor: UK based initiative

Documentation: http://www.projectcounter.org/code_practice.html

SUSHI (Standardized Usage Statistics Harvesting Initiative)

The SUSHI is a protocol designed for the transmission and sharing of COUNTER-compliant usage data from repositories, OA publishers, toll publishers, aggregators and other bibliographic service providers who are able to present usage data in COUNTER-compliant format. This protocol is a product of NISO (National Institute for Standards Organization, US) and aims to alleviate automated integration of large-scale usage data from different sources including open access service providers. The SUSHI protocol[77] also includes a Schema specification in XML format () that allows integration and aggregation of COUNTER-compliant usage reports quickly and easily by repository manager at local level. The protocol is a complete pack that includes Documentation, SUSHI Tools, SUSHI Schemas, SUSHI Reports Registry, SUSHI Server Registry, SUSHI Developers List and SUSHI FAQs.

URL: http://www.niso.org/workrooms/sushi/

Present version: Stable

Sponsor: NISO, US

Documentation: http://www.niso.org/workrooms/sushi/

[77] http://www.niso.org/ schemas/sushi/counterElements4_0.xsd

KE-USG (Knowledge Exchange Usage Statistics Guidelines)

The Knowledge Exchange Usage Statistics Guidelines (KE-USG) is an important initiative in aggregating and transferring usage data from OA journals and OA repositories. It is basically a set of guidelines that includes metadata format for usage data (the format is compliant with OpenURL Context Objects), prescribes protocol for transferring usage data across repositories (SUSHI or OAI/PMH), suggests rules for normalizing usage data (issues related with 'robot filtering' and 'double clicks') and provides a framework of interaction (how providers of usage data and service providers can interact and transfer data including legal boundaries). The three major national level initiatives in usage data namely PIRUS2 in UK, OA-Statistics in Germany, SURFSure in Netherlands have contributed considerably in developing KE-USG. These initiatives are completely compliant with KE-USG.

URL: http://www.knowledge-exchange.info/Default.aspx?ID=365

Present version: Version 2.0 stable

Sponsor: Knowledge Exchange (cooperation from JISC, DEFF, SURF and DFG)

Documentation: http://wiki.surf.nl/display/standards/KE+Usage+Statistics+Guidelines+Work+group

NEEO (Network of European Economist Online)

Network of European Economists Online (NEEO) is an international consortium of 18 universities. NEEO maintains a subject repository in the domain of Economics. This initiative, originated in Belgium, provides a set of guideline to aggregate item level usage data based on article identifier and user identifier. It also developed extensive guidelines for – i) creation of usage statistics; ii) aggregation of usage statistics; and granularity for usage statistics for designing recommender system. NEEO differs from COUNTER in view of the followings - typically publisher usage data is available in COUNTER format (NEEO provides platform for both OA and toll publishers); COUNTER uses journals as lowest level of granularity (NEEO emphasizes item-level usage and thereby more granular than COUNTER). NEEO uses three major standards SWUP (the Scholarly Works Usage Community Profile) for usage data description, OpenURL ContextObject for IR log analysis and OAI/PMH protocol for transferring usage data.

URL: http://www.neeoproject.eu/

Present version: Version 1.4

Sponsor: NEEO-WP 5

Documentation: http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/

OA-Statistik

Open Access Statistics (OAS) aims to support open access movement by promoting the usage data and statistics. OA-Statistik is a German project to aggregate globally available usage data mainly from open access service providers by providing technical infrastructure for collecting, processing and presenting usage data at item level. The infrastructure is a two-layer system – i) layer 1 collects from OAS data providers, processes usage data and presents processed data in standard interface to layer 2; ii) layer 2 includes a central OAS service provider which harvests usage data from OAS data providers, calculates statistics from usage data and finally makes analytical results available to participating repositories and other value-added service providers.

URL: www.dini.de/projekte/oa-statistik/english

Present version: Version 5

Sponsor: DINI, Germany

Documentation: http://www.dini.de/fileadmin/oa-statistik/projektergebnisse/Specification_V5.pdf

PIRUS (Publishers and Institutional Repository Usage Statistics)

Usage-based metrics is now accepted by research community as a tool to assess the impact of journal articles. PIRUS is a JISC, UK funded initiative in view of the emergence of online usage data as an alternative measure of article and journal value. PIRUS is a code of practice for managing usage data and is considered as open international standard in the domain. The aim of this initiative is to provide a set of guideline and standards in recording, exchange and interpretation of online usage data at the individual article level. In fact PIRUS is granular extension of COUNTER standard at the item level. Although it is primarily meant for COUNTER-compliant repositories, Non-COUNTER-compliant service providers may also use the Secondary Clearing House services to generate PIRUS compliant usage reports from their raw usage data. The major objectives of PIRUS are:

- To define a core set of standards for repositories for producing usage statistics;
- To collect and process usage statistics at the individual article level
- To derive consolidated PIRUS usage statistics per article;
- To provide a central source of validated, consolidated PIRUS usage statistics for individual articles; and
- To develop a suite of open source tools for generating COUNTER-compliant usage data at item level;

PIRUS has a close liaison with the project COUNTER. But PIRUS gives more emphasis on item level usage data through a framework of standards that include article types to be counted; article versions to be counted; data

elements to be measured; definitions of these data elements; content and format of usage reports; requirements for data processing; requirements for auditing; and guidelines to avoid duplicate counting. At the item level, PIRUS suggests to include following metadata elements – i) either print ISSN OR online ISSN; ii) article version, where available; iii) article DOI; iv) online publication date or date of first successful request; and v) monthly count of the number of successful full-text requests. Other optional but desirable metadata elements are - i) journal title; ii) publisher name; iii) platform name; iv) journal DOI; v) article title; and vi) article type. The item level granularity in PIRUS is achieved through two additional metadata – article DOI and ORCHID as author identifier.

URL: http://www.projectcounter.org/pirus.html

Present version: Release 1, October 2013

Sponsor: JISC, UK and Mimas (University of Manchester)

Documentation:
http://www.projectcounter.org/documents/Pirus_cop_OCT2013.pdf

SURE (Statistics on the Usage of Repositories)

The project SURE is an initiative by a group of Dutch universities. It is funded by SURF foundation. It aims to coordinate and aggregate usage data from repositories in Netherlands. The technical specification of the SURE project is fully compatible with the national and international initiatives in the area of usage statistics. The SURE project uses OpenURL Context Object Schema and the schema is compatible with other similar initiatives like PIRUS in UK, NEEO in Belgium and OAS in Germany. This project uses NARCIS portal (the gateway to scholarly objects in Netherlands) to store usage data. The dashboards provide usage statistics for individual objects (with different visualization facilities) to different service providers in OA. Local repositories can also use API or Widget (developed by SURE project) to integrate usage data at respective user interfaces (see http://repositorymetrics.narcis.nl/). The SURE project is also planning to provide matrices for individual contributor on the basis of Digital Author Identifiers (DAIs).

URL: http://wiki.surf.nl/display/statistics/Home

Present version: Draft version available

Sponsor: Open Society Institution (OSI) with technical support from Knowledge Exchange project

Documentation:
http://wiki.surf.nl/display/standards/KE+Usage+Statistics+Guidelines

2.3.5 Identifier-level Interoperability Initiatives

As already discussed, unique identification schemes are essential to ensure cross-system interoperability in terms of digital scholarly objects, contributors and datasets. Use of unique identifiers in achieving interoperability is not a new idea in the area of library and information science. As a library professional you know the value of name authority data to identify authors consistently and the role of ISBN/ISSN/ISMN etc to identify documentary resources uniquely. In last few years different standards have emerged in the digital scholarly resource landscape such as standards for unique author identification (e.g. ORCID and AuthorClaim), standards for object identification (e.g. DOI, Handle system, PersID) and recently standards for dataset identification (e.g. DataCite). All these emerging standards and services help to support open access interoperability.

Author Identifiers

There are some established author identifier schemes like RePEc Author Service (RAS) in the subject field Economics and some emerging services like AuthorClaim and ORCID. RePEc is mainly concerned with one particular discipline and therefore cannot be considered as universal standard in the area of author identification but it has already passed ten years of service and a large pool of researchers in Economics are registered members of RAS. Similarly E-LIS (a subject-specific repository in Library and Information Science) has its own author identification system. The two universal author identification systems (AuthorClaim and ORCID) have considerably been influenced by the said subject-specific author identification systems. For example, AuthorClaim is developed by Thomas Krichel, the creator of RAS. But we should remember that author identifier has no value if it is not linked with biographic and bibliographic information. The integration of author ID, his/her biographic data (institute – past and present, co-authors, subject area etc) and his/her scholarly works (with necessary bibliographic data elements) creates an author *profile*, and this can be done either by the system issuing the identifier, or by the systems that collect scholarly contributions, or by one or more other systems.

AuthorClaim Registration Service

Thomas Krichel (the creator of RePEc and RAS) developed AuthorClaim with the fund support from Open Society Institute (OSI) under ACIS project[78]. It aims to create profile of scholars in a bibliographic database for linking the authors uniquely with the scholarly contributions. This author identification system has following features:

- Allows authors to build a profile in AuthorClaim through registration (only prerequisite is e- mail ID of author);
- System searches for related publications and ask author to identify his/her

[78] http://acis.openlib.org/

publications (there is option for manual entry for different types of resources);

- Bibliographic databases that use AuthorClaim record can link the profile page of author;
- Distinguishes authors from each other even with the same name;
- Provides regular statistics about downloads and citations to authors; and
- Helps authors and other service providers to compute various rankings related to productivity (e.g. h-index).

URL: http://www.authorclaim.org

Status: Operational

Sponsor: Open Society Institute (OSI) under ACIS project (see http://acis.openlib.org/)

Open Researcher and Contributor ID (ORCID)

ORCID is an open international initiative to provide a registry of unique researcher identifiers at global scale. It offers a method to connect scholarly contributions with author identifiers. Additionally, ORCID also allows cooperation with other identifier systems. This author identification system has following features -

- Author creates their profile and links profile with his/her list of publications (allows import of bibliographic data elements through standards like RIS, bibtex etc.;
- Author can enter ORCID ID during submission of digital resource in repositories;
- Repositories can ingest ORCID record into repository software architecture; and
- Publishers can use ORCID ID of an author during submission of manuscript.

URL: http://www.orcid.org

Status: Operational, stable from October 2012

Sponsor: ORCID Society

In the library world, VIAF (Virtual Internet Authority File) is coming as a comprehensive name authority service. It is an OCLC initiative to aggregate name authority data from 25 national libraries and the interesting fact is that the whole dataset is now available as Linked Open Data (LOD). It means these datasets can be linked dynamically with the *DC.Creator* metadata field in different repository software. However, the major initiatives in the area of author identification are reported by Fenner (2011) and on the basis of that a report on chronological evolution of initiatives with relevant information is given in Table 8.

Table 8: Major Author Identification Systems (Source: Martin Fenner, 2011)

Initiatives	Funding agency (types of service)	Major features	Subject scope Geographic scope	Year of origin
AuthorClaim	Open Library Society (Nonprofit)	Started as RePEc Author Service (RAS) by Thomas Krichel, extended as AuthorClaim in 2008.	Subject scope: All Geographic scope: All	1999 (as RePEc) and 2008 (as AuthorClaim)
		Project/Service URL: http://authorclaim.org		
LATTES	National Council for Scientific and Technological Development (Government)	Links many bibliographic databases and mandatory in Brazil since 2002.	Subject scope: All Geographic scope: Brazil	1999
		Project/Service URL:http://lattes.cnpq.br/		
VIAF	Online Computer Library Center (OCLC) and 25 national libraries (Nonprofit)	An OCLC initiative to aggregate name authority files of 25 national libraries and presently available as linked open data	Subject scope: All Geographic scope: Global	2003
		Project/Service URL: http://viaf.org/		
NARCIS	Royal Netherlands Academy of Arts and Sciences (KNAW) (Government)	Integrates datasets of NARCIS as portal for scholarly resources in Netherlands	Subject scope: All Geographic scope: Netherlands	2004
		Project/Service URL: http://www.narcis.nl		
ArXiv Author ID	Cornell University Library (Academic)	ArXiv is the forerunner of e-prints archives and the services introduces it in 2005	Subject scope: Physics, mathematics, computer science Geographic scope: All	2005
		Project/Service URL: http://www.arxiv.org		
Scopus Author ID	Elsevier (Commercial)	Links with one of the most comprehensive bibliographic database Scopus	Subject scope: All Geographic scope: All	2006
		Project/Service URL: http://www.scopus.com		
Names Project	Mimas, British Library (Academic)	Acts as author identification system for UK based researchers	All Geographic scope: United Kingdom	2007
		Project/Service URL: http://names.mimas.ac.uk/		
Researcher ID	Thomson Reuters (Commercial)	Links with one of the most comprehensive bibliographic database,Web of Science	Subject scope: All Geographic scope: All	2008
		Project/Service URL: http://www.researcherid.com/		
ORCID	ORCID (Nonprofit)	Connects bibliographic database CrossRef and also links other author identifier systems.	Subject scope: All Geographic scope: All	2009
		Project/Service URL: http://www.orcid.org/		
PubMed Author ID	National Library of Medicine (NLM) (Government)	Integrates many biomedical bibliographic databases and services.	Subject scope: Life sciences Geographic scope: All	2010
		Project/Service URL: http://www.pubmed.gov/		

Object Identifiers

Digital ID is a necessary foundation of many forms of online exchange. Internet allocates a numeric identifier for host computers or servers, called IP address. Domain name system is textual representation of IP address. These two schemes uniquely identify servers (or any connected device) in the network. Apart from these interoperability standards for machines, standards like URL, URN, DOI CNRI handle, PersID, DataCite etc are in use to achieve interoperability in digital resource access and exchange. Interoperability requires persistent and actionable object names. The features of unique identifier for digital resources are (DOI[79] Foundation):

- Object names require mechanisms for persistence;
- Object identification requires action-ability (it means resolution from a name to some service);
- Object representation requires specification of an object (it may be achieved either through simple referencing or more formal description); and
- Object naming requires standard syntax (demands prescriptive rules for assigning identifiers in a standard format ensuring uniformity and uniqueness).

The major initiatives in unique object identification of scholarly resources are:

DOI (Digital Object Identifier) System

The International DOI Foundation developed a generic standard for unique identification of digital objects including scholarly digital resources. The features of the DOI system are as follows:

- The DOI System uses naming syntax on the basis of NISO standard Z39.84;
- DOI name persistence is guaranteed through social infrastructure which provides rules for registration, formal resilience procedures etc;
- The DOI System applies the Uniform Resource Name (URN) and the Uniform Resource Identifier (URI);
- URI and URN specifications in DOI deal only with syntax;
- Uniform Resource Identifier (URI) specification in DOI is based on IETF RFC 2396 standard;
- URN (Uniform Resource Name) specification in DOI is based on RFC 2141;
- DOI is Neutral as to implementation (the design of DOI is not specific to Web only and may work in non-Web environment);
- DOI allows granularity of naming and administration at the object level; and
- DOI is neutral as to language, script or character set (Unicode may represent DOI in any script).

[79] http://www.doi.org/

Handle System

Handle system is an initiative of Corporation for National Research Initiatives (CNRI) to manage unique and persistent identification of digital resources in a heterogeneous network environment. The handle system is based on the Digital Object Architecture of CNRI. The architecture has following features:

- Allows identification, access and protection (if required);
- Machine and platform independent;
- Incorporates not only digital object but also unique identifier and associated metadata;
- Metadata may include rights related information, licensing agreement and restriction on access (if any);
- Handle includes namespace, open set of protocols and necessary reference implementation;
- Protocols enable to persistent identifiers of digital resources (known as handles) in a distributed computing system;
- Handles can be resolved into set of information that are necessary to locate, access and authenticate the digital resources;
- Information set can be changed and modified (to suite current state of the resources) without changing the identifier;
- Ensures persistent access to digital objects in spite of changes in location and other related status information;
- Handles allow identification of digital objects with a persistent URL (means handle can identify a digital object even the URL of the object itself changes);
- Repositories or other service providers need to register in CNRI handle (handle.net) system and implement handle at local level (see Figure 15 and Fig 16).

The Handle System has significant advantages in unique and persistent identification of digital objects: i) it is a global resolution service; ii) the plug-in is available freely and tested across multiple platforms/applications; iii) URN plug-ins may be configured to provide server-side support; iv) platform independent implementation; iv) available with added security features; v) can be delivered through web browser. DSpace, a globally reputed open source repository management software incorporated CNRI handle system right from the beginning. After registering and obtaining CNRI handle, administrator of DSpace can enter handle obtained in configuration file.

```
##### Handle settings ######

# Canonical Handle URL prefix
#
# By default, DSpace is configured to use http://hdl.handle.net/
# as the canonical URL prefix when generating dc.identifier.uri
# during submission, and in the 'identifier' displayed in JSPUI
# item record pages.
#
# If you do not subscribe to CNRI's handle service, you can change this
# to match the persistent URL service you use, or you can force DSpace
# to use your site's URL, eg.
#handle.canonical.prefix = ${dspace.url}/handle/
#
# Note that this will not alter dc.identifer.uri metadata for existing
# items (only for subsequent submissions), but it will alter the URL
# in JSPUI's 'identifier' message on item record pages for existing items.
#
# If omitted, the canonical URL prefix will be http://hdl.handle.net/
handle.canonical.prefix = http://hdl.handle.net/

# CNRI Handle prefix
handle.prefix = 123456789

# Directory for installing Handle server files
handle.dir = ${dspace.dir}/handle-server
```

Figure 15: CNRI Handle Implementation in DSpace

Figure 16: Receiving of CNRI Handle by an Object in DSpace

After submission of digital objects in repository by authors/submitters, a handle is allotted by the software according to the handle prefix (here 123456789 – a fictitious handle). For example (see Figure 16) after successful submission one digital objects received handle 123456789/3 with URL http://hdl.handle.net/123456789/3.

The user interface also allows global unique and persistent access to digital object during retrieval irrespective of the URL of the repository (see Figure 17).

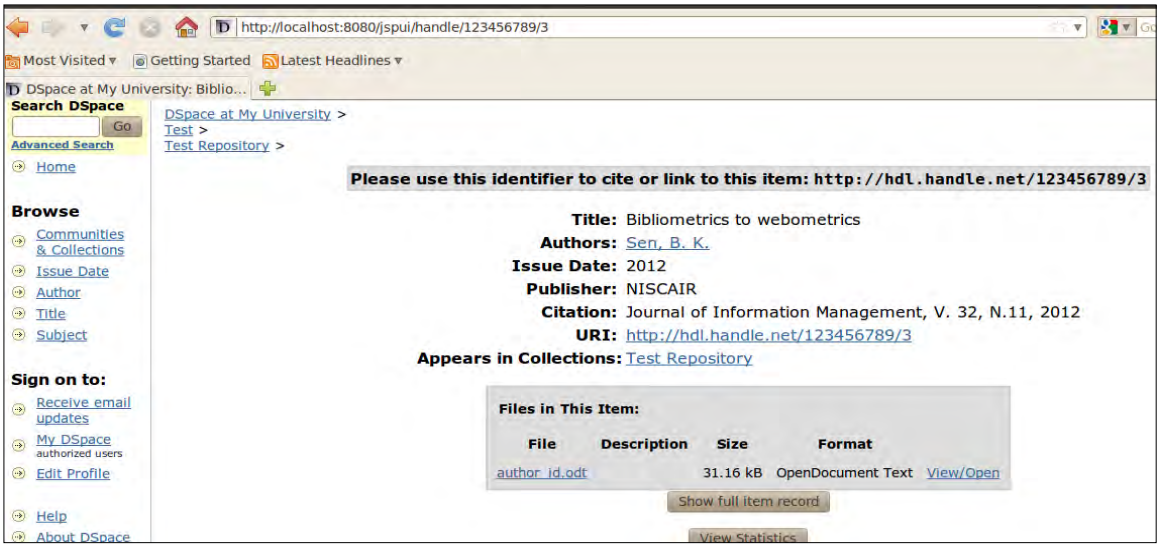

Figure 17: Persistent Access to Digital Object through CNRI handle
(See CNRI Handle System URL hdl.handle.net)

PersID

This unique identification system is a joint initiative of national libraries (National libraries of Sweden, Denmark, Germany), national research bodies (DANS, Netherlands; DEFF, Denmark; FDR, Italy, CNR, Italy) and some international projects (Knowledge Exchange, SURF Foundation). This identification system provides identifiers by combining URN (Uniform Resource Names) and NBN (National Bibliography Numbers) in the form of URN: NBN. It supports persistent identification of knowledge objects through an international infrastructure and knowledge base. URN is a specification of IETF (Internet Engineering Task Force), a W3C organ and bibliographic identification systems includes ISBN (a 13 digit number in the line of EAN – European Article Number), ISSN, ISMN etc. PersID[80] may be applied to a wide range of web resources.

DataCite

DataCite[81] is an international foundation working in the area of unique and persistent identification of published digital datasets since 2009. It is a membership based organization and is closely related to over hundred data centres all over the world. The founder members of DataCite include esteemed institutes like the British Library, Purdue University, National Library of Science and Technology, Germany, National research Council, Canada and many more. The partner data centres include California Digital Library, US; Australian National Data Service; Beijing Genomics Institute etc. The DataCite initiative has following features:

- Works in close liaison with data centres around the world-wide;
- Assigns persistent identifiers to datasets in consultation with data centres;

[80] http://www.persid.org/
[81] http://www.datacite.org/

- Includes method for data citation, data discovery and dataset linking with related resources such as journal papers;
- Persistent identifiers will be assigned against membership registration;
- Citable datasets (as scholarly contribution) may create a new method for measuring scientific productivity;
- Promotes data arching for future use and re-purposing.

DataCite is a member of the International DOI Foundation. The members of DataCite support registration for DOIs. Some DataCite members provide registration facilities through their own APIs and others use DataCite API directly for registration (see Figure 18).

[dspace]/config/spring/api/identifier-service.xml

```
<?
<!--
Copyright (c) 2002-2010, DuraSpace.  All rights reserved
Licensed under the DuraSpace License.
A copy of the DuraSpace License has been included in this
distribution and is available at: http://www.dspace.org/license
-->
<beans xmlns="http://www.springframework.org/schema/beans"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.springframework.org/schema/beans
http://www.springframework.org/schema/beans/spring-beans-2.5.xsd">

<bean id="org.dspace.identifier.IdentifierService"
class="org.dspace.identifier.IdentifierServiceImpl"
autowire="byType"
scope="singleton"/>

<bean id="org.dspace.identifier.DOIIdentifierProvider"
class="org.dspace.identifier.DOIIdentifierProvider"
scope="singleton">
<property name="configurationService"
ref="org.dspace.services.ConfigurationService" />
<property name="DOIConnector"
ref="org.dspace.identifier.doi.DOIConnector" />
</bean>
<bean id="org.dspace.identifier.doi.DOIConnector"
class="org.dspace.identifier.doi.DataCiteConnector"
scope="singleton">
<property name='DATACITE_SCHEME' value='https'/>
<property name='DATACITE_HOST' value='mds.datacite.org'/>
<property name='DATACITE_DOI_PATH' value='/mds/doi/' />
<property name='DATACITE_METADATA_PATH' value='/mds/metadata/' />
<property name='disseminationCrosswalkName' value="DataCite" />
</bean>
</beans>
```

Figure 18: Configuration of DataCite as Identifier service in DSpace version 4.0

DataCite will be utilized in DSpace from version 4.0 onwards. DSpace is planning to use DataCite in two alternative ways – i) administration of DOIs by using the DataCite API directly; or ii) by using the API from EZID (a service of the University of California Digital Library, an active member of DataCite Initiative).

2.3.6 Object-level Interoperability Initiatives

Enhanced publications are rapidly becoming a trend of the scholarly communication process. Research publications are now increasingly attached with datasets, models, algorithms, images, streaming videos, post publication materials (like comments, blog posting, citations, ranking etc). Enhanced publications are compound digital objects that include text, audio, video, image etc. The concept of enhanced publications in OA domain was first reported by DRIVER project and developed further by another OA project SURF to integrate open data and publication. Exchange and sharing of compound digital objects or enhanced publication failed due to two reasons – i) there is no standard way to identify an aggregation; and ii) there is no standard way to describe the constituents or boundary of an aggregation.

OAI-ORE, as an interoperability standard for compound digital objects, aims to provide solution that supports aggregations of Web resources. Open Archives Initiative (OAI) - Object Reuse and Exchange (ORE) is a standard developed by Open Archive Initiative under the leadership of Pete Johnston of Eduserv Foundation. OAI-ORE works on the basis of following principles:

- Based on Web architecture with four basic components – i) Resource (an item of interest); ii) URI (a global resource identifier); iii) Representation (a data stream accessible through URI by using a protocol like HTTP); and iv) Link (a connection between two resources);
- Supports Semantic web, Linked data and Cool URI;
- Provides XML-based serialization for the Resource Description Framework (RDF);
- Can unambiguously refer to an aggregation of Web resources through Aggregation URI (represents a set or collection of other Resources);
- Web Aggregation is also called Resource Map (provides machine-readable representation about the Aggregation and it has a URI);
- Resource Maps can be expressed in different formats including Atom XML, RDF/XML, RDFa, n3, turtle, and other RDF serialization formats;
- Resource Map is able to return an RDF/XML or Atom XML document against HTTP request and clients/agents can then interpret resource map to provide enhanced services like navigation, printing, archiving, visualizing, and transforming the Aggregation.

Almost all major open source repository management software like DSpace, Eprint and Fedora are supporting OAI-ORE for harvesting compound digital objects. Eprint archive repository management software allows both harvesting compound digital objects by using OAI-ORE and can also export items in OAI-ORE compatible format.

OAI-ORE is presently in version 1.0 and it has all the capabilities to emerge as de facto global standard for the interoperability of aggregated digital resources. It follows a simple but robust ORE data model and compliant with Linked Open Data (LOD) and Semantic Web technologies.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

3) List the initiative related with usage statistics of open access resources.

...……..……..……...

……..……..……...

……..……..……...

4) What is OAI-ORE?

...……..……..……...

……..……..……...

……..……..……...

2.4 MAJOR INTEROPERABILITY STANDARDS

Order, collaboration and interoperability are three most important prerequisites for effective information services. All these requirements depend on effective standards. Library services have long depended on shared standards. The case of open access interoperability is no exception. ANSI defined a standard as a specification accepted by recognized authority as the most practical and appropriate current solution of a recurring problem. ISO/IEC Guide 2:2004 defines a standard as a document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context. In the area of interoperability, there are many initiatives to cover different areas with specifications and standards. However, three major interoperability standards are accepted widely by information professionals all over the world. These are Z 39.50 for distributed cataloguing, OAI/PMH for metadata harvesting and OAI-ORE for sharing compound digital objects.

2.4.1 Z 39.50

The ANSI/NISO standard Z39.50-2003 (Information Retrieval: Application Service Definition & Protocol Specification) is adopted widely by library systems, library automation software vendors, and digital library developers (such as Greenstone Digital Library Software) as a protocol for searching catalogue databases in different library systems and software across the globe. The retrieved results may be saved in desired format and may also be edited before inclusion in the local catalogue database. ISO and other major national SDOs adopted this standard widely in developing equivalent standards such as

IS 15390:2003 (by the Bureau of Indian Standards) and ISO 23950:1998 (by ISO). Z39.50 was developed over Open System Interaction (OSI) protocol. It is basically a program-to-protocol and divided into two parts – client (called an 'origin' in the standard) and server (called a 'target' in the standard). Z39.50 supports most of the major MARC formats and can operate over WWW through http- Z39.50 gateway.

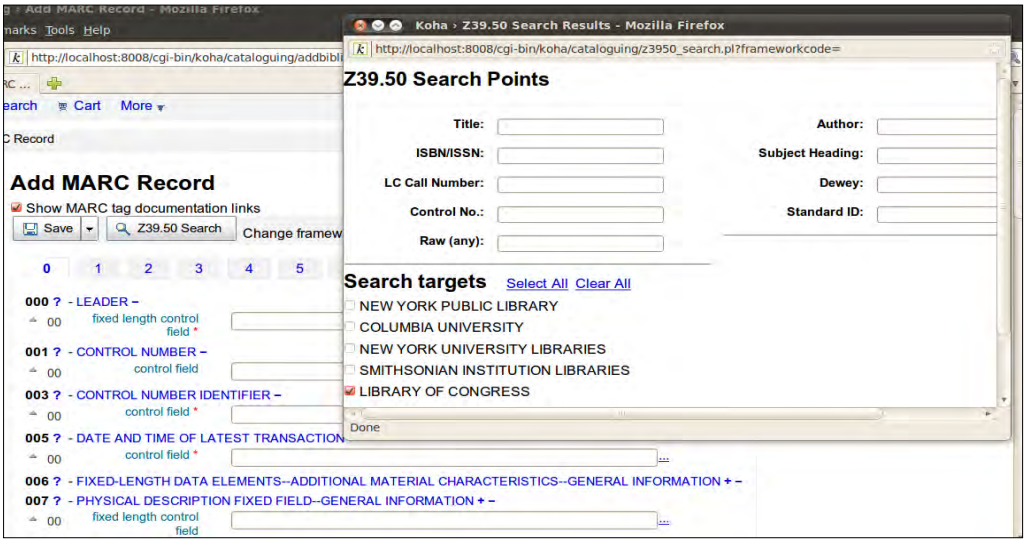

Figure 19: Application of Z 39.50 in Distributed Searching in Koha

A Z39.50 service can be implemented in libraries in diverse ways. It can offer library resources through Z39.50 server and by using Z39.50 client software, a library can go for intersystem searching and bibliographic record transfer irrespective of different hardware and software (See Figure19). Considering the endless advantages of Z39.50 protocol, modern LMSs are incorporating Z39.50 client protocol suite in the catalogue module. Z 39.50 standard is now being replaced by emerging interoperability standards like SRU/SRW[82] (Search and Retrieve URL/Web Service - Web services for search and retrieval based on Z39.50, developed by Library of Congress) and ZING (Z39.50-International: Next Generation covers a number of initiatives by Z39.50 implementers to make the semantic content more broadly available).

2.4.2 OAI/PMH

The OAI/PMH is a light-weight standard protocol developed by Open Archive Initiative (OAI) for harvesting metadata records from 'data providers' to 'service providers'. It provides a set of rules to harvest metadata of knowledge objects from a repository not the full-text objects. The full-text resource may be retrieved form source repository or data provider. There are two groups of operators in the OAI-PMH framework:

- **Service Providers** use metadata harvested via the OAI-PMH as a basis for building value-added services; and

[82] http://www.loc.gov/standards/sru/

- **Data Providers** administer systems that support the OAI-PMH as a means of exposing metadata.

At the time of harvesting, service providers send requests to other repository i.e. data provider in the form of OAI/PMH verbs (request type). The OAI/PMH includes Six Verbs and these are:

- **Identify** (return general information about the archive and its policies);
- **ListSet** (provide a listing of sets in which records may be organized);
- **ListMetadataFormats** (list metadata formats supported by the archive as well as their schema locations and namespaces);
- **ListIdentifiers (**list headers for all items corresponding to the specified parameters);
- **GetRecord (**returns the metadata for a single item in the form of an OAI record); and
- **ListRecord (**retrieves metadata records for multiple items).

Service provider can get specific type of metadata by GetRecord and ListRecord. The request is transferred on the basis of the rule of HTTP over the Web (see Figure 20).

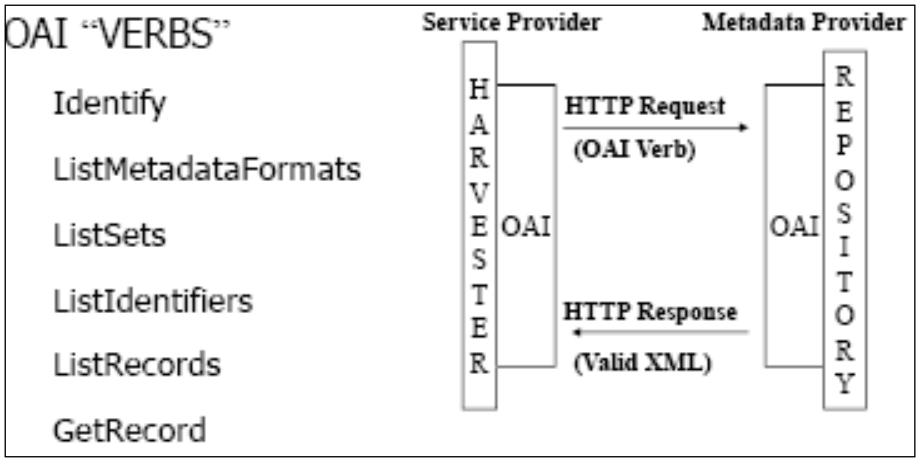

Figure 20: OAI/PMH Protocol (*Source*: http://www.oaforum.org/tutorial/)

OAI/PMH is a matured interoperability standard now. Almost all the OA repository software are compliant with this standard. It is also applied for harvesting usage data by the initiatives like KE-USG, NEEO and OA-Statistik. Many open source harvesting software are also compliant with OAI/PMH.

2.4.3 ORE

You already know that OAI-ORE, as an interoperability standard for compound digital objects, aims to provide solution that supports aggregations of Web resources. Open Archives Initiative (OAI) - Object Reuse and Exchange (ORE) is an open interoperability standard developed by Open Archive Initiative. OAI-ORE standardizes the description of the relationship between digital objects. This relationship could be between versions of an object, such as might be found in a repository record, or aggregations of objects, such as a Web page with images, or a collection of chapters in a book.

The OAI-ORE standard has four basic components to support web aggregation.

- **OAI-ORE Model**: The model advocated using RDF model to annotate objects with metadata at the repository level to support semantic web technologies like Linked Data and Cool URI.
- **Aggregations**: Alongside the RDF model, OAI-ORE specifies the concept of Web Aggregations and Aggregated Digital Resources (an Aggregation is simply a set of Aggregated Resources, all of which are represented by URIs.)
- **Resource Maps**: The next level of the standard suggests Resource Map. A Resource Map describes a single Aggregation with unique URI. In OAI-ORE model a Resource Map can only link to a single Aggregation in the OAI-ORE model.
- **Representations**: A Resource Map in the ORE standard requires representation for interoperability. This can be done in two ways (as prescribed by OAI-ORE) – i) *RDF/XML Serialization;* and ii) *Atom/XML Serialization*.

OAI-ORE is an operational standard and can be used to transfer resources from one repository software to another. For example both EPrints and Fedora (two reputed open source repository software platforms) are now compliant with OAI-ORE by applying OAI-ORE plug-ins in terms of cross-system import and export of compound digital objects.

2.4.4 Others

The other interoperability standards (related directly or indirectly with open access interoperability issues), developed by Standard Development Organizations (SDOs), National Libraries and Library Associations are (illustrative list not comprehensive)

- ISO 10957:1993 (Information and documentation -- International standard music number (ISMN))
- ISO 15706-1:2002 (Information and documentation -- International Standard Audiovisual Number) (ISAN) -- Part 1: Audiovisual work identifier);
- ISO 15706-2:2007 (Information and documentation -- International Standard Audiovisual Number (ISAN) -- Part 2: Version identifier);
- ISO 2108:2005 (Information and documentation -- International standard book number (ISBN));
- ISO 21127:2006 (Information and documentation -- A reference ontology for the interchange of cultural heritage information);
- ISO 3297:2007 (Information and documentation -- International standard serial number (ISSN))
- ISO 3901:2001 (Information and documentation -- International Standard

Recording Code (ISRC))

- ISO/AWI TR 19934 (Information and documentation -- Statistics for the use of electronic library services);;
- ISO/CD 27729(Information and documentation -- International Standard Party Identifier (ISPI));
- ISO/CD TR 26102 (Information and documentation -- Requirements for long-term preservation of electronic records)
- ISO/CD 26324 for Digital Object Identifier system;
- ISO/NP 27730 (Information and documentation -- International standard collection identifier (ISCI));
- ISO/TR 21449:2004 (Content Delivery and Rights Management: Functional requirements for identifiers and descriptors for use in the music, film, video, sound recording and publishing industries);
- ISO/TR 21449:2004 (Content Delivery and Rights Management: Functional requirements for identifiers and descriptors for use in the music, film, video, sound recording and publishing industries);
- ISO/TR 26122:2008(Information and documentation -- Work process analysis for records);
- MADS (Metadata Authority Description Standard) - XML markup for selected authority data from MARC21 records as well as original authority data ;
- METS (Metadata Encoding & Transmission Standard) - Structure for encoding descriptive, administrative, and structural metadata;
- MODS (Metadata Object Description Standard) - XML markup for selected metadata from existing MARC 21 records as well as original resource description
- SRU/SRW (Search and Retrieve URL/Web Service) - Web services for search and retrieval based on Z39.50;
- Z39.53 (Codes for the Representation of Languages for Information Interchange);
- Z39.56 (Serial Item and Contribution Identifier (SICI)).
- Z39.93 (The Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol);

The interoperability standards are originated from two groups of activities – i) cooperative standards developed by learned societies, library associations and national libraries; and ii) standards developed by national and international standard organizations (like ISO, NISO, BSI etc). The cooperative interoperability standards are open standards. The W3C (2006) provides a set of six pack criteria in defining *Open Standards* - **transparency** (due process is public, and all technical discussions, meeting minutes, are archived and citable in decision making), **relevance** (new standardization is started upon due analysis of the market needs, including requirements phase, e.g. accessibility,

multilingualism), **openness** (anybody can participate, and everybody does: industry, individual, public, government bodies, academia, on a worldwide scale), **impartiality and consensus** (guaranteed fairness by the process and the neutral hosting of the W3C organization, with equal weight for each participant), **availability** (free access to the standard text, both during development and at final stage, translations, and clear IPR rules for implementation, allowing open source development in the case of Web technologies) and **maintenance** (ongoing process for testing, errata, revision, permanent access).

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

5) What do you mean by 'Six Verbs' of OAI/PMH?

...

...

...

...

6) Mention the standards related with identifier-level interoperability.

...

...

...

...

2.5 APPLICATION OF INTEROPERABILITY: METADATA HARVESTING

The Open Archives Initiative Metadata Harvesting Protocol (OAI/PMH) supports interoperability and sharing of metadata across an array of open access repositories. The creation of large repositories by using OAI/PMH protocol is advantageous to bring together scholarly information bearing objects and cultural resources. However, the mixing of metadata from a variety of institutions and communities poses difficulties for discovery and interoperability. OAI/PMH differs from Z 39.50 in many aspects as interoperability standard (Table 9).

Table 9: Difference[83] between Z39.50 and OAI

Features	Z39.50	OAI
Content (Objects)	Distributed	Distributed
World View	Bibliographic	Bibliographic
Object Presentation	Data provider	Data provider
Searching is	Distributed	Centralized
Search done by	Data provider	Service provider
Metadata search is	Up to date	Stale
Semantic Mapping	When searching	Metadata delivery

Open source OAI harvesting tools provide opportunities to make the difficult job an easy one. There is an array of open source harvester software (compatible with OAI/PMH V.2) such as

- Arc (Old Dominion University, URL: http://arc.cs.odu.edu/)
- my.OAI (FS Consulting, Inc., URL: http://www.myoai.com/)
- OAIster (University of Michigan, URL: http://www.oaister.org/)
- PKP Harvester (Public Knowledge Project, URL: http://pkp.sfu.ca/harvester2/)

PKP (Public Knowledge Project) harvester developed by University of British Columbia has already been proved as an excellent metadata harvesting and presentation tool. This multi-platform Web-based tool extracts data and presents it in a coherent manner. It employs an intuitive user interface to organize data (see Evaluation of Open Source Spidering Tools[84]). PKP harvester (presently in version 2.3.x) is a platform independent **A** (Apache) - **M** (MySQL) – **P** (PHP) based application software. The AMP requirements are as follows:

- PHP >= 4.2.x (including PHP 5.x);
- MySQL >= 3.23.23 (including MySQL 4.x/5.x)
- Apache >= 2.0.4x or 2.0.5x; and
- Operating system: Any OS that supports the above software, including Linux, BSD, Solaris, Mac OS X, Windows (preferably NT based Windows flavors).

[83] http://hdl.handle.net/2142/147
[84] https://diva.cdlib.org/projects/ harvesting_crawling/recall_ crawl/spider_eval.pdf

Design and development of harvesting framework by using PKP requires an array of steps, strategies and planning. The three major components of such a framework design are:

i) **Development of software architecture** (Installation of Apache, MySQL and PHP and linking these tools for seamless interaction);

ii) **Selection, installation and configuration of harvesting too**l (Selection of PKP harvester and configuration settings like proxy server settings, homepage customizing etc); and

iii) **Selection of repositories and collection of essential attributes** for harvesting

The data elements like title, resource URL, OAI base URL, mail id of repository administrator are essential to start harvesting a selected repository. OpenDOAR is an excellent source to collect all the required data related to OAI/PMH compliant open access repositories such as title of repository, repository URL, and OAI base URL.

After successful harvesting, PKP harvester gathers metadata from selected repositories through OAI/PMH protocol and allows users to browse or search (both simple and advanced search interfaces are available) all metadata elements collected from different selected repositories through a single-window search interface and thereby helps users to get rid of drudgery of moving from one repository to another repository.

OAI/PMH supports harvesting not only the metadata formatted in DCMES (Dublin Core Metadata Elements Set) but also rich metadata sets like ETD-MS, Qualified DCMES etc. The interoperability initiatives like KE-USG, NEEO, SURE, PIRUS, OA-Statistik etc harvesting usage data through OAI/PMH. The exclusive open access search engine BASE (base-search.net) and services like OAISter depends on OAI/PMH for collecting metadata from different resources. Open source repository management software like Dspace, Eprint archive, Greenstone are fully compatible with OAI/PMH version 2.0 and these softwares can act as data providers as well as service providers.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

7) Differentiate Z 39.50 and OAI/PMH.

...……..……..……...

……..……..……...

……..……..……...

……..……..……...

8) List the steps of harvesting an OAI/PMH compliant open access repository.

……..……..……...

……..……..……...

……..……..……...

2.6 INTEROPERABILITY: TRENDS AND FUTURE

Open access resources, open source software and open standards are changing the interoperability scenario and most importantly these three distinct but interrelated movements are flying forward in harmony and through coordination. Most of the interoperability standards are open standards developed by learned societies, library associations and voluntary groups. Some of these open standards are accepted all over the world and are considered as *de facto* global standards in the area of interoperability (such as OAI/PMH, OAI-ORE, DataCite, etc). In open access domain heterogeneity is the norm and therefore techniques for interoperability are extremely crucial in reconciling distributed and diverse open access sources. We already know different levels of interoperability along with the initiatives and standards associated with each level. There are seven levels of interoperability of which six levels are already established. The seventh level i.e. semantic interoperability is presently the most challenging and the most promising area of interoperability. Semantic interoperability ensures meaningful exchange of information consistently among machines and people. It helps end users in general and researchers in particular to retrieve relevant items from diverse sources in concerted way. A combination of Resource Description Framework (RDF), XML and Ontology are being implemented to express digital objects relationships in a machine understandable manner. The object relationships are important elements of semantic interoperability. It allows creating machine-generated services – i) to support global representation of knowledge objects; ii) to make cross-discipline connections; and iii) to combine related resources on-the-fly to develop new information services. SIMILE (Semantic Interoperability of Metadata and Information in unlike Environments) is another promising initiative in semantic interoperability. SIMILE (see simile.mit.edu) is an initiative of MIT and it aims to enhance interoperability among digital assets, metadata schemas, integrated vocabularies, domain ontology, metadata, and services. This initiative also aims to develop comprehensive open source tools that allow open access systems to access, manage, visualize and reuse digital assets. Another emerging area in the domain of interoperability is Linked Open Data (LOD). Libraries all over the world are exploring the possibilities to export own bibliographic data in RDF triples and also investigating paths to integrate external linked datasets into their collections. LOD provides great opportunities to create new levels of user services and at the same time inviting challenges in developing interoperability standards for integrating LOD into local service framework (presently most of the LOD integrations are based on content negotiation). With the rising

importance of OA movement throughout the world, interoperability issues related with languages and scripts are major concern for OA service providers. Although Unicode (a 2 Byte character coding standard to cover all the scripts and languages of the world) standard is performing exceptionally well in scripts representation, lack of interoperability standards in transliteration and translation is creating problems for multilingual content integration.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this Module.

9) What is Linked Open Data (LOD)? How can we achieve interoperability in LOD?

...……..……..……...

……..……..……...

……..……..……...

10) Discuss three major trends in interoperability.

...……..……..……...

……..……..……...

……..……..……...

2.7 LET US SUM UP

Interoperability is a means to achieve global aggregation of open knowledge objects. We need internationally agreed upon standards to realize the dream of global open access infrastructure. So far there are seven levels of interoperability to handle metadata, multi-deposits, compound digital objects, usage data related to OA resources, unique resource identification, persistent author identifiers, network level exchange of OA resources and semantic level interoperability. Each of these interoperability levels are fortunately supported by various initiatives and standards. Most of these initiatives are coming with innovative solutions and standards. Many standards are considered as global *de facto* standards in the domain. Although all the levels are equally important, presently the metadata interoperability is the most matured area and the semantic interoperability is possibly the most futuristic in nature. It promises a new era of machine-generated meaningful exchange of OA resources across global service providers. The challenges in interoperability includes persistent identification of resources and contributors, multilingual contents transfer, managing interoperability of web aggregation, integrating OA networks operating at global scale and support for implementing interoperability standards to individual OA service providers at local level.

UNIT 3 RETRIEVAL OF INFORMATION FOR OA RESOURCES

Structure

3.0 INTRODUCTION

Content development is the most important aspect for information retrieval, whether it is in traditional environment or in web-information retrieval context. The content development encompasses varieties of activities from recording, managing, processing, and organizing to offering different services and ultimately the retrieving information. The organization and retrieving the information in open access environment is no exception to it. The philosophy and fundamentals remain the same. Only the techniques vary depending upon the development and availability of technology. In the previous two unites of

this module we have discussed the resource description and interoperability issues. This unit provides you an insight to the intricacies of information retrieval for open access literature.

3.1 LEARNING OUTCOMES

After going through this unit, you are expected to be able to:

- Describe the evolution of retrieval processes from traditional to Web-enabled IR;
- Understand basic concepts related to retrieval techniques, retrieval approach and retrieval models in OA retrieval systems as Web-enabled IR;
- Critically examine problems, prospects and services related to OA retrieval;
- Explain the role of text retrieval engines in OA retrieval; and
- Explore the use of support tools in OA retrieval including multilingual retrieval.

3.2 INFORMATION REPRESENTATION AND RETRIEVAL IN OA CONTEXT

Manning et al (2008) reported that information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). Library professionals all over the world are increasingly taking part in dissemination of open contents by setting up open access repositories, by publishing online open access journals and by creating single-window web-scale discovery services for open contents (Crow, 2002; Yang & Hofmann, 2011). All of these dissemination services are essentially based on information representation, processing and retrieval. The process of Information Representation and Retrieval (IRR) involves three primary stakeholders – the users, the intermediary (submitters, editors and content managers in open access retrieval system) and the information retrieval system. These three intertwining elements act jointly in developing and functioning of IRR system. However, in any type or size of IRR setup (including IRR for open contents), the last primary element i.e. information retrieval system consists of four major components – **the database** (includes information represented and organized through systematic process – both metadata and full-text objects); **the search mechanism** (determines how information stored in databases can be retrieved); **the language** (a crucial component in information representation and query formulation that can either be natural or controlled language; determines specificity, flexibility and artificiality in IRR); and **the interface** (that allows users to interact with the IR system and thereby represents human dimension in IRR).

Open access knowledge systems (such as Open Access Journals and Open Access Repositories) are essentially information representation and retrieval (IRR) systems where full-text knowledge objects are stored and made available for open and free access to the end users.

3.2.1 Retrieval: From Conventional to Neo-conventional System

Information representation is essential pre-requisite for information retrieval. Open knowledge objects, irrespective of forms and formats, need to be represented in a standard manner before it can be retrieved. Quality of information representation has direct impact on retrieval efficiency. This aspect has drawn attention of the experts in the field over the ages. Now, the processes of information representation and retrieval have changed fundamentally with the advent and application of ICT. Open access contents are no exceptions. Before making a quantum jump into retrieval of open contents, let's have a brief discussion on evolution of information representation and retrieval (IRR).

The term *information retrieval* was first coined by Calvin Mooers in 1952 but research and development on Information Representation and Retrieval (IRR) started right from the time of Panizzi. The conventional processes of information representation include two major activities – a) Identification and extraction of elements (concepts important for retrieval) from the documents e.g. keywords, phrases etc. representing the concepts; and b) assignment of terms (appropriate for retrieval) to a document e.g. descriptors or subject headings. Information representation, in other words, is a combination of these two processes and is an array of activities like indexing, abstracting, categorization, summarization etc.

Indexing

Indexing is a widely adopted method for information representation. It involves selection and use of terms (derived or assigned) to represent important facets of the original document (bibliographical or full text). Indexing may be grouped on the basis of how the terms are obtained.

- **Derivative indexing**: Here, terms are extracted from the original documents. It can also be treated as similar to keyword indexing and there is no control of the terms. As a result, there is no need of any vocabulary control mechanism either at the indexing or at the retrieval stage.
- **Assignment indexing**: Here, terms are assigned to represent the documents. Scheme(s) of controlled vocabulary is/are used for choosing appropriate terms which can also be used at the time of search.
- **Automated and Automatic indexing**: The automatic indexing was developed by H.P.Luhn with the invention of Key Word in Context (KWIC) indexing system and subsequently the methods developed by him primarily using statistical techniques. Presently, mechanical activities related with indexing (such as alphabetizing, formatting, chronological sorting etc.) can be done by using computers but intellectual activities are

accomplished by human beings though various methods are being experimented for selection of terms from the documents without human efforts. If computers are applied only for mechanical operations of indexing and human indexers are employed for intellectual activities of indexing, we call it automated indexing. If computer systems are applied to perform both mechanical and intellectual operations, we call it automatic indexing (also termed as machine indexing).

- **Hyper structure indexing**: In Web environment, index terms are recorded as hyperlinks that embody both the index terms and the locator mechanism (Chu, 1997). In other words, indexing of Web documents uses hyperlink names as index terms and help users to locate index terms in hyper documents.

IRR for organizing open access resources utilizes all four major indexing methods as mentioned above. It extracts terms from the body of knowledge objects through derivative indexing, manages metadata assigned by indexer, sorts and arranges browsing keys by automated process, and highlights and hyperlinks search elements (keyword or phrase) by hyper structure indexing.

Categorization

In library world, it is another widely adopted method for information representation. In simple words, it may be termed as successive and hierarchical representation of information by categories. We generally use established classification schemes (e.g. DDC, LCC, CC etc.) for traditional information resources. But in Web environment where documents are of mixed quality, huge in quantity and ephemeral, application of the library classification scheme(s) for information representation becomes expensive and inappropriate. Categorization of Web documents is done by taxonomy based on loosely structured categories. Most of the institutional repository software support organization of open contents through the use of subject taxonomies.

Summarization

It is the process of developing condensed copy of the original document. Different types of summarization are possible on the basis of degree of condensation. These are – Abstracts (a concise and accurate representation of the contents of a document), Summaries (a restatement of the main points of the original document) and Extracts (comprises one or more selected part(s) of a document to represent it). Application of computers in summarization is fairly successful for extracts e.g. Internet retrieval systems like Google, Altavista, NorthernLight etc. employed auto-extraction process for information representation. But computerized summarization is moderately successful for auto-summary and not at all satisfactory for auto-abstracting.

Citation indexing

Citations are bibliographical information about documents, and therefore can be considered as a source for information representation. As a result, citations can be used as means of information representation by citing authors for their

own publications. Eugene Garfield introduced this method of information representation through publication of citation indexes. Citation indexing can be carried out entirely by computers without human intervention. Most of the open content search services like Google Scholar (includes both open access and restricted-access document), OAIster[85], and OAN-Search[86] include "*Cited By*" as value-added feature of retrieval systems on the basis of citation indexing methods. Moreover, in some retrieval systems citation frequency is a major parameter for evaluation of the quality of the documents.

String indexing

It is based on the concept of representing a document by a suitable phrase or a statement, or in some cases by a group of phrases. String indexing is a special kind of indexing. There are different types of string indexing (e.g. Chain indexing, PRECIS, POPSI, NEPHIS etc.) and each of these systems includes two basic steps – a) human indexer creates an input string to summarize the content/theme of a document; and b) computer generates index entries from input string on the basis of rules of string indexing system. String indexing, as an integration of manual selection of input string and computer generated index entries, is particularly useful for generating printed index and not quite an attractive option in information representation for digital open contents. Chu (2009) framed a comparison chart for conventional methods of information representation on the basis of four parameters namely types and entity of representation, framework of representation and production mode. The chart (Figure 21) framed by Heting Chu is quite helpful in selecting suitable method(s) for different purposes.

Methods / Features	Indexing		Categorization		Summarization			Others	
Representation Type	Derivative	Assignment	Classification	Taxonomy	Abstracts	Summary	Extracts	Citation	String
Entity Represented	Part		Whole		Whole		Part	Whole	
Representation Framework *	No	Yes	Yes	No	No			No	Yes
Production Method	Automatic	Manual & Automatic	Manual	Manual & Automatic	Manual	Manual & Automatic		Automatic	Automated

Figure 21: Chu Framework for information representation

(* Controlled Vocabulary, if any)

Retrieval systems related to open contents are also following the above-mentioned four conventional processes including neo-conventional processes like citation indexing, string indexing etc.

[85] http://www.oclc.org/oaister

[86] http://oansuche.open-access.net

Full text information representation

The advent of ICT in general and storage technology in particular over the last decade made full text representation of digitally stored objects a bit easier. Fugmann (1993) advocated that full text representation should avoid two extremes i.e. “every word a descriptor” and “no indexing is necessary”. Most of the open contents retrieval systems are full text information retrieval systems. These systems generally have two levels of information representation. The first level contains metadata representation (see unit 2 for resource description) and second level that includes full text representation. Presently, almost all the open source institutional repository software (such as DSpace, Greenstone, E-Print archive) support full text representation including generation of thumbnail image of the format (e.g. PDF, HTML, ASCII text, MSWORD etc.) in which the full text information object is available within the system. These software also support automatic association of the format with appropriate software for display and reading of the full text document. Representation of full text is a sort of derivative indexing, where retrieval software can extract keywords automatically after exclusion of junk words (on the basis of a predefined list of stop words). Naturally, full text information representation and retrieval systems are limited by low precision, high recall and cross-disciplinary semantic drift. These problems of full text retrieval are under active investigation by researchers working in the domain of AI (artificial intelligence), NLP (natural language processing), and semantic Web.

Multimedia information representation

Open access resources do not contain only textual materials (although the percentage of textual resources is still very high in all types of digital content management systems). The domain of OA is increasingly populated by slides, MP3 files, video clips, animated pictures, photographs etc.

Figure 22: Open Access Biomedical Image Search[87]

On the other hand, a single open digital object may contain text, image, video, and audio in linked environment. You already know from the previous unit the use of OAI-ORE in sharing and exchange of compound digital objects. These compound digital objects are also called multimedia digital object and retrieval

[87] http://openi.nlm.nih.gov/index.php

processes are different from textual retrieval systems. Multimedia information retrieval systems for open contents are maturing day-by-day. For example, Open-i project of NLM (National Library of Medicine, US) aims to provide image search service for open access biomedical resources (Figure 22). It includes biomedical articles from the full text collections such as PubMed Central and retrieves both the text and images in the articles. The support is provided on the basis of extensive image analysis & indexing and deep text analysis & indexing.

3.2.2 Retrieval Approaches

Retrieval approaches may be categorized as structured retrieval and unstructured retrieval. As a whole, the retrieval methods as classified by Luhn (1958) are:

- *Browsing:* Retrieval of information by look-up in an ordered array of stored records;
- *Searching*: Retrieval of information by finding/locating in a non-ordered array of stored records; and
- *A combination of searching and browsing.*

As you know, searching is the prime retrieval approach for most of the IR systems. Fenichel & Hogan (1981) identified a total of four types of searching. These four basic search strategies are quite relevant for retrieving open contents. These are

- **Building block approach**: It starts with a single concept search. In case of a complex query, this strategy advises to decompose search statement into required number of single concepts and then integrating retrieved result sets through appropriate search operators. This strategy is very helpful for novice users.
- **Snowballing approach**: This strategy advices searcher to conduct a search first and then modify the search query on the basis of the retrieved results.
- **Successive fraction approach**: This strategy advices searcher to start a search with a broad concept and then narrow down the search by applying different limiting techniques.
- **Most specific facet approach**: This approach directs that in case of multiple concept query string, identify the most specific term/concept first and conduct search against it.

Convenient approach: If full-text IR users just enter terms by leaving a space in between (space automatically incorporate default Boolean operator e.g. AND or OR) or pick up different filtering parameters (file type, language, year range etc.) from drop down lists. This method is termed as Quick or Convenient search approach.

3.2.3 Retrieval Techniques

Retrieval techniques are search operators or devices that help users in resource discovery through searching. A typical online IR for open contents supports

different retrieval techniques. These techniques may broadly be divided into two groups – basic set and advance set.

Basic Set

These retrieval techniques are supported by most of the information retrieval systems. These are:

- **Boolean operators**: The Boolean search operators help addition of concepts, exclusion of concepts and inclusion of concepts through AND operators, NOT operators and OR operators respectively.
- **Truncation**: Truncation technique supports retrieval of different forms of a term but all with one part in common. For example, DSpace uses * as Wildcard Operator. The characters "arch*" matches with archive, archival, archiving etc.
- **Proximity operators**: These operators help to specify distance between two search terms precisely. DSpace uses tilde symbol, "~", at the end of a phrase as proximity operator. For example, the query *"library science"~3* in DSpace will rretrieve records where the words 'library' and 'science' are separated by three spaces.
- **Case sensitive search**: It helps searchers to specify case of a search term i.e. upper case or lower case.
- **Range search**: It helps in selecting/filtering records within certain data ranges. The search query *author:[rao TO rath]* in DSpace will retrieve documents authored by names that fall between 'rao' and 'rath
- **Field search**: It helps searchers to limit the search in one or more fields.
- **String search**: It is a kind of free-text searching that allows searchers to search those terms that a searcher thinks but have not been indexed.

Advance Set

These techniques are provided selectively in some modern retrieval systems. Most of these techniques are still in research bed and their efficiency level is increasing day-by-day. These are

- **Fuzzy searching**: It is a unique search technique that can tolerate errors committed during data entry or query input. This technique can detect and correct spelling errors, errors related to OCRing and text compression. DSpace uses tilde symbol, "~" for Fuzzy searching. For example, a search by Fredarick~ (a misspelled author name) will retrieve the author named "frederick" (exact name of author).
- **Weighted searching**: The weightage technique helps to assign different weights to search terms during query formulation to indicate proportionality of their significance or the emphasis the user placed upon them. Both symbols (e.g. * in ERIC system) and numerals (e.g. 1 to 10 in GSDL) may be used to indicate relative weighting.
- **Query expansion**: It allows searchers to improve search results by

modifying query string on the basis of retrieved result set.

- **Multiple database searching**: This method helps in searching two or more information retrieval systems simultaneously. It helps to get rid of different query syntax (e.g. use of different symbols for different operators), different encoding standards (e.g. ASCII, Unicode), and different data formats (e.g. MARC, CCF, Dublin Core etc.) in different retrieval systems. Distributed searching on the basis of Z 39.50 protocol is a classical example of multiple databases searching (that use different retrieval software and different data formats like MARC, CCF, UNIMARC etc). For example, the search on author: Ranganathan, S. R. can be forwarded to three or more Z 39.50 servers at the same time (see Figure 23).

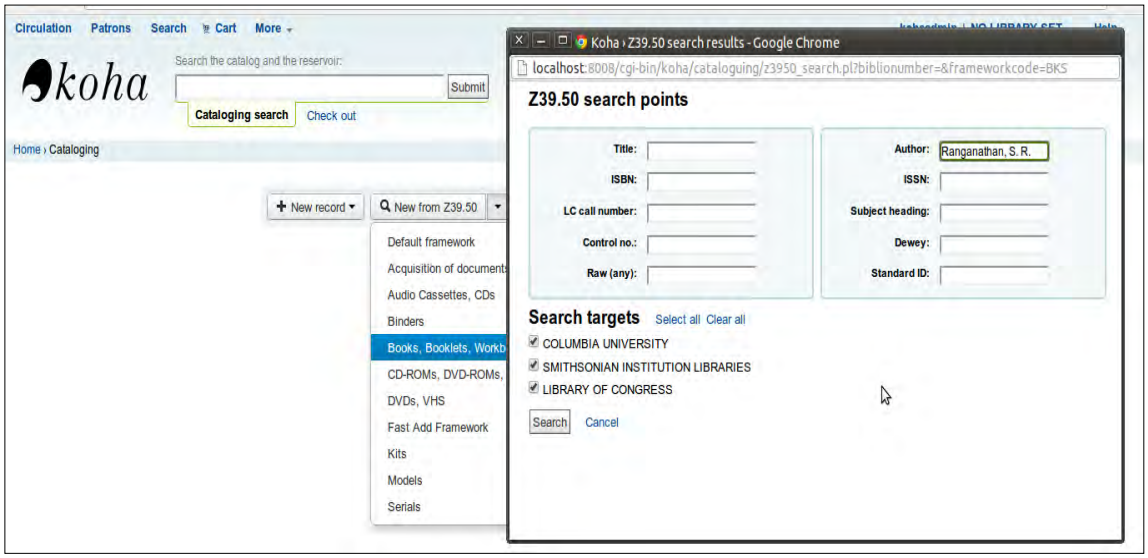

Figure 23: Multiple Database Searching through Distributed Search Protocol

3.2.4 Retrieval Models

IR models are theory based approaches to cover different aspects of information retrieval systems. Different IR models have been developed over the years but matching mechanisms form the basis of all these models. Matching can be done between terms or between similarity measurement (e.g. distance, term frequency etc.).

Term matching

Term matching is a direct matching of terms derived from or assigned to documents, document representation and queries. There are four types of term matching as mentioned below:

- **Exact match**: It means query representation exactly matches with document representation in IR system e.g. case sensitive search and phrase search;
- **Partial match**: In this case part of the term being matched with the document representation in information retrieval system e.g. truncation;
- **Positional match**: It takes into consideration the positional information of

what is being matched in retrieval process e.g. proximity search; and

- **Range match**: It takes into consideration what is being matched in a given range e.g. searching of bibliographic records by publication dates.

Similarity matching

It is an indirect matching process in which final matching is made on the basis of similarity measurement. For example, in Vector Space model matching is based on the distance between vectors or degree of vector angle. Again, in probabilistic model, similarity is measured on the basis of term frequency. It determines the probability of relevance between queries and documents.

Beaza-Yates and Ribeiro_Neto (1999) grouped IR models into two categories – system oriented models and user oriented models. The classification may be represented as in Figure 24.

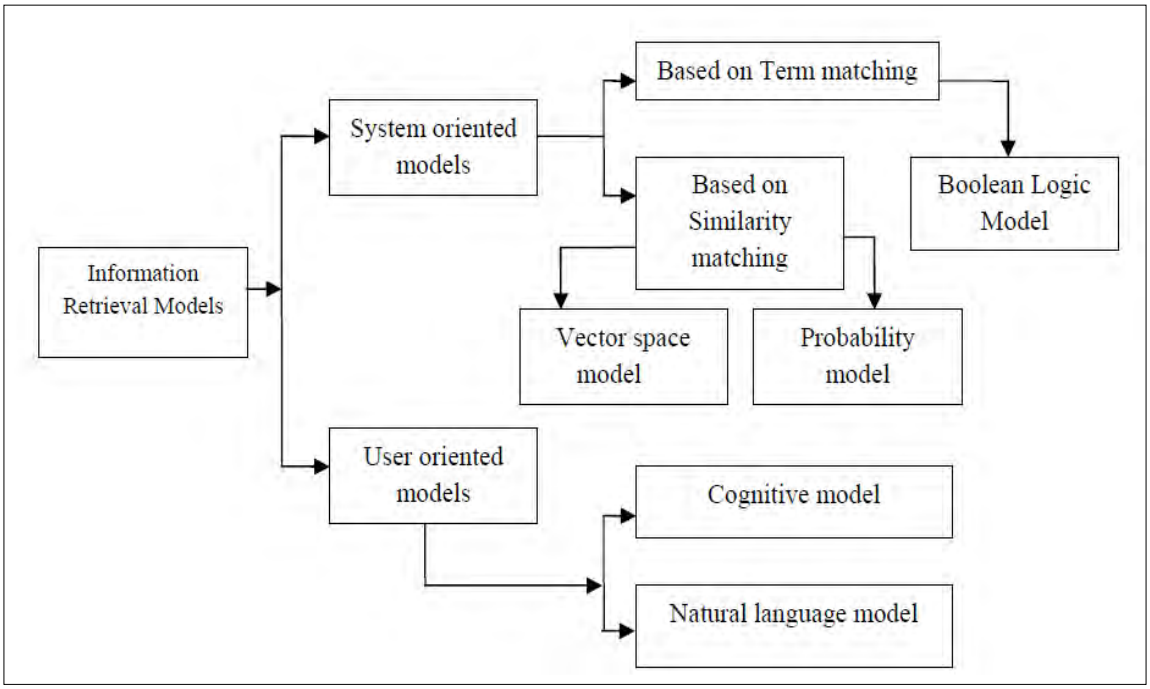

Figure 24: Models of Information Retrieval

Most of the software that manage open access repositories are using open source text retrieval engines like Lucene, Solr MGPP etc. Vector Space Model (or it's modified version) is probably the most common in these retrieval engines. These text retrieval engines (based on Vector Space model) works in the following manner – I) Extract tokens from content or primary bit-stream; ii) Transform extracted tokens on the basis of indexing parameters as set by indexer; iii) Stemming of tokens; iv) Expand with synonyms (to support query formulation); v) Remove tokens which are stop words or junks; vi) Add metadata elements in indexing; vii) Store tokens and related metadata as structured data for search optimization; and viii) Creation and maintenance of Inverted Index. The process of information representation, query formulation and matching is shown in Figure 25.

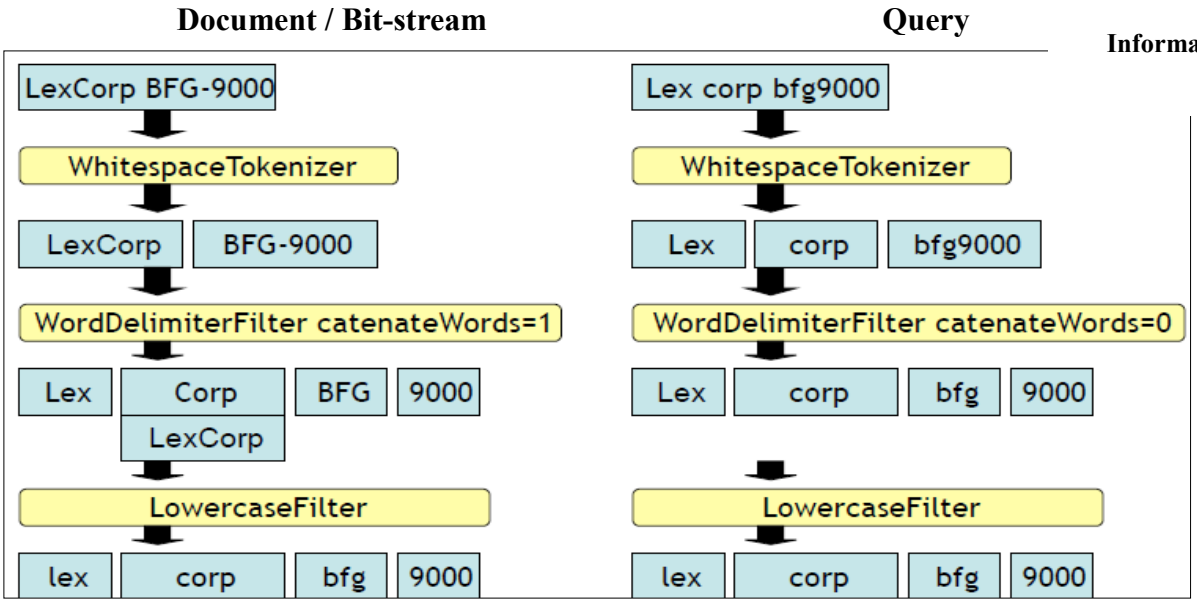

Figure 25: Workflow in Vector Space Model (Source: Yonik Seeley)

A comparative study for three basic information retrieval models may be presented as Figure 26.

Model / Features	Boolean Logic	Vector Space	Probability
Boolean logic	Yes	No	No
Term Weighting	No	Yes	Yes
Ranking of results	No	Yes	Yes
Matching mechanism	Term matching	Similarity matching (Vector distance)	Similarity matching (Term frequency)
Special features	None	Relevance feedback	None

Figure 26: Comparison of basic information retrieval models

3.2.5 Evaluating Retrieval Systems

Researchers (Keen, 1971; Large, Tedd & Hartley, 1999) formulated a common set of evaluation parameters irrespective of any type or size of IR systems. This set is equally applicable for different kinds of IRR (including IR systems related to open access resources) and includes parameters like accuracy (exact representation of original documents through surrogates), brevity (briefness of representation), consistency (uniform representation), objectivity (authentic

description of original document), and other parameters (clarity, readability and usability). These parameters may be termed as Generic measures. Other evaluation measures can be grouped into two categories – measures related to *retrieval performance* and measures related to *retrieval process*. The evaluation measures that concentrate on retrieval performance are as follows:

- **Recall and Precision**: This measure (proposed first by Kent in 1955) is a combination of two factors. The Recall factor measures retrievability of an IR system and Precision factor measures the ability of an IR system in separating the non-relevant from the relevant items. Salton (1992) observed that these two factors, although not quite perfect, have formed the basis for many evaluation projects. There are many extensions of these two factors such as E-measure (Swets, 1969), Average recall and precision (Harman, 1995), Normalized recall and precision (Foskett, 1996; Korfhage, 1997), and Relative recall and precision (Harter & Hert, 1997).
- **Fallout ratio**: This measure, proposed by Swets (1963), is ratio between non-relevant documents retrieved and all non-relevant documents in a system database. The smaller fallout value ensures better IR system.
- **Generality measure**: It is defined as the proportion of documents in a system database that is relevant to a particular topic. Lancaster & Warner (1993) reported that the higher generality number is associated with the easier searching.
- **Single measure**: Recall and precision (including their extensions and modifications) factors are criticized for their incompleteness as evaluation measures. In view of this limitation Cooper (1973) suggested a utility measure on the basis of user's subjective judgment about usefulness of an IR system.
- **Other measures**: Griffith (1986) proposed that only three numbers namely relevant retrieved, non-relevant retrieved and total number of documents in an IR system should be considered in evaluating.

But retrieval performance is not the only factor to evaluate an IR system completely. The evaluation studies for an IR system are again designed in different ways by different researchers considering different evaluation parameters. A sum up table may be designed to list common evaluation parameters for open access IR systems (Table 10).

Table 10: Evaluation Criteria for OA Retrieval System

Open Access IR System	Evaluation Criteria
General	1. Coverage (types of documents, number of documents, update frequency, retrospectiveness) 2. Recall and Precision (an optimal point for both recall and precision is required) 3. Response time (time lapse between submission of query and return of results) 4. User effort (ease of learning the IR system) 5. Output (flexibility in forms and formats for display and obtaining of results)
Specific	6. Index composition (index methods, query handling, extent of indexing i.e. title, first paragraph, full document, coverage, update frequency, cache version availability etc.) 7. Search capability (Boolean search, fuzzy search, phrase search, positional and relational operators, truncation, filtering etc.) 8. Retrieval performance (a combination of three factors i.e. recall, precision and response time) 9. Output (a sum of three perspectives i.e. accessibility, contents and formats including format interoperability) 10. User effort (structured online help, proper documentation and appropriate use of icons in interface) 11. Other factors (multilingual search, cross language search, clustering of results, passage retrieval, web 2.0 tools like RSS, Faceted navigation etc)

Many projects have been accomplished to evaluate different types of IR but till date we don't have any specific evaluation study related to OA retrieval systems. However, TREC (Text Retrieval Conference) and FIRE (Forum for Information Retrieval Evaluation) initiated some evaluation studies related with OA retrieval systems such as TREC TRACK-8 for Web-enabled and Integrated IR, TREC TRACK 10 for video retrieval and Federated search TREC. The major evaluation projects may be categorized under two groups – Accomplished projects (Table 11) and Ongoing projects. In the first group, Cranfield tests may be considered as the most influential and in the second group, TREC (Text Retrieval Conference) is the most comprehensive evaluation project in the history of IRR.

Table 11: Accomplished Retrieval Evaluation Projects

Project, Year of origin and Personnel/Organization	Objectives and Methods	Outcome
Cranfield I; 1957; C.W. Cleverdon Ref:Cleverdon, C.W. (1962). Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Cranfield: College of Aeronautics	Objectives To compare effectiveness of four indexing systems (Subject Heading List, UDC, Faceted Classification, Uniterm coordinate indexing) Method 100 documents indexed five times by three indexers in the field of aeronautics; Developed a set of 18000 index entries; Searched by 400 queries in 3 rounds (1200 queries) designed by users.	• Subject background/knowledge was not a significant factor for indexing • There is an inverse relationship between recall and precision • A 1% increase in precision could be achieved at a cost of 3% loss in recall • Uniterm system outperformed three controlled vocabulary based systems • Increased time in indexing would not necessarily increase recall • Faceted classification based system performed poorly in comparison with other three systems
Cranfield II; 1967; C.W. Cleverdon Ref: Cleverdon,C.W. and Mills, J. (1963). The testing of indexing language devices. ASLIB proceedings, 15(4), 106-130	Objectives To assess effects of different indexing devices (synonyms, generic relations, coordination links, and role) on retrieval performance Method 200 research papers in the field of aerodynamics were gathered each having 5-10 references; Each author was asked to frame one question related with the main area of the paper and three subsidiary questions related with the investigation;	• Single term indexing languages were superior in comparison with others • When single terms were used for indexing, the inclusion of collateral classes (quasi-synonyms in particular) reduced retrieval performance • When concepts were used for indexing, inclusion of superordinate, subordinate and collateral classes reduced retrieval performance • When controlled terms were used, inclusion of narrower and broader term reduced

	Each author also asked to judge relevance of cited papers against their questions in a 5 point scale; A total of 1400 documents (papers + cited documents) and 221 questions were finally selected; Each selected document indexed in three ways (concept recorded in natural language, single words in each concepts were listed and concepts were combined to form the main themes of the documents); Each term given weight to indicate its relative importance; Searching was done by single term, simple concept and controlled term index language.	retrieval performance • Index languages formed on the basis of titles performed better than those formed on the basis of abstracts • Best performing index languages were composed of uncontrolled single words derived from documents

Apart from these two major IRR evaluation projects, there were SMART project conducted in 1964 (Salton, 1981), MEDLARS project in 1967 (Lancaster, 1968) and STAIRS in 1985 (Blair & Maron, 1985) for evaluating IR systems.

Ongoing Projects

The TREC (Text Retrieval Conference) is an ongoing evaluation project jointly sponsored by NIST and DARPA. The TREC structure includes two major categories – CORE (main activities of TREC) and TRACKS (subsidiary activities of TREC). CORE category of TREC experiments are again divided into two groups - Ad hoc (related to retrospective retrieval) and Routing (related to SDI type services). Ad hoc retrieval search is an unknown item search where the user is not aware of the existence of the documents and wants to retrieve them. Such kind of search produces a ranked list of items from databases. On the other hand in routing search user's interest remains stable but the document set changes. Such a search is useful for researchers who want to keep track of the latest developments in their field of interest. In ad hoc search, an IR system searches a static set of documents using new questions. In routing IR system it makes a decision whether or not a particular document is of relevance to the user's query. It produces an unordered set of documents. The area of major retrieval experiments (TRACKS) of TREC are as given in Table 12.

Table 12: Major areas of Text Retrieval Conference

Conference	TRACKS
TREC-1	Bibliographic data structuring and system engineering
TREC-2	Natural Language Processing (NLP) and Automatic query representation/formulation
TREC-3	Interactive system design and Query formulation in multiple databases
TREC-4	Problems of short user statements
TREC-5	Information retrieval of non-English languages (non-Roman scripts representation and encoding)
TREC-5	Information retrieval of non-English languages
TREC-6	Cross-language and spoken document information retrieval
TREC-7	Large query formulation
TREC-8	Web-enabled and Integrated IR
TREC-9	IR related to image and NLP interface
TREC-10	IR related to video objects
TREC-11	Fine tune searching within the ranked set of documents
TREC-12	IR specific to bioinformatics and genomics

Apart from TREC, there are some other ongoing IR evaluation projects like

- CLEF[88] (Cross-Language Evaluation Forum)
- NTCIR[89] (NII Test Collection for IR Systems) Project
- Chinese Web test collection[90]
- FIRE[91] (Forum for Information Retrieval Evaluation)

[88] http://www.clef-initiative.eu/
[89] http://research.nii.ac.jp/ntcir/index-en.html
[90] http://net.pku.edu.cn/~webg/cwt/en_index.html
[91] http://www.isical.ac.in/~clia/

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this unit.

1) What is a Retrieval Model? What model do you think is suitable for OA retrieval?

……..……..……...

……..……..……...

……..……..……...

2) What is the role of TREC in OA retrieval?

……..……..……...

……..……..……...

……..……..……...

3.3 RETRIEVAL OF OPEN CONTENTS: A STATE-OF-THE ART REPORT

Users generally perform information retrieval tasks in three ways. These are searching, browsing and a combination of searching and browsing. Searching is a structured retrieval process. It intends to find out the resources that would match with the query terms by using available retrieval techniques. On the other hand, browsing is finding and selecting resources by skimming and scanning. Browsing did not receive much attention in the regime of online IR (dominated by commercial database vendors and database aggregators) because of high connection charges. But it started getting attention with the advent of CDROM based IR and gained popularity in Web based IR. Open access retrieval systems are essentially Web-enabled IR and support all these three retrieval tasks.

The role of these three basic retrieval tasks may be understood in a better way through an analogy. Koll (2000) proposed a structured analogy between information retrieval and finding needle in haystack (Figure 27). In this proposition, needle stands for information resources and haystack represents IR system. Koll enumerated a total of *twelve possibilities*, which can be matched with three retrieval methods i.e. searching, browsing and combination of searching and browsing.

1	A **known** needle in a known haystack	Searching
2	A **known** needle in an unknown haystack	
3	An **unknown** needle in an unknown haystack	
4	**Any** needle in a haystack	
5	The **sharpest** needle in a haystack	Searching & Browsing
6	**Most** of the sharpest needles in haystack	
7	**All** the needles in a haystack	
8	Affirmation of **no needles** in the haystack	
9	Things **like** needles in any haystack	Browsing
10	Let me know **whenever** a new needle shows up	
11	**Where** are the haystacks	
12	Needles, haystacks – **whatever**	

Figure 27: Koll's Analogy

Almost fifty years back, Luhn (1958) first grouped retrieval methods as – i) Retrieval of information by look-up in an ordered array of records; ii) Retrieval of information by search in a non-ordered array of records; and iii) Combination of I and II. The first and second approaches represent browsing and searching respectively. The third approach of Luhn is an integration of searching and browsing. This integrated approach holds key to successful retrieval in digital IRR including OA retrieval. The following section covers different aspects of these two retrieval methods including the applications of text retrieval tools in retrieving contents from local repositories and retrieval features of global open search services.

3.3.1 Organization of Open Contents

Information Representation and Retrieval (IRR) activities entered into digital age with the advent of ICT in general and the Web in particular. ICT influenced the design and development of four major components of any IR systems (any type or size) including OA retrieval. These components are database, search process, language of IRR and user interface. OA retrieval system is essentially based on database and language of IRR at the core. The search processes support matching of search queries and documents on the

basis of metadata and contents of documents through an intuitive user interface.

Database

Databases form the core of Web-enabled OA retrieval system. Bibliographic database technologies exclusively deal with textual objects. Traditional bibliographic databases (online and CDROM databases) include two parts. The first part is sequential file (field-record-database) and the second part is inverted file (indexes to sequential file). On the other hand, Web-enabled IR systems also contain two parts but the sequential files are generally made of field-less information entity (i.e. full-text resources in Web page (HTML, XML) format, PDF format etc).

Search process

Database determines what can be retrieved from the OA retrieval system, whereas search mechanism determines how open access resources stored in databases can be retrieved. It provides search algorithms and procedures for retrieving open contents. Generally search mechanism of an IR system provides two sets of retrieval techniques – basic retrieval techniques (Boolean, relational and positional search operators) and advance techniques (weighted searching, fuzzy searching, term boosting, soundex search, relevance ranking etc.). Text retrieval engine plays a pivotal role here in OA retrieval.

Language of IRR

Search mechanism determines what retrieval techniques will be available to searchers for retrieving open contents, whereas language of IRR, to a great extent, determines the flexibility in information representation (metadata encoding and content description by library professional) and query representation (query formulation by searcher). Language in IRR may be grouped as natural language and controlled vocabulary (classification schemes, subject headings list and thesauri). The debate about natural language vs. controlled vocabulary is an ongoing event in IRR for many years.

User interface

It is a layer of interaction between users and IRR activities in an OA retrieval system. The utility of user interface depends on mode of interaction, display features, online help, provision of feedback, availability of statistics, web 2.0 supports to ensure participation and collaboration, RSS feeds etc. It is considered as the human dimension of IRR. The components of an OA retrieval system and their relationships may be illustrated as below (Figure28).

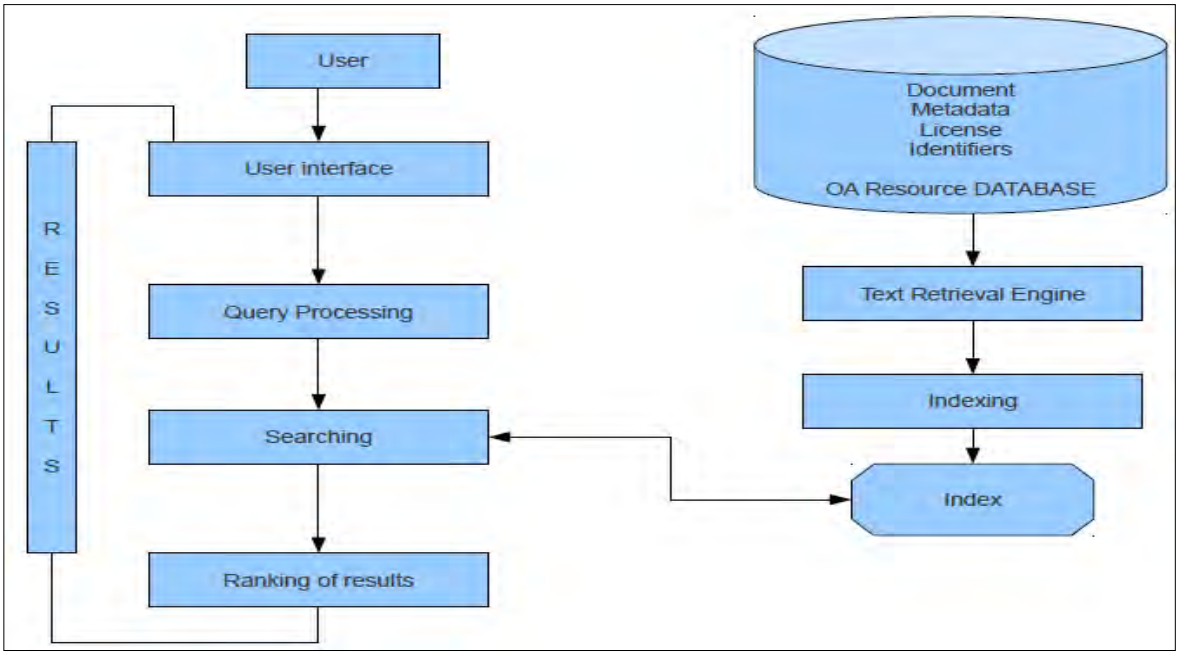

Figure 28: Components of an OA Retrieval System

3.3.2 Retrieving Open Contents: Problems and Prospects

The toll based scholarly communication process limits, rather than expands the wide availability and global sharing of research resources. In such a research communication process, research publications are also obliterating their institutional origins. Exorbitant increase of journals prices and resultant subscription cancellations is affecting readership considerably. Libraries and academic communities in developing countries are worst affected. In the age of print publication, open access was physically, economically and technically impossible. But thanks to the distributed information system in general and Web in particular, OA is an emerging reality for providing viable alternative to toll-based system. OA movement promotes availability of scholarly communications in public domain through digital publishing system and thereby offers an unprecedented public good: the free online availability of peer-reviewed scientific and scholarly digital resources. The obvious advantages of OA are the widespread sharing of knowledge and the acceleration of research. OA repositories and OA journals are both practical and lawful. The emergence of OA services around the world are proving that OA can do much better than traditional subscription-based journals in their cost-effectiveness and service to science and scholarship. Moreover, OA retrieval systems are adding values in services through personalized alert services, federated search for distributed open access repositories, e-SDI service notifying a user the availability of new open contents, ontology-driven retrieval, usage data and statistics, citation linking, aggregation of OA resources by multiple logical approaches (discipline-wise, country-wise, institutional group-wise etc).

But, at the same time, organization of OA resources involves some serious problems. The basic problems in retrieval of open contents may be summarized as follows -

Distributed OA resources at global scale

OA resources are available under different software, represented by different metadata sachems, and distributed in different types of services across the globe. Till date there is no comprehensive listing of OA resources subject-wise, language-wise, country-wise (although DOAJ, DOAR and ROAR are providing basic lists).

High percentage of volatile OA resources

Like other Web resource OA resources are volatile in comparison with their commercial counterparts. The change of URL of OA repositories, disappearance of OA journals, non-availability of persistent URLs for most of the resources, no universal standard for unique author identification, missing hyper-links are some of the serious problems in organization and retrieval of open contents.

Large volume of OA resources

OA resources are increasing rapidly in magnitude and in variety but organizing capabilities of search services are failing to keep pace with such geometric growth. For example, a multimedia and multilingual OA resource requires fundamental restructuring of retrieval mechanisms. BASE, an exclusive search service for OA resources, recently reported coverage of 52 million OA resources.

Unstructured OA resources and datasets

Most of the open content service providers like OA repositories and OA journals are not quite serious in policy formulation and in following standard metadata encoding rules, metadata element refinement (e.g. DC. Date may represent date of publication, date of modification, date of uploading etc; therefore element refinement like DC.Date.publication is required for effective organization of OA resources).

Redundant OA resources

As multiple deposit standards (like SWOARD, CRIS-OAR, OA-RJ) are not quite matured yet, authors tend to submit OA resources in many OA retrieval services and thereby leading to redundancy of OA resources. This results in placing unnecessary loads on retrieval systems.

Quality of description datasets for OA resources

Most of the repositories apply simple DCMES (Dublin Core Metadata Elements Set) for describing all sorts of OA resources like journal papers, technical reports, research datasets, thesis and dissertations, images, learning

objects, video objects etc. But these special types of resources require domain-specific metadata schemas for describing specific attributes of resources like ETD-MS for describing thesis and dissertations, VRA-Core for image resources, IEEE-LOM for learning objects.

Heterogeneous OA resources

Heterogeneity is the norm in OA. These resources differ in formats, forms, degree of complexity, nature of contents, metadata standards, software in use, back-end database, character encoding, degree of completeness in metadata, supports for interoperability standards and so on. These differences affect efficiency in retrieval considerably.

3.3.3 Retrieval Facilities in Gold OA and Green OA

The OA retrieval systems may broadly be categorized into three major groups on the basis of services rendered by these entities. The major groups are – i) Path Finder services; ii) Federated search services; and iii) Localized search services. The first two groups of services are mostly operating at global scale. The third group of services are developed and maintained by OA publishers, institution-specific repository managers, subject-specific repository managers and volunteer groups. These three broad groups of services are generally using three groups of retrieval utilities – i) using utility of global search engines; ii) use of own search engines; and iii) use of open source text retrieval engines. The above structure may be represented in Fig 29.

	Path Finder Services	Federated Search Services	Localized Search Services
Utility of global search engines	openDOAR	NA	NA
Use of own search engines	DOAJ DOAB OATD	BASE (Bielefeld Academic Search Engine) OAIster Database	Most of the OA journals listed in DOAJ
Use of open source text retrieval engines	SHERPA/RoMEO and SHERPA/JULIET	CORE (COnnecting REpositories)	Most of the OA repositories listed in DOAR and ROAR

Figure 29: OA Retrieval Systems and Services

This is just an illustrative list of OA retrieval systems under different categories. You may consult Wikipedia[92] and OAD[93] for a comprehensive list. As it is not possible to discuss here all the OA retrieval systems, the following section provides you brief overview on facilities and services of major OA retrieval systems.

Directory of Open Access Journals[94] (DOAJ)

DOAJ provides path finder service to quality controlled Open Access Journals. DOAJ started with the directory services only and later extended retrieval service to search contents of many OA journals listed in DOAJ. It means from DOAJ search interface users can search OA journals at content level. In 2013, DOAJ celebrated its tenth year of operation and the number of articles accessible through the Directory surpassed 1.6 million. DOAJ uses it own search engine for retrieval of contents at two levels – search journal title and search journal articles. It provides (Figure 30) two search interfaces – simple (with provision to search keywords) and advanced search (with provision of using fielded search, Boolean operators, range search etc). It provides no scope for sophisticated search operators like term boosting, fuzzy searching, multilingual search,

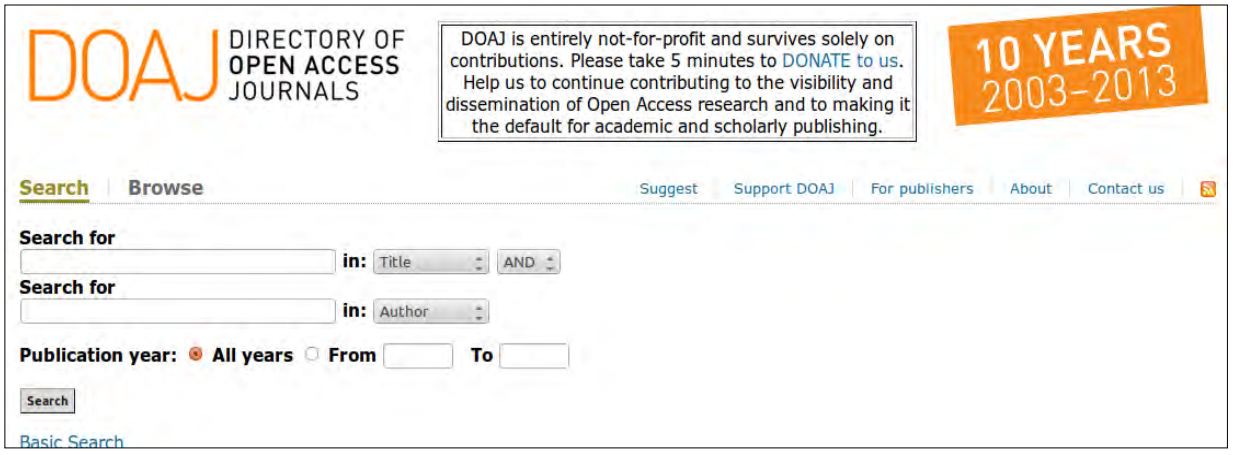

Figure 30: OA Retrieval System of DOAJ

OpenDOAR[95]

It is a directory of academic open access repositories, maintained by the University of Nottingham. This OA service lists institutional and subject-based repositories, while also providing a service to search the contents of these repositories. It is an authoritative worldwide directory of academic open access repositories with over 2200 listings.

[92] http://en.wikipedia.org/wiki/List_of_academic_databases_and_search_engines
[93] http://oad.simmons.edu/oadwiki/Main_Page
[94] http://www.doaj.org
[95] http://www.opendoar.org/

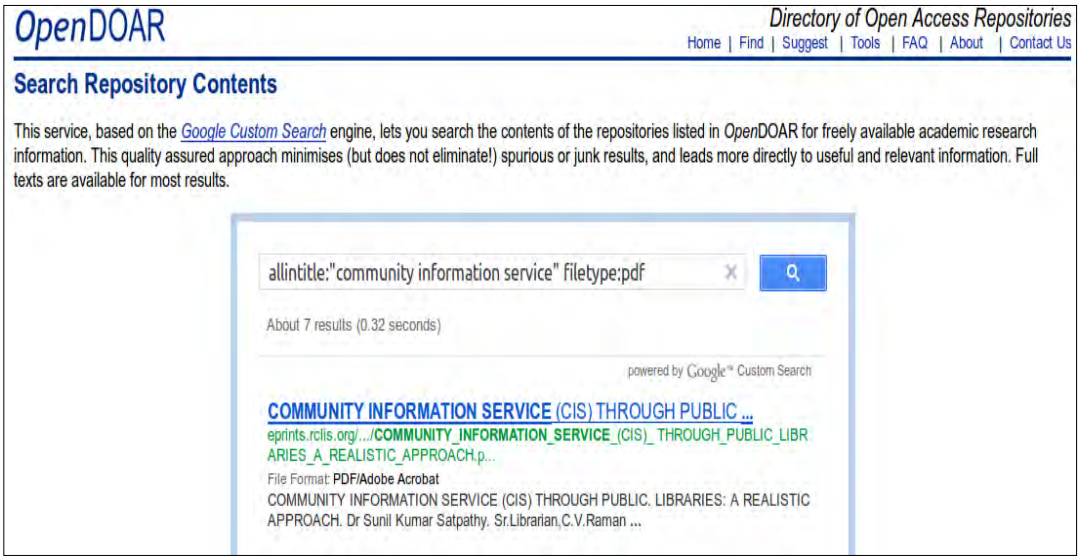

Figure 31: Google CSE Based OA Retrieval System of Open DOAR

OpenDOAR stared with a simple repository listing of OA repositories but later on started providing content retrieval service by using Google custom search service (CSE). It is now possible to use OpenDOAR to search for repositories as well as to search repository contents[96]. As it is using one of the most comprehensive generic search engine, an array of special keywords of Google search are available for fine tuning the query representation for efficient content retrieval. These special search operators are phrase search (e.g. "open access journals"), Boolean operators ("open access" AND benefits), allintitle (for multi word search in title filed only e.g. allintitle:"open access journals"), intitle (single word in title e.g. intitle:ORCID), filetype (format of file e..g filetype:pdf), site (to retrieve documents from a specific domain e.g. allintitle:"open access journals" AND site:.ac.in), related (to find sites that are similar to a URL e.g. related:opendoar.org), link (to find pages that link to a certain page e.g. link:eprints.roar.org) etc. The retrieval of contents from OA repositories from openDOAR interface by using special keyword is given in Figure 31.

Directory of Open Access Books[97] (DOAB)

DOAB is a retrieval service of academic, peer-reviewed books from a variety of publishers and available under an Open Access license. It is a service of OAPEN Foundation. This OA service was launched in July 2013. Presently, it contains over 1600 OA books and resources growing at a rapid rate. DOAB supports libraries to integrate the directory into OPAC, helping library users to discover the books. DOAB also supports metadata harvesting through OAI-PMH interoperability standard. Service providers and libraries can harvest the metadata of the records from DOAB for inclusion in their collections and catalogues. The retrieval facilities of DOAB is quite simple and supports only keyword based search. It also provides limited browsing facilities.

[96] http://www.opendoar.org/search.php
[97] http://www.doabooks.org

Open Access Theses and Dissertations[98] (OATD)

OATD is a valuable retrieval tool for open access graduate theses and dissertations published around the world. Metadata sets of ETDs come from over 800 colleges, universities, and research institutions across the globe. OATD currently indexes 1,839,584 theses and dissertations.

Figure 32: OA Retrieval System of OATD

It provides two levels of search interfaces – simple and advanced (Figure 32). This OA retrieval service supports fielded search, Boolean operators and other sophisticated search operators but recent advances in retrieval technology like relevance ranking, fuzzy searching, term boosting are not available in OATD. The use of Web 2.0 utilities are also missing in OATD.

BASE[99] (Bielefeld Academic Search Engine)

BASE (is one of the world's largest retrieval services for academic open access web resources. It also supports an array of sophisticated search operators and end user services. In 2001, Bielefeld University Library started development of federated search service for OA contents on the basis of OAI-PMH interoperability standard and Bielefeld Academic Search Engine (BASE, http://base-search.net) finally appeared in public domain in 2004. Presently, BASE indexes more that 52 million OA resources at global scale (number of documents: 52,615,190; number of content sources: 2,776 as on 18.11.2013). BASE provides two interfaces (a single search field and an advanced search with multiple search fields and sophisticated search options). But the real achievement of BASE is development of Automatic Enhancement of OAI Metadata (AEOM). This AEOM mechanism helps in assigning Dewey Decimal Classification numbers to documents indexed by BASE automatically (Figure 33).

[98] http://oatd.org/
[99] http://www.base-search.net/)

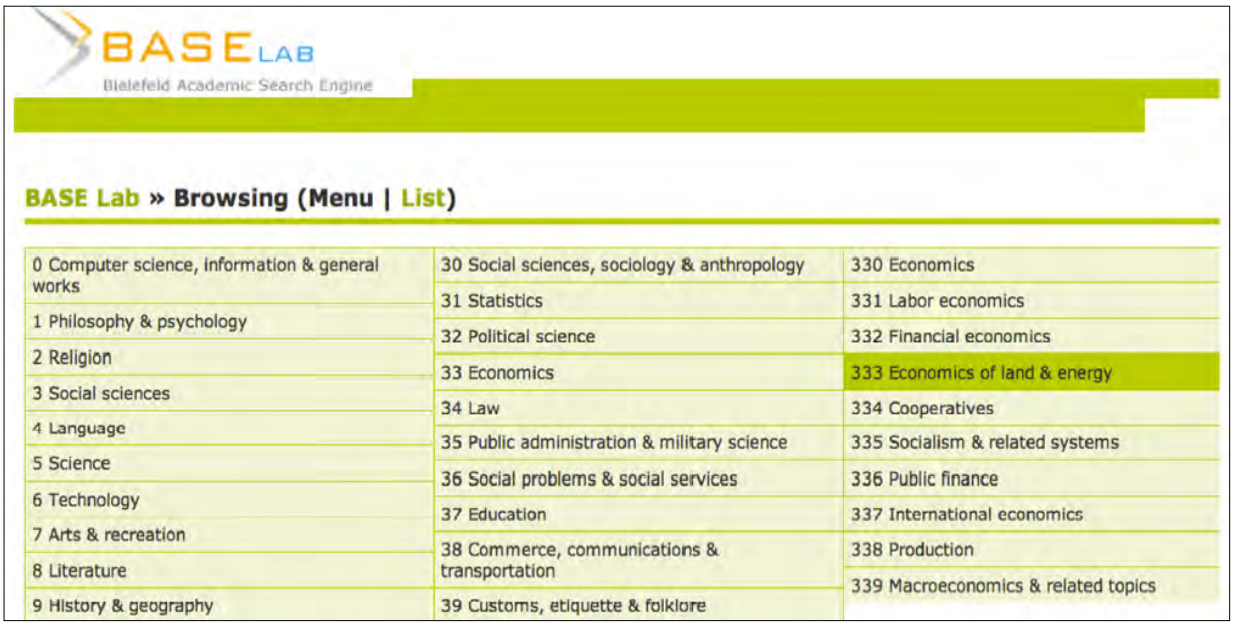

Figure 33: Browsing by DDC in BASE Retrieval System

Other major features of this premier OA retrieval system are -

- *Multilingual*: Multi-lingual search through integration of Eurovoc (end users can search for synonyms and translations from a dataset containing 239,000 terms from 21 languages);
- *Multi-modal*: Automatic redirection to mobile website and support for all modern platforms like Android, IOS, Windows Phone;
- *Multi source*: About 75% of the indexed documents in BASE are OA resources, the rest can be accessed up to metadata level.
- *Multi operators*: Supports for all basic and advanced level search like fielded search, wild card, truncation, range search, positional operators and relational operators;
- *Ranking:* Sorting of results is by relevance (determined by occurrence of the search term in the title or in the metadata);
- *Search refine*: Search results can be refined by author, subject, DDC (classification), year of publication, content source, language and document type.
- *Search history*: Search history for the last ten search queries are displayed, along with the number of retrieved hits;
- *RSS feed*; Creates an RSS Feed for each query;
- *Browsing:* Two kinds of browsing is supported - by Dewey Decimal Classification (DDC) and by document types;
- *Search plug-in*: Provides search plug-in for BASE so that users can directly access search toolbar in browser at user end;
- *Personal Search Environment (PSE)*: Users can create PSE to add favorites

and to save search history permanently;

- *API:* An application programming interface (API) exists which allows integrating the BASE index into local search services like library OPAC;
- *Zotero interface*: Supports transferring results from BASE to Zotero (an open source citation management software) through add-on;
- *User interaction*: Users can correct existing DDC class number or suggest DDC class numbers for unassigned contents;
- *Filtering:* Advanced search interface provides scope to filter results by document types, geographic area and year range;
- *Display*: Users can control ranking of results by number of options (by relevance, by author, by title, by chronological order etc);

BASE is a feature-rich OA retrieval system (Figure 34) and is acting as model for other such services. BASE is a perfect combination of Vector-space information retrieval model and its integration with controlled vocabulary (Eurovoc) and subject access system (DDC). Moreover, BASE provides facilities to integrate BASE search interface within a local open access repository through an easy-to-implement API and thereby leading to globalization of OA retrieval system.

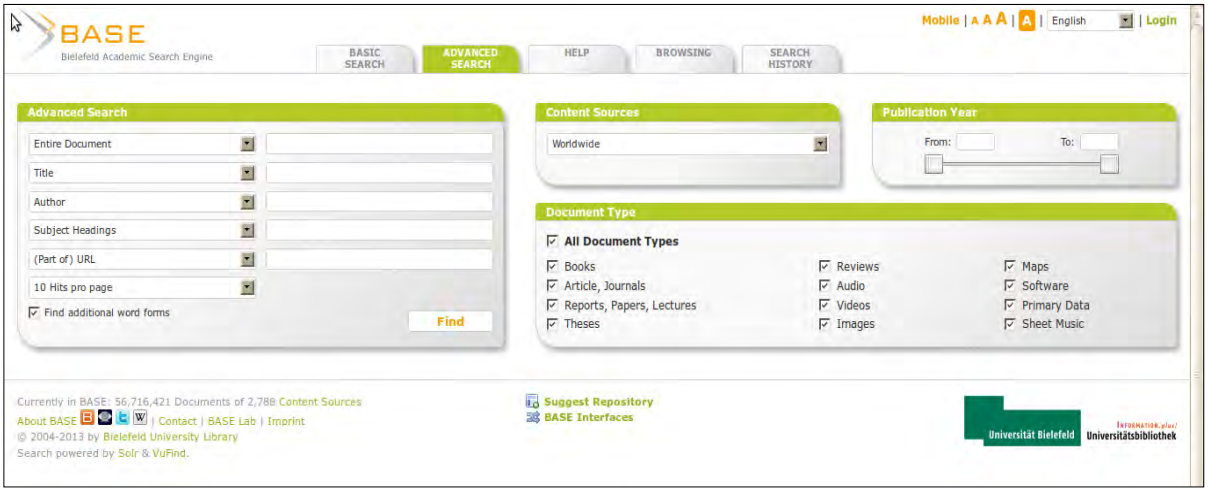

Figure 34: OA Retrieval System of BASE

CORE[100]

CORE (COnnecting REpositories) presently facilitates OA retrieval system for scholarly publications distributed across many systems. CORE depends on metadata harvesting through OAI/PMH. JISC, UK initially developed CORE as an aggregation of open access repositories in UK (142 approved OA repositories to be exact) but later on it was extended to cover OA resources at global scale. CORE is accessible through number of options like online portal, mobile device interface, and through repositories and libraries that have integrated CORE with local search service. As a whole, the interfaces may be grouped into five groups – i) CORE Portal (allows to search and browse OA resources harvested from a wide range of OA repositories through OAI/PMH;

[100] http://core.kmi.open.ac.uk/

ii) CORE Mobile (an Android application to search, browse and download OA resources); iii) CORE Plugin (script to integrate CORE with local repositories to extend search query to CORE); iv) CORE API (allows external systems and services like library OPAC to forward search query to CORE); and v) Repository Analytics (a value-added service to monitor the ingestion of metadata and content from repositories and provides usage statistics).

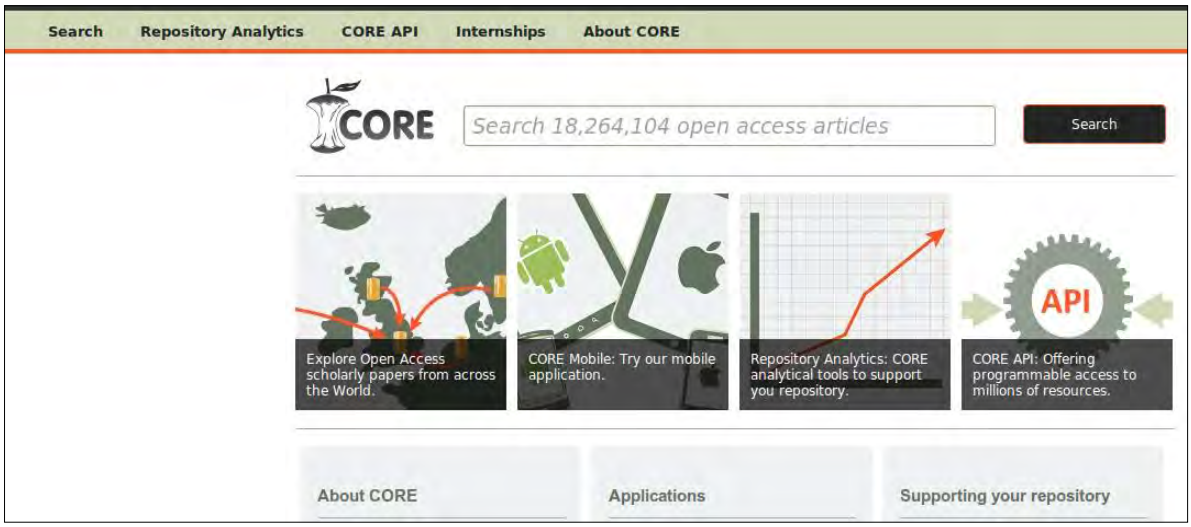

Figure 35: OA Retrieval System of CORE[101]

CORE supports almost all sophisticated search operators but only through simple search interface. But the availability of controlled vocabularies and subject category based browsing are not available till date. CORE retrieval system is supporting only monolingual retrieval of English-language OA contents.

OAISter[102]

OAIster is union catalogue of millions of OA resources, developed by OCLC through OAI/PMH based harvesting from collections across the globe. OAIster includes more than 25 million OA from more than 1,100 sources. Anyone can access OAISter through registration. The retrieval of features of OAISter is powered by WorldCat search services and provides almost all required search operators. It also supports a limited number of web 2.0 tools (like RSS and information mashup).

VOA^{3}R[103] (Virtual Open Access Agriculture & Aquaculture Repository)

VOA^{3}R is a social platform with OA retrieval system for students and researchers in agriculture and aquaculture. It integrates OA resources and uses AGROVOC thesaurus to support subject cataloguing and end-user retrieval. Apart from supporting search (simple and advanced) and browse (by author, tile, date etc), it has two unique experimental features - *Map view* (retrieved

[101] http://core.kmi.open.ac.uk/search
[102] http://www.oclc.org/oaister.en.html
[103] http://voa3r.eu/

results can be shown in a geographical map on the basis of the author's country or city as mentioned in author affiliation section of the OA resource); and *Time-line view* (allows items to be categorized by date of publication and this retrieval feature is integrated with AGROVOC thesaurus). The VOA^{3}R OA retrieval system presently covers 2656148 items and uses existing and metadata sets of OA resources to develop an advanced, community-focused integrated service for the retrieval of relevant open contents in the domain of agriculture and aquaculture. The integration of AGROVOC in the retrieval system helps researchers to formulate search query in terms of methods, variables, scientific techniques etc in combination with subject descriptors. The time-line view (see Figure 36) and map view are two important experimental features that may be trend-setters for other OA retrieval systems.

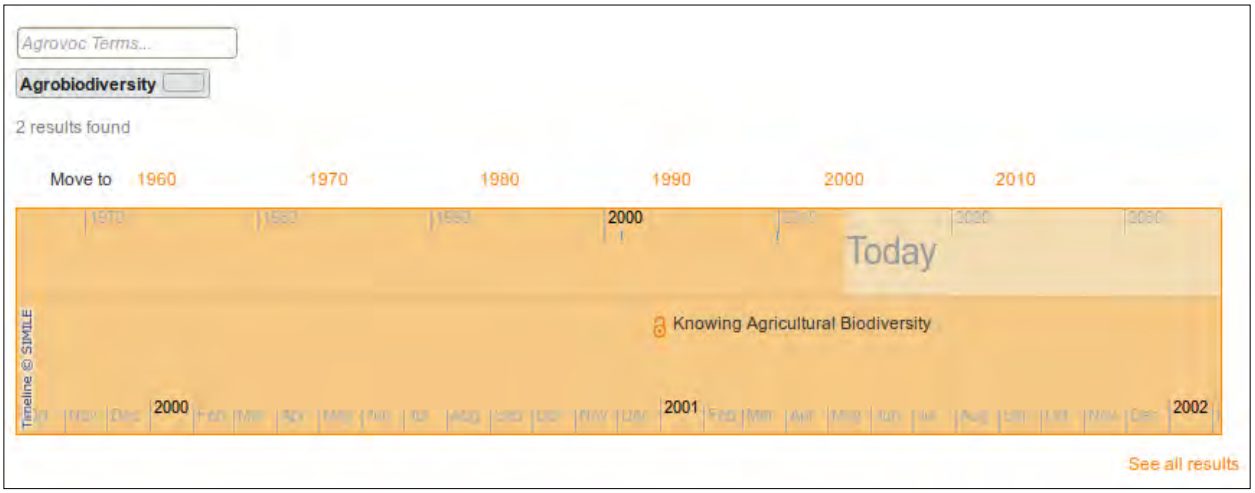

Figure 36: Time-line View in Retrieval Interface of VOA^{3}R

The localized repositories are mostly using open source text retrieval engines. The next section of this unit deals exclusively with the features of text retrieval engines in general and four major retrieval engines (namely Solr, Lucene, Zebra and MGPP) in particular.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this unit.

3) Discuss the core components of OA retrieval system.

……...

……...

……...

4) Enumerate the features of BASE as OA retrieval system.

……...

……..……..……...

……..……..……...

3.4 TEXT-RETRIEVAL ENGINES AND OPEN CONTENTS

Almost all OA content providers are using text retrieval engines or simply search engines for contents indexing, searching of index and ranking of retrieved results. An OA content manager must know the operational features of these text retrieval engines for two reasons – i) to select appropriate text retrieval engine for indexing open contents; and ii) to help users to guide in using search operators for content retrieval. There are three ways to adopt text retrieval engine – i) in house development of text retrieval engine; ii) using a commercial retrieval engine; and iii) using an open source based text retrieval engine. The main problems associated with in-house development of search engine are maintenance, regular up-gradation and total cost of ownership. Commercial search engine is not an attractive proposition for OA service providers both philosophically and economically. On the other hand, open source retrieval engines provide enhanced features, scope of customization and available free of cost. Most of the Green OA software (like DSpace, Greenstone, EPrint etc) and Gold OA software (like Open Journal System, Open Monograph Press) are using open source retrieval engines like Apache-Solr (DSpace version 4.0) Lucene (DSpace upto version 3.2) MGPP (Greenstone version 2.x), Zebra (Koha version 3.x). These open source retrieval engines may be categorized on the basis of following parameters – i) programming language in which it is implemented; ii) how it stores the index (inverted file, database, other file structure), iii) searching capabilities (Boolean operators, fuzzy search, use of stemming, etc), iv) ranking of retrieved results; v) document type handling capabilities (HTML, PDF, plain text, etc); vi) abilities to manage incremental indexes; vii) abilities to integrate related resources on-the-fly; and viii) generic factors such as user base, frequency of update of the software, the current version and the activity of the initiative. The next section discusses features of the major retrieval engines.

3.4.1 Apache-Solr

Solr was created by Yonik Seeley in 2004 as an in-house initiative at CNET Networks and donated to the Apache Software Foundation in early 2006. Solr is presently part of the Apache Lucene project. Solr is a standalone enterprise-grade full text search engine with high performance search server. It can be integrated with web-service through API. Solr is highly scalable, providing distributed search and index replication. It is written in Java and runs as a standalone full-text search server within a servlet container (such as Tomcat). Solr uses the Lucene (a project of Apache Software Foundation) library for full-text search, supports faceted navigation, provides hit highlighting utility and allows query language as well as textual search. The other prominent

features of Solr are - HTML administration interface; distributed and scaling of contents volume; search results clustering; plug-in integration; relevance ranking; caching and suitability in embedding in a Java Application. The marked advantages of Solr in comparison with other open source retrieval engines are – i) can drive more intelligent processing through the use of declarative Lucene Analyzer specifications; iii) CopyField functionality that allows indexing a single field multiple ways, or combining multiple fields into a single searchable field; iv) explicit field types that eliminates the need for guessing types of fields during search; v) external file-based configuration of stop word lists, synonym lists, and protected word lists; vi) many additional text analysis components including word splitting, regex and sounds-like filters. Presently, there are a few limitations of Solr – i) does not support relational joins; and ii) does not support wild card at the beginning of a search term. In 2010 Apache Lucene and Apache Solr are merged together by Apache Software Foundation[104].

3.4.2 Lucene

Lucene is a simple but robust and powerful text retrieval engine. This retrieval engine is quite suitable for nearly a decade now and especially useful for cross-platform applications. It provides the capabilities of fielded searching, stop world removal, stemming, and the ability to incrementally add new indexed contents without regenerating the entire index. DSpace presently uses Lucene as default search engine. Lucene comes with two main services available: indexing and searching. The indexing tasks are done independently from the search tasks. Both the index and search services are available so that developers can extend them to meet their needs. There are two varieties of Lucene – PyLucene (Java Lucene integrated with Python) and NXLucene (XML based query formulation, indexing and searching). Lucene supports several types of searches that are useful in retrieving open contents. Some of the major features of Lucene are listed below:

- Supports Boolean logic (Boolean operators allow terms to be combined through logic operators. Lucene supports AND, "+", OR, NOT and "-" as Boolean operators);
- Supports Exact Term search or Phrase search (The search term can be a world or a phrase. In phrase search the phrase should be within double quotes. Ex. “institutional repository”);
- Allows Proximity search (Lucene supports finding words are within a specific distance away);
- Provides Range search facility. Range queries allow one to match documents whose field values are between the lower and upper bound specified by the range query. Range search can be applied to any field including Date range. For example, if the search query is: Author:[Mishra to Mukhopadhyay], then the system shows those documents only written by names that fall between ‘Mishra’ to ‘Mukhopadhyay’ only;

[104] http://lucene.apache.org/solr

- Allows Field Search/ Field-specific Queries (One can search for a term in a particular field. Such as Author:mishra or Title:institutional repository);
- Supports Case sensitive searching, Relevancy ranking, Browsing of indexes, Truncation etc;
- Supports Wildcard and Stemming (Lucene supports single or multiple wildcard searches. The symbol (?) is used for a single character i.e. 'bo?k'. It may be the word like 'book'. The symbol '*' is used for multiple characters i.e. 'bio*'. It may be word like biology or biography); and
- Allows Fuzzy searching (Fuzzy search mechanism in Lucene is based on the Levenshtein Distance, or Edit Distance algorithm)

Lucene was developed by the Apache Software Foundation. It handles field and proximity searching, but only at a single level (e.g. complete documents or individual sections, but not both). Therefore, document and section indexes for a collection require two separate index files. It provides a similar range of search functionality to MGPP with the addition of single-character wildcards, range searching and sorting of search results by metadata fields. Another important feature of Lucene is its ability of term Boosting. Query-time boosts allow searcher to specify which terms are "more important". In other words, Boosting allows users to control the relevance of a document by boosting its term or phrase terms (e.g. "resource description"^4 "metadata encoding" means preference of phrase one over the second phrase). By default, the boost factor is 1. Although the boost factor must be positive, it can be less than 1 (e.g. 0.2). The higher the boost factor, the more relevant the term will be, and therefore the higher the corresponding document scores. A typical boosting technique may assign higher boosts to title matches than to body content matches: (title:interoperability OR title:"open access")^1.5 (body:interoperability OR body:"open access").

3.4.3 MGPP

MGPP (MG plus plus) is a new version of MG (Managing Gigabyte), developed by the New Zealand Digital Library Project as an open source retrieval engine. MGPP allows word level indexing to provide fielded, phrase and proximity searching facilities to end users. It supports Boolean operators and Boolean searches can be ranked. Greenstone, an open source digital archive software, is using MGPP as retrieval engine. The granular indexing of MGPP allows integrating document/section levels and text/metadata fields in one index file. MGPP has limitations like - i) no support for Fuzzy searching; and ii) searching may be a bit slower for large collection due to the index being word level rather than section level. The major features of MGPP are:

- text compression using a Huffman-coded semi-static word-based scheme;
- two-level context-based compression of bi-level images;
- lossless compression of gray-scale images for creating image collection;
- indexing algorithms for large volumes of text in limited main memory;

- index compression and processes for Boolean and ranked queries; and
- available with GUI interface to the retrieval system.

Apart from the above features, MGPP provides different search enhancements like folded search, stemming, term-weighted search and improvements over the fielded searching and proximity searching.

3.4.4 Zebra

Zebra is a powerful tool for indexing and searching highly structured data such as MARC records, and GILS records. The Zebra server is freely available for noncommercial applications. Zebra is licensed as Open Source, and can be deployed by anyone for any purpose without license fees. The C source code is open to anybody to read and change under the GPL license. The open source ILS Koha is using Zebra as retrieval engine. Apart from supporting basic search operators and techniques (like Boolean, Relational, Positional operators etc.), Zebra supports following advance and state-of-the art search techniques:

- *Term truncation* (left, right, left-and-right) and Fuzzy searches (spelling correction);
- *Scan* (Scan on a given named index returns all the indexed terms in lexicographical order near the given start term. This can be used to create drop-down menus and search suggestions);
- *Faceted browsing* (allows retrieval of facets for a result set);
- *Refine-search* (scanning in result sets can be used to implement drill-down in search clients);
- *Record Syntaxes* (Multiple record syntaxes for data retrieval: GRS-1, SUTRS, XML, ISO2709 (MARC), etc.);
- *Sort* (Sorting on the basis of alpha-numeric and numeric data is supported);
- *Combined sorting* (Sorting on the basis of combined sorts e.g. combinations of ascending/descending sorts of lexicographical/numeric/date field data is supported);
- *Relevance ranking* (Relevance-ranking of free-text queries is supported using a TF-IDF like algorithm.); and
- *Static pre-ranking* (Enables pre-index time ranking of documents).

3.4.5 Other Retrieval Engines

The above-mentioned four open source text retrieval engines are mostly in use to support OA retrieval systems. But, there are many other open source retrieval engines that need to be mentioned either because of their historical role or because of their experimental features. For example, SWISH-E is historically important as first plug-n-play text retrieval engine. The HTDig full-text search service was developed by using SWISH-E. It may be considered as pre-runner of modern text retrieval engines. On the other hand, Lemur is an experimental retrieval engine to develop auto summarization and

clustering of retrieved results. The following are the examples of open source text retrieval engines that are deployed by different software for retrieval of open contents.

Cheshire II: It's a logistic regression model based search engine available through FTP from the University of California at Berkeley. It supports Z39.50 protocol to avail distributed search features. The source code of the retrieval engine is available[105].

Glimpse: Freely available retrieval engine from the University of Arizona that is designed for efficient indexing (at some cost in retrieval efficiency). Glimpse is not configured for TREC-style evaluations, but these features can be introduced through customization.

IRF: It's a Java toolkit based open source retrieval engine for building IR systems for small applications. The strength of IRF is the object oriented framework that greatly simplifies tasks to modify source code. As Java is designed for platform independence rather than efficiency, the size of the collections that can be handled is quite limited.

Lemur: Lemur is an integrated retrieval engine with Lemur Toolkit, Indri, Galago, Lemur Query Log Toolbar and ClueWeb09 Dataset. It's an open source retrieval engine toolkit[106] for developing search engines, text analysis tools, browser toolbars, and data resources in the area of information retrieval. Apart from supporting regular search features, it supports query based sampling, database based ranking, result merging and summarization.

PRISE: It's a public domain vector space model based retrieval engine developed at NIST. PRISE includes Z39.50 interface for distributed searching. PRISE is configured to run TREC-style evaluations and the source code is available.

SMART: A vector space retrieval engine that is freely available by FTP from Cornell University. Library world knows SMART because of its association with retrieval experiments. It is configured to run TREC-style evaluations and the source code is available.

Xapian: An open source IR system that is designed to handle multilingual text processing and retrieval and available under GPL. It supports structured Boolean queries, relevance feedback, spelling suggestion and many other advanced search features, the popular social bookmarking tool. Delicious is using Xapian as retrieval engine.

In addition, there are also some powerful text retrieval engines such as DataparkSearch Engine; nutch;

Swish-e - Simple Web Indexing System for Humans – Enhanced; Webglimpse; and OpenFTS (Open Source Full Text Search engine).

[105] http://cheshire.berkeley.edu/
[106] http://www.lemurproject.org/

3.4.6 Comparison of Search Features

This section aims to help you in selecting appropriate text retrieval engine for development of Open Access searching and retrieval. The selection framework includes important parameters that required to be supported by text retrieval engine. The framework is divided into three groups – i) Core parameters; ii) Enhanced parameters and iii) Value-added parameters.

A. Core Parameters

It includes the features that are essential for selected text retrieval engine. The features are listed and compared against four major text retrieval engines (Table 13).

Table 13: Comparison of Core Parameters of Text Retrieval Engines

Features	***Apache Lucene***	***Apache Solr***	***MGPP***	***Zebra***
Query languages support	√	√	X	√
Fielded search	√	√	√	√
Data normalization	√	√	X	√
Term truncation	√	√	√	√
Fuzzy searches	√	√	X	√
Regular expression matching	√	√	X	√
Phrase Search	√	√	√	√
Wild cards	√	√	√	√
Proximity Search	√	√	X	X
Soundex search	X	√	X	X
Stemming	√	√	√	√
Duplicate detection	√	√	X	X
Relevance ranking	√	√	X	√
Search set manipulation	√	√	X	√
Search Result Filtering	√	√	X	√
Thesaurus/concept searching	X	√	X	√
Search statistics report	X	√	X	X
Recommended link	√	√	X	X
Term boosting	√	√	√	X

B. Enhanced Parameters

These features are added advantages of a text retrieval engine to help searchers in finding and displaying results according to the needs (Table 14).

Table 14: Enhanced Features of Text Retrieval Engines

Features	*Apache Lucene*	*Apache Solr*	*MGPP*	*Zebra*
Faceted browsing	X	√	X	√
Drill-down or refine-search	X	√	X	X
Sort by ascending or descending	√	√	X	√
Indexing speed control	X	√	X	√
Index size control	X	X	√	X
Reasonable response time (from request to results)	√	√		√
Granularity / Whole-doc, Section	√	√	√	X

C. Value-added Parameters

These parameters are additional utilities meant for both indexers and searchers (Table 15).

Table 15: Value Added Parameters

Features	*Apache Lucene*	*Apache Solr*	*MGPP*	*Zebra*
Query spelling correction	√	√	X	√
Download / save records (with format options)	X	√	X	√
Full record display	√	√	√	√
Brief display list	√	√	√	√
Highlights the corresponding field	√	√	√	√
Search result clustering	X	√	X	√
Numeric field statistics	X	√	X	X
Robust Updating	√	√	X	√
Browse collection (author, title, etc)	√	√	√	X
License (as open tool)	√	√	√	√
Stop word	√	√	X	√
Web Service support (through API)	√	√	√	√
Weighting and boosting options	√	√	√	X
Configure images (Icons and thumbnail)/ Image indexing	X	√	√	X
Content filtering	√	√	√	√
Query expansion and modifications	√	√	X	X

Please remember that support against a particular parameter by a text retrieval engine may change over the time as these open source text retrieval engines are under continuous up gradation.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this unit.

5) What is Text Retrieval Engine (TRE)? Enumerate the advantages of using open source TRE.

……..……..……...

……..……..……...

……..……..……...

6) Discuss the features of any TRE that you wish to use in your OA retrieval system.

……..……..……...

……..……..……...

……..……..……...

3.5 RETRIEVAL OF OPEN CONTENTS: SUPPORT TOOLS

As you know, language is a basic element of the Information Representation and Retrieval (IRR). It may take the form of either natural language or controlled vocabulary (a relatively latecomer in comparison to natural language). The applications and uses of languages in IRR may be studied under four groups, called four era of IRR language (Svenonious, 1986; Rowley, 1994; and Chu, 2009). The characteristics of these four eras may be summarized as below:

Period	**Use of language in IRR**
Era I	Natural language was the only language in IRR during the early days of information retrieval. In this era people started to realize the problems of synonymous and homographs in IRR by using natural language.
Era II	This era is characterized by the following events – • Introduction of controlled vocabulary; • Wide use of pre-coordinated vocabulary control devices (e.g. Classification Schemes); • Debate on Natural language Vs. Controlled vocabulary started but both of these languages coexisted.
Era III	This era is known for the following important events – • Resurgence of natural language on the basis of keyword based retrieval techniques and development of full-text

databases;
- Wide use of thesauri in developing bibliographic databases;
- Debate on Natural language Vs. Controlled vocabulary continued and intensified.

Era IV This era is dominated by the development of Natural Language Processing (NLP) techniques and characterized by a combined approach such as
- Use of Controlled vocabulary at backend of information retrieval system;
- System of invisible vocabulary control devices in natural language retrieval environment started in full swing;
- Advances in NLP and Artificial Intelligence (AI) contributed in developing retrieval systems in natural language (e.g. WIN retrieval system of Westlaw and AskJeev search engine)

Digital IRR including OA retrieval system generally use controlled vocabulary for populating subject access fields (like *DC.Subject*) but the use of natural language is increasing with the improvement of Natural Language Processing (NLP) technologies. Application of NLP may broadly be categorized into three groups:

Group I: Use of terms taken from titles, topic sentences, abstracts, and other important components (Assigned indexing)

Group II: Use of terms that are derived from any part of the document (Derived indexing)

Group III: Use of words or phrases from query representation of searchers

Activities of these three groups, related with natural language based information representation, are associated with inclusion of significant or desired words (i.e. candidate terms for indexing or query) and exclusion of non-significant or junk words (such as articles, prepositions, conjunctions etc.). In an automatic IRR a stop-word list is compiled and configured in system to stop indexing of the junk words. Some automatic IRR systems create go-list or desired word list as semi-structured vocabulary, which includes significant terms (along with synonyms etc.) from established vocabulary control devices (like thesauri, subject headings list etc.).

3.5.1 Vocabulary Control Devices

Controlled vocabularies are artificial languages with their own vocabulary, syntax and semantics. The vocabulary in a controlled vocabulary device is based on literary warrant and users warrant. Controlled vocabularies available in IRR domain may be divided into three groups – thesaurus, subject headings list, and classification scheme. As information professional you are already familiar with these devices. Therefore, a comparison of these devices may help you in determining their suitability for different applications (Table 16).

Table 16: Comparison of Vocabulary Controlled Devices

Controlled Vocabulary Features	**Thesauri**	**Subject Headings List**	**Classification Schemes**
Term representation	Descriptors	Subject headings	Classification labels
Reference mechanisms	U, UF, SN, BT, NT, RT	See, See also, X, XX	See, See also
Analysis process	Synthesis	Synthesis + Enumeration	Enumeration
Coordination process	Post-coordination	Post and Pre coordination	Pre-coordination
Specificity	High	Moderate	Low
Flexibility	High	Moderate	Low

Natural language or Controlled vocabulary: Which Way?

The four era of language in IRR (as described in foregoing paragraph) shows the trend of using natural language and controlled vocabulary in a combined way. Both of these groups have their own advantages and disadvantages. As a result, these two groups of language are applied in complimentary and supplementary basis for developing information retrieval system. A comparison of suitability for these two groups of language against major selected issues may be presented in Table 17.

As a whole we can conclude that advantages of using controlled vocabulary are related with efficient handling of synonyms, homographs and term association (syntax), and these are weak points of natural language. The advantages of using natural language are concerned with updating, accuracy, maintenance cost and compatibility, and these are weak points of controlled vocabulary. As a consequence of the relative merits of each of these systems, both have found their own places in IRR. The next sections shows you the use of controlled vocabularies in OA retrieval systems at two levels – use of controlled vocabulary for populating subject access fields and use of ontology in retrieval.

Table 17: Difference between Natural language and Controlled Vocabulary

Issues	Natural Language (NL)	Controlled Vocabulary
Synonym issue: Different terms referring to the same entity	This issue is a source of concern in NL based IRR	Only one term selected as candidate term and rest are non-preferred terms
Homograph issue: Same term carrying different meaning in different context	In natural language IRR this may lead to ambiguity	The context for interpreting homograph is provided in controlled vocabulary often by using parentheses e.g. bank (finance) and bank (river)
Syntax issue: Association of terms properly to convey meaning	There is a danger of false drop trough wrong association	Role operators are used to indicate relationships between or among terms
Accuracy issue: Exact representation of concepts	Attainable if NL is chosen as IRR language	Lacks specificity as a result of the language manipulation process
Currency issue: Updating issues related with IRR language	NL requires no updating and there is no problem of serving query that contain new terms	Requires continuous updating (lengthy and costly process) and query with new terms cannot be satisfied
Cost issue: Cost in terms of time, energy and manpower to learn, create and maintain	Neither training nor maintenance is required	High cost involvement is a characteristic feature of controlled vocabulary
Compatibility issue: Issues related with switching, migration and mapping	Switching and migration is seamless	Seamless compatibility is difficult to achieve

3.5.2 Subject Access Systems

There are broadly three parallel IR systems. These are traditional or manual IRR, online and optical disk based IRR, and Web-enabled IRR. In the first two IR systems, controlled vocabulary has taken a dominant role as IRR language.

But in Web-enabled IRR, application of controlled vocabulary is a costly option in view of the ever increasing magnitude of digital resources coupled with the factors like uneven quality and short life expectancy for these resources. Most of the Web-enabled retrieval systems make no use of controlled vocabulary apart from stop-lists or go-lists. The lack of controlled vocabulary as IRR language could be one of the reasons for non-satisfactory performance of these retrieval systems. Under such circumstances one question is gaining serious attention from library professionals – *what will be the future of controlled vocabulary as a language in digital IRR*. Lancaster & Warner (1993) advocated four possible approaches in this direction:

Solution I: Controlled vocabulary for both representation and retrieval

Solution II: Natural language for both representation and retrieval (by using role operators and pre-coordination processes from controlled vocabulary)

Solution III: Controlled vocabulary for representation only (use of invisible vocabulary control device at the back-end of the retrieval system)

Solution IV: Controlled vocabulary for retrieval only (use of vocabulary control device at the front-end of the retrieval system i.e. in search interface)

The last three approaches are equally viable in digital IRR environment as far as cost of creation and maintenance of the IR systems is concerned. But the third and fourth approaches require settling the issues related with switching and mapping of vocabularies. A comparison of these two mechanisms gives us following result:

Vocabulary switching	**Invisible vocabulary**
• Mechanism for automatically changing from one IRR language to another across different subject domains	• Invisible vocabulary handles translation between natural language and one controlled vocabulary stored online
• Multiple subject domains are covered	• Only one subject domain is covered
• Based on NLP techniques	• Based on semantic mapping of concepts

Digital repository software are increasingly aware of the advantages of using controlled vocabularies in retrieval particularly in populating subject access fields. The Eprint archive software right from the beginning using standard subject access system (by default LC Subject categories but it may be

configured to include any such subject categorization) for at the time indexing as well as at the time of searching. Figure 37 shows the use of standard subject access system at time of submitting open knowledge objects.

Figure 37: Subject Categories in Eprint in Submission Interface

After submission process is over, the subject field is populated by selected subject category or subcategory and it becomes ready for searching by subject category in indexing process (see Figure 38).

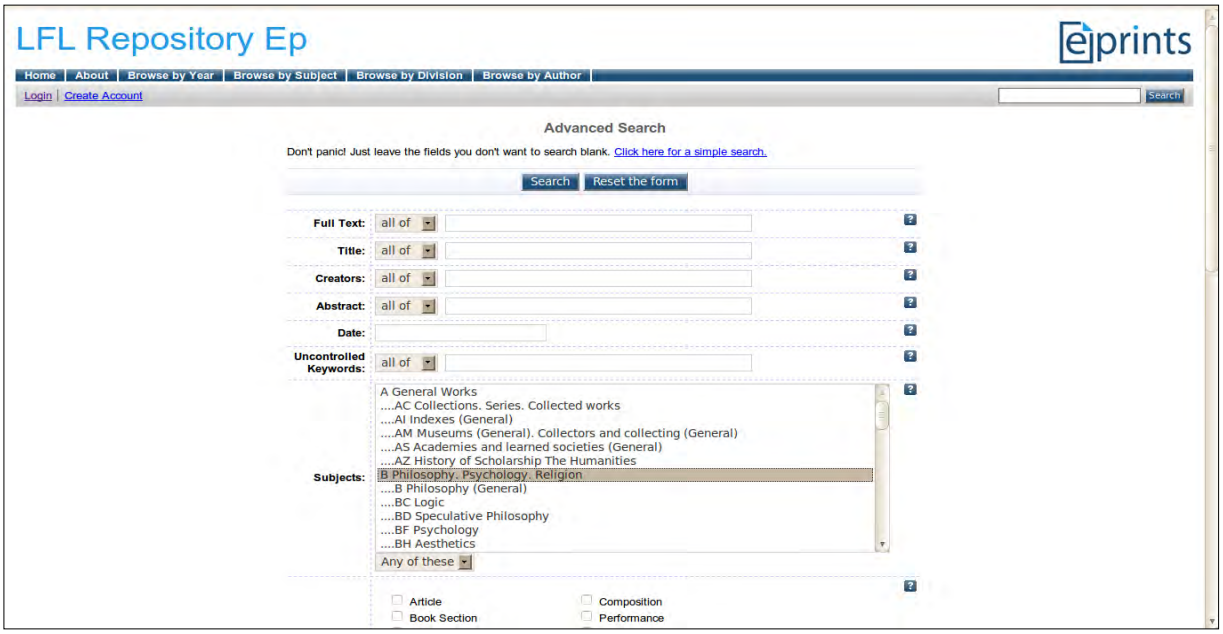

Figure 38: Subject Categories in Eprint in User Interface

DSpace also supports the use of controlled vocabularies in indexing and searching but the process of integration is much more flexible. It supports

SKOS (Simple Knowledge Organization System), a W3C recommendation, for representing and formatting subject hierarchy. As a result, integration of subject categorization in DSpace and interoperability of subject categories from/to DSpace is standardized.

Step 1: Enabling Controlled vocabulary

It involves opening controlled vocabulary option in *dspace.cfg* file

Step 2: Creating SKOS-enabled Subject Access System

This step involves representation of subject access system in SKOS format in XML. This XML based representation of subject hierarchy also provides scope for inclusion of multilingual subject heading. It means subject heading / preferred term may be represented in more than one language or scripts.

Step 3: Linking Submission interface with Subject Access System

This step links the XML-formatted standard subject access system with submission interface (for indexing).

These three steps integrate controlled vocabularies in DSpace for managing retrieval of open contents in both interfaces – indexing phase and end user searching phase.

3.5.3 Ontology Support

Ontology is a formal, explicit specification of a shared conceptualization. In simple words, it is a model of organized knowledge in a given domain (e.g. fisheries). Ontologies consist of components called “concepts, attributes, relations and instances”. Ontology is considerably different from taxonomy and thesaurus. Taxonomy is a hierarchical tree structure which models a domain from abstract to specific. On the other hand a thesaurus is a structured vocabulary that defines each term by three major types of relationships – hierarchical (as in a taxonomy), associative and equivalent. But ontology is the most formal model as it defines the meaning of concepts by modeling constraints that restrict the number of possible interpretation. Therefore, these three schemes differ mainly in their degree of precision. However, a comparison is given here in Table 18 to help you in understanding the features.

Table 18: Comparison of Taxonomy-Thesaurus-Ontology

Features	Taxonomy	Thesaurus	Ontology
Background	Natural Sciences and Universe of Subjects	Library and Information Science	Metaphysics, AI and NLP, Knowledge modeling
Modeling standard	None	ISO-2788 (equivalent standards are BS 5723, BS 6723, Z 39.19)	No official standard yet
Notational standard	Graphical tree and Mixed-base notation	BT, NT, RT, UF, USE etc.	RDF schema and OWL
Relationships	Basically hierarchical but all types of relationships are modelled	Untyped hierarchical, associative and equivalence	Typed hierarchical and associative
Properties	None	Scope Notes (SN) device	Domain and Range (in RDF schema)
Application	Classification, Navigation, Search	Classification, Navigation, Search	Classification, Navigation, Search, Visualization and Automated reasoning
Popular tools for creation	Mind manager	MultiTES	Protégé

Thesauri are structured according to an international standard (ISO-2788), and, therefore, these schemes can be transferred to ontologies through the application of ontology representation language (such as RDF schema). In Semantic Web environment, we need an element which can unequivocally describe the meaning of a concept or word for the software agent. This role is performed by ontologies. In practice, desired words/concepts/terms are marked by a tag that refers to the ontology. A software agent who comes across the tag can consult the ontology for meaning of the term. The Semantic Web extends the present form of Web by giving meaning and context to information bearing

objects, allowing people and software agents to share and process data more competently. Ontology helps to boost the effectiveness and uniformity of describing resources i.e. they allow more sophisticated functionalities in IRR. The use of standards, such as the Resource Description Framework (RDF) and Web Ontology Language (OWL), provide structures and methods for descriptions, definitions and relations within a given domain. In OA domain, some of the content retrieval systems support ontology-driven retrieval of knowledge objects. For example, *sciencewise.info* an experimental OA retrieval system (presently covers Physics, Life Sciences, Humanities and Information Technologies disciplines) provides ontology-driven search interface. A search query is automatically linked with available domain ontology and user allows navigating from one Node to another. It also gives users links to open contents (preprints/post prints).

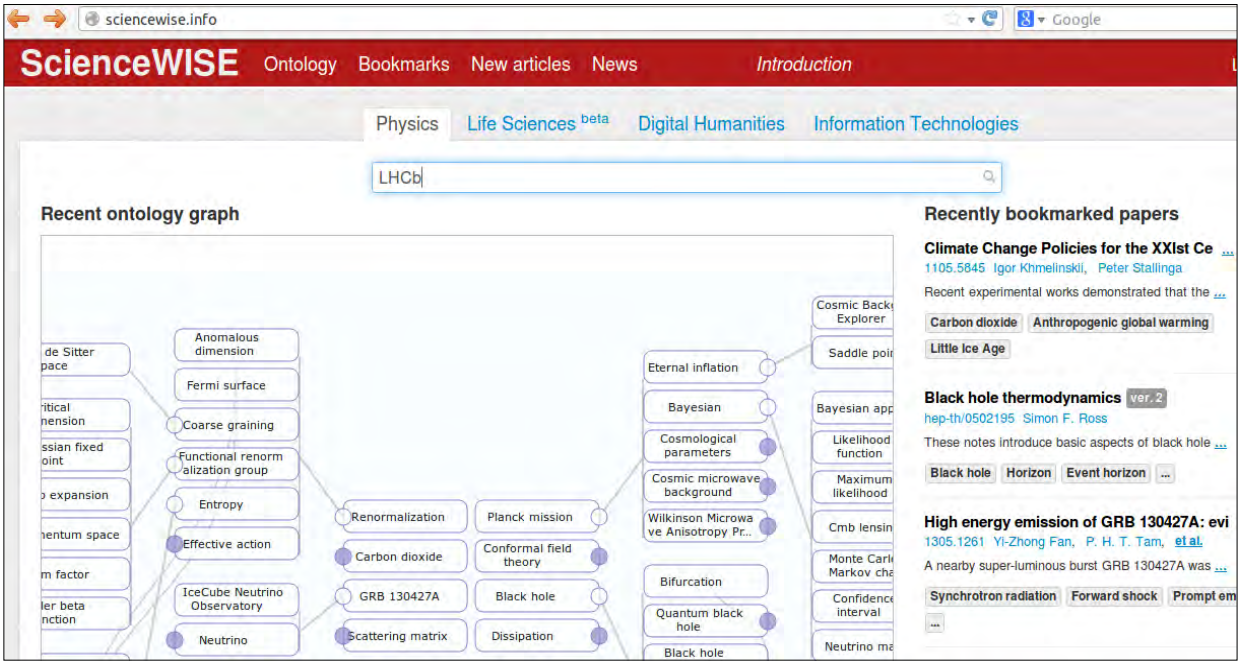

Figure 39: Ontology-driven Retrieval in sciencewise.info

For example, a search on LHCb in *sciencewise.info* shows position of the query term in domain ontology (including it's relationships with other concepts) and provides link to available open access journal papers related to LHCb (Figure 39 & 40). This is also a participative retrieval architecture which allows user scope to define a new concept or to edit an existing concept in domain ontology.

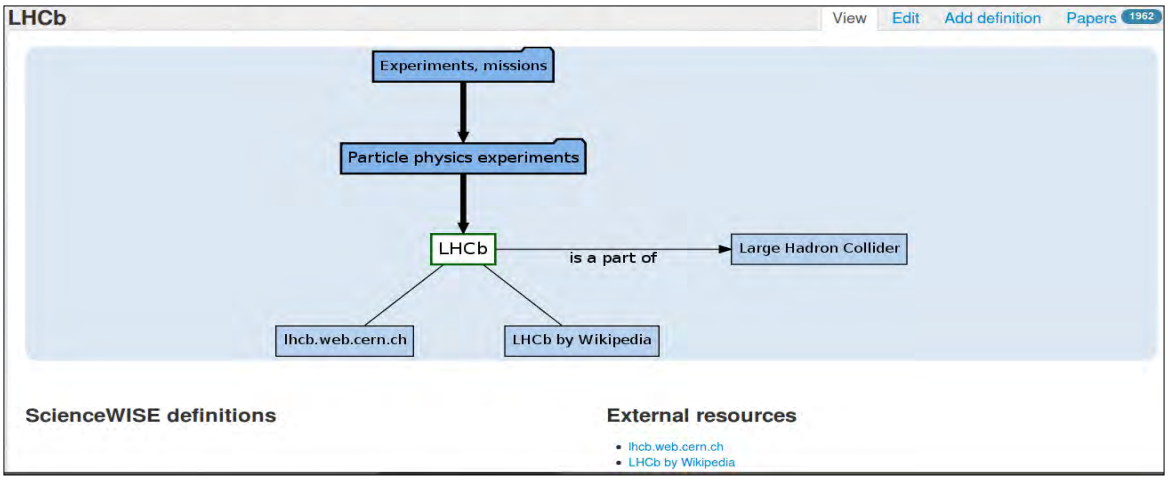

Figure 40: Linking of Query Term in Domain Ontology in sciencewise.info

3.5.4 Statistical and Other Tools

You already know in Unit 2 of this Module that usage data and statistics is considered as a value-added feature for any OA retrieval system. Many repository software are attempting to implement the statistics add-on by using usage data stored in retrieval engine. For example, the statistics add-on in the DSpace platform allows gathering, processing and presenting usage data, contents related data and administrative statistics by utilizing Apache Solr (text retrieval engine in use in DSpace version 4.0) underlying application layer for harvesting vast array of usage data. Some of the statistical datasets displayed by DSpace are – top ten countries and cities from where visits originate, total number of visits for community, collection and items, search history, work-flow related statistics, item download statistics etc (Figure 41).

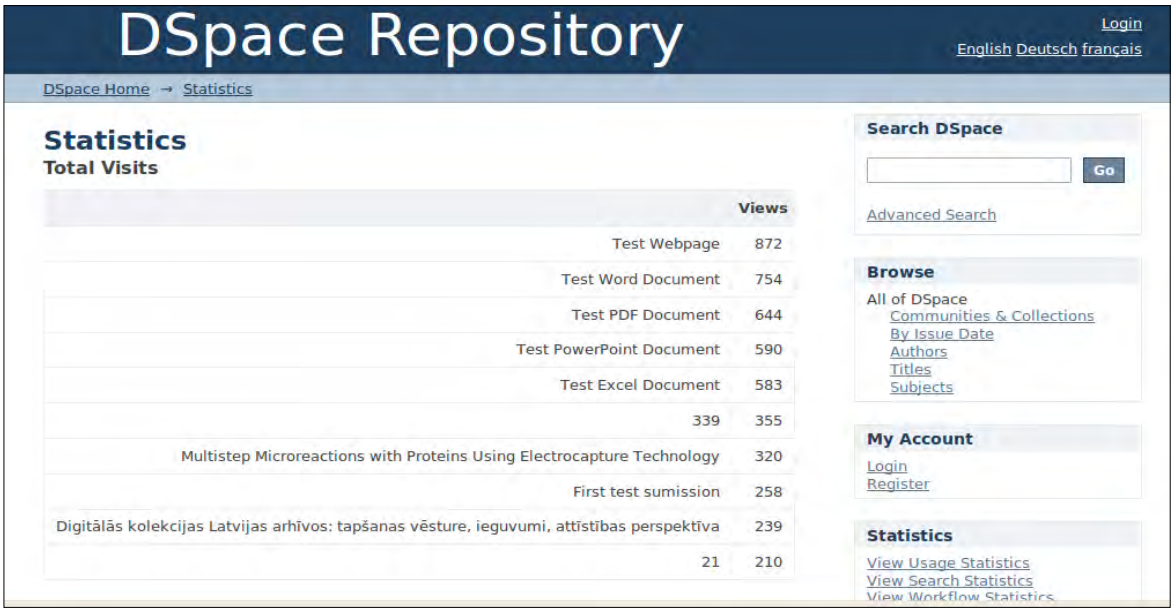

Figure 41: Usage Data and Other Statistics in DSpace Retrieval System

The other associated services that support OA retrieval system are Web 2.0 tools for achieving interactive, collaborative and participative architecture in content retrieval. These are use of RSS feeds, content rating, folksonomy, review submission, social networking tools etc. Shafi, Gul and Shah (2012) conducted a study in 2012 to measure the use of Web 2.0 tools and services in OA repositories listed in openDOAR (1977 to be exact). The finding of this research provides shows that the use of RSS is the most popular Web 2.0 application in OA retrieval (possibly the use of RSS as automatic alerting service for updated contents makes it very useful support tool in OA retrieval) and social bookmarking occupies the next position (again because of scholarly reasons). The other useful Web 2.0 tools are social networking tools (Twitter, Face book, and YouTube) and collaborative tools (like Blog, Flickr, and Podcasting). In a total of 1,412 accessible repositories (in 1977 total listed repositories), 57 percent (804 number of repositories) applied Web 2.0 tools and the remaining 43 percent (608 number of repositories) have not yet applied Web 2.0 tools. Again a country-wise distribution of the use of Web 2.0 tools in OA repositories shows that US based OA retrieval systems ranked first and UK and Germany occupied the next positions respectively. One interesting fact is that use of web 2.0 tools in Asian OA retrieval systems are increasing (Taiwan – 83.33%, India – 60% and Japan – 41.56%) in comparison with European and American OA retrieval systems.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this unit.

7) Discuss the use of Controlled Vocabulary in OA retrieval systems.

……..……..……...

……..……..……...

……..……..……...

……..……..……...

……..……..……...

8) What is Ontology? Discuss how is it helping to improve retrieving OA resources.

……..……..……...

……..……..……...

……..……..……...

……..……..……...

……..……..……...

3.6 RETRIEVAL OF SPECIALIZED OPEN CONTENTS

Michel Lesk (1995) in his seminal paper reported a comparison between development in the domain of Information Retrieval and seven ages of man as described by Shakespeare in *As You Like It (Act 2, Scene 7, lines 143-166*). Lesk predicted many possible achievements of IR in the first decade of 21st century. These are – i) Resource Description Framework (RDF) and XML supported Web-enabled IR; ii) Centralized/Federated search services through harvesting; iii) Influence of Semantic Web and Web 2.0 in Information Representation and Retrieval (IRR); iv) Matured multimedia IR systems with information mashup support; v) Integration of digital libraries with online learning environments; vi) Sophisticated multilingual IR with Unicode support; vii) Interactive and collaborative IRR; and viii) Application of Ontology in IRR. Many of these predictions are still in research bed but multimedia IR and multilingual IR are quite matured now. This section covers major aspects of these two retrieval systems.

3.6.1 Multimedia Contents Retrieval

Full text information representation cannot handle non-textual information objects like diagrams, charts, sound, image etc. But web is now increasingly populated by slides, MP3 files, video clips, animated pictures, photographs etc. Moreover, a single digital object may contain text, image, video, and audio. These information bearing objects are called multimedia information. Multimedia information representation and retrieval is one of the hardest challenges to the domain of information retrieval. Multimedia information representation involves three approaches namely - i) Description-based; ii) Content-based; and iii) a combined approach. Description-based approach takes care of information representation through enumeration of descriptive elements like creator, caption, image size, keywords, theme etc. The problem of this approach is that in most of the cases multimedia objects can hardly be described explicitly and objectively. In Content-based approach information representation is based on intrinsic attributes of multimedia objects such as image color, bit-depth, shapes, texture, sound pitch etc. Combined approach is an integration of description-based and Content-based approaches. Researchers of multimedia information retrieval strongly recommend application of integrated or combined approach for Web-enabled access to multimedia based information objects.

3.6.2 Multilingual Contents Retrieval

Text is the most prominent form of information representation, though other representation techniques such as symbols, signs, pictures, sound etc. are also playing important roles. With the progress of multimedia technology, many formats came into existence to deal with multimedia files. However, ASCII remained *de facto* standard for textual data processing for a long time. ASCII is

an 8-bit (1 Byte) code and can represent maximum of 2^8 or 256 characters. Most of the ASCII values are reserved for Roman scripts. Although there are instances where ASCII or extended ASCII (such as ISCII) has been in use to represent scripts other than Roman scripts, it is crystal clear that ASCII is inadequate for a multilingual approach to represent various characters from different scripts of the world. The reason is quite simple – 256 characters cannot cover all the scripts of the world. Unicode is a promising open text encoding standard for processing and retrieval of multilingual data. The Unicode Consortium was incorporated in January 1991 to promote the Unicode standard as an international encoding system for information interchange. The Unicode Standard is the universal character-encoding scheme for written characters and text. It defines a consistent way of encoding multilingual text that enables the exchange of text data internationally and creates the foundation for global software. The Unicode Technical Committee (UTC) is the working group within the Consortium responsible for the creation, maintenance, and quality of the Unicode Standard. The UTC follows an open process in developing the Unicode Standard and its other technical publications. In the beginning Unicode was a simple, fixed-width 16 bit encoding. Over the time, Unicode changed this fixed-width encoding style and presently allows three different forms of encoding to meet different requirements:

- UTF-8 attempts to allow legacy systems to use Unicode by coding the characters in the ASCII character set with only eight bits, and encoding characters that are not in the ASCII character set with 16 bits. This is commonly used for Web pages.
- UTF-16 is supplementary characters outside the basic multilingual plane. It encodes most of the world's major languages in a fixed 16-bit character representation (2 bytes). This is the most common implementation.
- UTF-32 is an actually UCS 4, given a new name. It uses four bytes (32 bits) to encode all possible characters (rarely used).

Many major Web-based search services are Unicode-compliant and support multilingual information retrieval e.g. Google and Yahoo provides Indic-script based retrieval.

Apart from the use of Unicode as text-encoding standard, there are two sets of requirements for developing Unicode-compliant Indic script based information retrieval systems. These are - i) system specific requirements and ii) language specific requirements. The first group needs Unicode-compliant Operating System, Text editor, Programming environment and Database management system (Unicode-compliant DBMSs support UTF-8 as standard for native character set). The second set requires language specific tools like Virtual keyboard, Rendering engine and Open type font(s) for respective language. Conjuncts and ligatures are the most font dependent of any scripts. They could be at different positions in different fonts. A rendering engine should be using each font's glyph substitution tables to contextually render the characters. On the other hand, an open type font has two distinct advantages in a multilingual

environment – its cross-platform compatibility and its ability to support widely expanded character sets and layout features.

Let's see how multilingual user interface and retrieval achieved in DSpace repository management software (with reference to an Indic script but this methodology may be extended to any script in the world). The methodology includes three basic steps – i) use of UTF-8 as default character set in backend RDBMS; ii) preparing Java servlet engine to support transaction of multilingual data in UTF-8 encoding; and iii) translation of messages and menus (English language messages and menus stored in DSpace in a central place). This methodology with these three steps create language-specific user interface in DSpace and supports simple and advanced search and retrieval for DSpace.

Step 1: Setting native character set as UTF-8 in back-end RDBMS

The first logical step to achieve multilingual retrieval is to set native character set as UTF-8 in back-end RDBMS (here PostGreSQL) (Figure 42).

Step 2: Setting URIENCODING in Web transactions

The URIEncoding value need to be set as UTF-8 (in DSpace the *server.xml* file need to be modified) to support multi-script data transaction.

Connector

port="8080" maxHttpHeaderSize="8192"

maxThreads="150" minSpareThreads="25" maxSpareThreads="75"

enableLookups="false" redirectPort="8443" acceptCount="100"

connectionTimeout="20000" disableUploadTimeout="true"

URIEncoding="UTF-8"/>

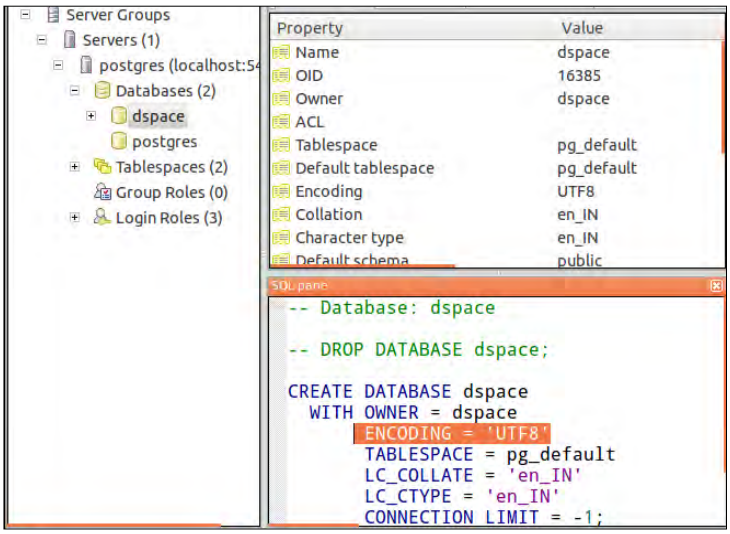

Figure 42: UIF-8 as Native Character Set

Step 3: Language-specific translation

Translation of messages into target language and script is next logical step. In DSpace we can set different *messages.properties* file for different languages. The file name must be set by using ISO language code (here file name for Bengali translation is messages_bn.xml). The Figure 43 shows the translation of messages in DSpace in Bengali. This translated *messages.properties* file must be saved as Unicode file.

```
jsp.layout.navbar-default.about = ডিস্পেস সর্ম্পকিত
jsp.layout.navbar-default.advanced = উন্নত অনুসন্ধান
jsp.layout.navbar-default.authors = লেখক
jsp.layout.navbar-default.browse = ব্রাউজ
jsp.layout.navbar-default.communities-collections = সংগ্রহ ও  সংগ্রহ বিভা
jsp.layout.navbar-default.date = তারিখ অনুযায়ী
jsp.layout.navbar-default.edit = প্রোফাইল পরিবর্তন
jsp.layout.navbar-default.go = অনুসন্ধান
jsp.layout.navbar-default.help =সহায়তা
jsp.layout.navbar-default.home = প্রাথমিক অবস্থান
jsp.layout.navbar-default.loggedin = সদস্যের নাম {0}
jsp.layout.navbar-default.logout = লগ আউট
jsp.layout.navbar-default.receive = ই-মেল আপডেট
jsp.layout.navbar-default.search = অনুসন্ধান
jsp.layout.navbar-default.sign = সদস্য হন
jsp.layout.navbar-default.subjects = বিষয় শিরোনাম
jsp.layout.navbar-default.titles = আখ্যা
jsp.layout.navbar-default.users =সদস্য এলাকা
jsp.layout.navbar-default.users-authorized = নিয়ন্ত্রিত প্রবেশ
jsp.layout.navbar-default.subjectsearch = বিষয় শিরোনাম
```

Figure 43: UTF-8 as Native Character Set

Step 4: Retrieval interfaces

The multilingual user interface along with search and retrieval is shown in Figure 44 (as product of the above-mentioned three steps). It supports field-specific search, Boolean operators and sorting order of results.

Figure 44: UTF-8 based User Interface in Dspace

This interface supports display of Boolean operators in target language, relevance-ranking and sorting in ascending/descending order. The only problem of this method is that it cannot sort chronologically when date value entered in other than Indo-Arabic numbers.

CHECK YOUR PROGRESS

Notes: *a) Write your answers in the space given below.*
b) Compare your answers with those given at the end of this unit.

9) What are the requirements for developing multilingual OA retrieval system?

……

……

……

10) Mention the steps to develop multilingual IR through DSpace.

……

……

……

3.7 LET US SUM UP

Evidences of organizing and archiving written form of information is dated back to around 3000 BC during Sumerian civilization but modern IRR is greatly influenced by the great visionary Vanneuvar Bush. In 1945 he envisioned a single window user interface for fast access to the contents of the

world's libraries by 2010 and reported in his seminal article. Information retrieval deals with the problems related with the storage, access and searching of information sources by persons in need of information. In this digital network era, information sources are growing at an exorbitant rate, available in many forms and formats, and accessible through various channels. Moreover, recent advancements in ICT help in integration of different information sources and process them on a larger scale. OA retrieval systems, as part of the great landscape of IRR are promising a new era of content retrieval. This unit on OA retrieval is an attempt to provide systematic exposition of OA retrieval system from traditional to modern era with facets on techniques, approaches, models and evaluation processes of retrieval systems. It briefly discusses features of major OA retrieval systems at global scale including many ongoing experiments on OA retrieval like map-view of results, integration of controlled vocabularies, subject categories and ontology-driven organization. It dedicates a complete section on text retrieval engines as most of the OA contents providers and OA content management software are increasingly using open source text retrieval engines for dissemination of OA resources. This unit also deals with multilingual OA content retrieval in relation with necessary configuration of OA repository management software.

MODEL ANSWERS TO CHECK YOUR PROGRESS

Unit 1

1) A metadata schema includes metadata elements, encoding rules for description and prescribes possible use of standards for some metadata elements. As a whole, it specifies three independent but related aspects of metadata – semantics, content rules and syntax. Semantics refers to the metadata elements that are included in the scheme by giving each of them a name and definition. Content rules indicate how values for metadata elements are selected and represented and Syntax of a metadata schema is concerned with the encoding of metadata elements in machine-readable form.

2) The most important role of metadata in OA context is to inform the status of a piece of content as open access. Apart from this vital function, OA metadata helps librarians in data mining, pattern identification (organization and usage), clarity over licensing agreements, discovering of OA, and accessing open access contents within hybrid journals. On the other hand, metadata helps end user in finding and accessing OA contents, in setting priority of OA contents over paid contents (filtering of results by OA status), in knowing access and re-use permissions, and in getting help to cite OA resources. The other stakeholders like publishers and funders are also benefitted from OA metadata such as i) publishers want to clearly convey what readers can and cannot do with the objects they publish; ii) research funders want to promote research output they sponsor; and iii) search engines, A&I databases, and other discovery services use OA metadata to help users in finding OA resources.

3) Metadata policy is an important component of OA resource management system. Such a policy framework helps repository managers to solve issues like – i) Who can enter or edit metadata? ii) Which metadata standards are to be followed? iii) Whether different metadata schemas are required for describing different type of documents? iv) Whether or not the repository systems allow metadata harvesting by service providers? v) Which protocols should OA system support for metadata harvesting? vi) Which fields require support for authority files and standards lists? vii) How to deal with rights management description? viii) How and to what extent metadata be exposed for reuse?

4) The usage metadata may serve as an important value-added service for users of open contents. Apart from the contributors and users of open access resources, funding agencies are also interested in availability of integrated usage data to measure research impact and to analyze trends over time. There are many standards and initiatives for describing and storage and usage metadata in the domain of OA such as SURE (Statistics on the Usage of Repositories), (Network of European Economists Online), KE-USG (Knowledge Exchange Usage Statistics

Guidelines), and OpenAIRE that specify metadata formats to be used to (Publishers and Institutional Repository Usage Statistics), OA-Statistik, NEEOncorporate information of usage events. Most of these initiatives are based on the OpenURL ContextObject format.

5) DCMES is a generic metadata schema and meant for OA contributors. Therefore it follows a set of principles that help easy encoding of OA resources by contributors themselves. It is not heavyweight schema like MARC 21 that requires skills and training for metadata encoding. DC metadata by following *Six Principles* – i) Intrinsicality: DC metadata is based on intrinsic data; ii) Extensibility: It allows inclusion of extra descriptive materials for specialized requirements; iii) Syntax Independence: It is applicable to a wide range of disciplines and application program; iv) Optionality: All the DC elements are optional; v) Repeatability: All the DC elements are repeatable, and vi) Modifiability: Each element in the Dublin Core has a definition, which is self-explanatory. Each element can be modified by an optional qualifier and in such cases the definition of the element is modified by the value of the qualifier.

6) DCMES may be categorized into two groups as far as encoding level is concerned. "Simple Dublin Core" is DC metadata that uses no qualifiers. It applies only main 15 elements without any qualifier. On the other hand, "Qualified Dublin Core" uses additional qualifiers to increase specificity or precision of the metadata. There are two broad classes of qualifier – i) Element Refinement (these qualifiers make the meaning of an element specific); and ii) Encoding Schemes (these qualifiers identify schemes that aid in the interpretation of an element value; these schemes include controlled vocabularies and formal notations.

7) ETD domain: ETD-MS, UK-ETD, Shodhganga

Image domain: VRA-Core

Maps: FGDC

Learning Objects: IEEE-LOM, GEMS, IMS Global, CanCore

Cultural Objects: CCO

Compound Digital Objects: OAI-ORE

8) AGLS: Government information resources

SWAP: Knowledge objects in Green OA

MIDAS: Cultural heritage

ONIX: Book industry

GILS: Government records

e-GMS: E-governance

9) Data value standards advocates use of controlled terms to ensure consistency and to achieve collocation of resources related the same

topic or person through the application of thesauri, controlled vocabularies, and authority files. These standards ensures both finding and collocations functions.

10) MODS is meant for exposing MARC 21 bibliographic format in XML syntax, MADS aims to release MARC 21 authority data in XML format. METS acts as metadata wrapper. MODS and MADS are often used in harmony to describe bibliographic and authority datasets in XML and METS provides a metadata wrapper to store, deliver and sharing of resource description datasets.

Unit 2

1) Interoperability may fundamentally be grouped into two categories – i) Syntactic interoperability; and ii) Semantic interoperability. COAR (Confederation of Open Access Repositories) identified seven major areas of interoperability – metadata level for transferring metadata, content level for supporting multiple deposits, identifier level for unique identification of resources and contributors, usage data level for sharing and aggregating usage statistics, network level for cross-system interoperability, object level for transferring compound digital objects and semantic level.

2) The DELOS Digital Library Reference Model prescribes architecture level interoperability to support two components: i) component profile and ii) application framework. The first one prescribes that each architectural component must be associated with a profile to describe functionality of the software component. The application framework prescribes that seamless ex

3) Aggregating of usage statistics is emerging as an important area in open access interoperability. It allows measuring impact of individual open knowledge objects (e.g. Research articles) and supports aggregation and exchange of usage information from different repositories and information systems. Many protocols and standards are being developed in the area of cross-repository usage statistics like COUNTER, KE-USG, SURE (Statistics on the Usage of Repositories) and PIRUS (Publishers and Institutional Repository Usage Statistics).

4) OAI-ORE, as an interoperability standard for compound digital objects, aims to provide solution that supports aggregations of Web resources. The OAI-ORE standard has four basic components to support transferring of compound digital objects in heterogeneous network environment – I) A model that includes use of RDF, XML, Ontology and Cool URI; ii) Resource aggregation; iii) Resource Map; and iv) Resource map representation in RDF/XML or Atom/XML.

5) The OAI/PMH is a matured open standard in the area of metadata interoperability. It has two components – data provider and service provider. The content negotiation between these two groups takes place on the basis of Six Verbs. These are: Identify (return general information about the archive and its policies); ListSet (provide a

listing of sets in which records may be organized); ListMetadataFormats (list metadata formats supported by the archive as well as their schema locations and namespaces); ListIdentifiers (list headers for all items in repository corresponding to the specified parameters); GetRecord (returns the metadata for a single item in the form of an OAI record); and ListRecord (retrieves metadata records for multiple items).

6) The identifier-level interoperability area deals with standards for unique author identification (e.g. ORCID, AuthorClaim, VIAF etc) and standards for object identification (e.g. DOI, Handle system of CNRI, PersID etc). Recently standards for dataset identification (e.g. DataCite) are also emerging in a big way.

7) There are many similarities and differences between Z 39.50 standard and OAI/PMH standard. Both are meant for bibliographic domain, both standards deal with metadata sharing and transferring. But these two standards also differ from each other in many contexts – Z 39.50 is standard developed by NISO (A SDO in US) and OAI/PMH is a cooperative open standard. Technically, these two standards differ in search type (Z 39.50 is distributed searching and OA/PMH is centralized searching); and search agent (Z 39.50 search is data provider activity and OA/PMH is using service provider for searching).

8) There are three groups of activities for implementing harvesting services – Group I deals with selection of harvester, installation of required software environment, installation and configuration of harvester. Group II tasks include selection of repositories to be harvested and collection of required dataset (e.g. Title of repository, repository URL, OAI base URL, of repository and mail id of repository administrator). The Group III activities include addition of repository, management of repository, harvesting metadata from repositories and designing user interfaces for retrieval of harvested metadata.

9) The term Linked Data refers to connecting structured data on the Web. The three key technologies that support Linked Data are – i) URIs (a generic means to identify entities or concepts in the world), ii) HTTP (a simple yet universal mechanism for retrieving resources, or descriptions of resources), and RDF (a generic graph-based data model with which to structure and link data describes relationships). Linked Open Data can be accessed by using SPARQL in a machine-readable format that could immediately be integrated automatically with similar data from other sources, rather than available only to human like online catalogue. Presently interoperability standards are targeting this very important area of automatic content integration.

10) Interoperability is a complex area for technologies. However three visible trends may be identified in interoperability. These are – i) Semantic interoperability by combining Resource Description Framework (RDF), XML and Ontology to express digital objects relationships in a machine understandable manner. The object

relationships is an important element of semantic interoperability; ii) Linked Open Data (LOD) interoperability for integrating LOD into local service framework (presently most of the LOD integration are based on content negotiation); and iii) Multilingual interoperability to achieve cross-lingual and multilingual resources integration at content level and at semantic level.

Unit 3

1) An information retrieval (IR) model is theoretical framework to cover different aspects of information retrieval systems. There are many IR models but these can be grouped fundamentally into two groups – System-oriented models and User-oriented cognitive models. Different IR models have been developed over the years but matching mechanisms form the basis of all these models. IR models can be grouped into two categories on the basis matching mechanisms – i) matching can be done between terms; or ii) between similarity measurement (e.g. distance, term frequency etc.). Term matching is a direct matching of terms derived from or assigned to documents, document representation and queries. Similarity matching is an indirect matching process in which final matching is made on the basis of similarity measurement. For example, in Vector Space model matching is based on the distance between vectors or degree of vector angle. Vector Space model was developed by Salton during SMART experiments related to IR. In this model each term is defined as a dimension while each query or document is expressed as a vector. The complete set of term values in a vector describes the position of the query of document it represents in the space. Almost all the open source text retrieval engines are either using vector space model or modified vector space model.

2) The TREC (Text Retrieval Conference) is an ongoing evaluation project jointly sponsored by NIST and DARPA for – i) encouraging research in text retrieval based on large text data collection; ii) increasing communication between academia and IR practitioners for exchange of research ideas; iii) developing retrieval solution to solve real life problems; and iv) developing evaluating methodologies for IR systems. The TREC TRACKS are dedicated for particular IR problems like cross-language retrieval, multilingual retrieval, natural language processing, query representation, application of filtering in IR etc. All these research datasets are utilizing by developers to improve retrieval efficiencies of OA retrieval systems.

3) OA retrieval system, as digital IRR, is essentially based on database and language of IRR at the core. The search processes support matching of search queries and documents on the basis of metadata and contents of documents through an intuitive user interface. OA retrieval system of any type or size has five basic components – Database (Databases form the core of Web-enabled OA retrieval system. Search process (Database determines what can be retrieved

from the OA retrieval system, whereas search mechanism determines how open access resources stored in databases can be retrieved. Language in IRR (may be grouped as natural language and controlled vocabulary like classification, subject heading and thesauri); and User interface (It is a layer of interaction between users and IRR activities in an OA retrieval system).

4) Bielefeld Academic Search Engine (BASE, http://base-search.net) appeared in public domain in 2004. Presently, BASE indexes more that 52 million OA resources at global scale (number of documents: 52,615,190; number of content sources: 2,776 as on 18.11.2013) and is considered as the largest OA retrieval service. BASE is a feature-rich OA retrieval system and acting as model for other such services. BASE is a perfect combination of Vector-space information retrieval model and its integration with controlled vocabulary (Eurovoc) and subject access system (DDC). Apart from supporting all the required search operators, BASE offers Web 2.0-enabled retrieval, DDC based browsing, many filtering and ranking tools.

5) A text retrieval engine or simply search engine is a tool for contents indexing, searching of index and ranking of retrieved results. These tools can handle both structured data (metadata, cataloguing data etc) and unstructured data like full-text objects. Open source retrieval engines provide enhanced features, scope of customization, continuous up-gradation, rapid use of cutting-edge retrieval techniques and available free of cost. Most of the Green OA software (like DSpace, Greenstone, EPrint etc) and Gold OA software (like Open Journal System, Open Monograph Press) are using open source retrieval engines like Apache-Solr (DSpace version 4.0) Lucene (DSpace upto version 3.2) MGPP (Greenstone version 2.x), Zebra (Koha version 3.x).

6) Apache-Solr may be a good choice as TRE for developing OA retrieval systems. As a part of the Apache Lucene project, Solr provides enterprise-grade full text search engine with high performance search server. It can be integrated with web-service through API. Solr is highly scalable, providing distributed search and index replication. The reasons for selecting Solr in comparison with other open source retrieval engines are as follows – i) can drive more intelligent processing through the use of declarative Lucene Analyzer specifications; iii) CopyField functionality that allows indexing a single field multiple ways, or combining multiple fields into a single searchable field; iv) explicit field types that eliminates the need for guessing types of fields during search; v) external file-based configuration of stopword lists, synonym lists, and protected word lists; vi) many additional text analysis components including term boosting, fuzzy searching, word splitting, regex and sounds-like filters.

7) Controlled vocabularies available in IRR domain may be divided into three groups – thesaurus, subject heading list, and classification scheme. These tools support efficient handling of synonyms, homographs and term association (syntax). Most of the repository

management software like, DSpace, Eprint etc are using standard vocabulary control devices for populating subject access fields.

8) Ontologies help to boost the effectiveness and uniformity of describing resources i.e. they allow for more sophisticated functionalities in IRR. The use of standards, such as the Resource Description Framework (RDF) and Web Ontology Language (OWL), provide structures and methods for descriptions, definitions and relations within a given domain. In OA retrieval systems some of the services support ontology-driven retrieval of knowledge objects. For example, *sciencewise.info* an experimental OA retrieval systems (presently covers Physics, Life Sciences, Humanities and Information Technologies disciplines) provides ontology-driven search interface. A search query is automatically linked with available domain ontology and user allows navigating from one Node to another.

9) Multilingual IR is now quite matured with application of an array of standards like Unicode. But Unicode is only a text encoding standard. We need to apply other standards and tools for developing multilingual IA for open access resources. The requirements can be grouped broadly as - i) system specific requirements and ii) language specific requirements. The first group needs Unicode-compliant Operating System, Text editor, Programming environment and Database management system (Unicode-compliant DBMSs support UTF-8 as standard for native character set). The second set requires language specific tools like Virtual keyboard, Rendering engine and Open type font(s) for respective language. Conjuncts and ligatures are the most font dependent of any scripts. They could be at different positions in different fonts. A rendering engine should be using each font's glyph substitution tables to contextually render the characters. Presently most of the OA repository software like Greenstone, DSpace and Eprint are supporting multilingual contents retrieval.

10) DSpace repository management software (the most popular OAR software) may be configured to support retrieval of contents in any script in the world. The methodology includes three basic steps – i) use of UTF-8 as default character set in back-end RDBMS; (PostGreSQL in DSpace) ii) preparing Java servlet engine to support transaction of multilingual data in UTF-8 encoding (Apache Tomcat is mostly used); and iii) translation of messages and menus (English language messages and menus stored in DSpace in a central place that need to converted into target language). This methodology with these three steps create language-specific user interface in DSpace and supports simple and advanced search and retrieval for contents deposited in DSpace.

KEY WORDS & ABBREVIATIONS

AGLS (Australian Government Locator Service): is Australian government metadata standard intended for the description of government resources on the Web.

Assigned indexing: is an indexing technique where terms are assigned to documents through the use of scheme(s) of controlled vocabulary in choosing appropriate terms.

AuthorClaim: is an initiative related to the unique identification of authors.

BASE: is an exclusive search engine for OA resource developed by FAO though harvesting technology.

Boolean operators: allow terms to be combined through logic operators with AND, "+", OR, NOT and "-" as Boolean operators. These search operators help addition of concepts (AND), separation of concepts (NOT) and inclusion of concepts (OR).

CanCore: a Canadian standard for the implementation of the IEEE LOM metadata standard for describing learning resources.

Cataloguing Cultural Objects (CCO): is a schema for cultural objects developed by the Getty Research Institute.

Citation indexing: is means of information representation by citing and cited authors.

CNRI Handle: is an initiative of Corporation for National research Initiatives (CNRI) to manage unique and persistent identification of digital resources in a heterogeneous network environment.

COAR (Confederation of Open Access Repositories): supports promoting greater visibility and application of research through global networks of Open Access repositories.

Controlled vocabulary: are artificial languages with their own vocabulary (based on literary warrant and users warrant), syntax and semantics such as – thesaurus, subject heading list, and classification scheme.

COUNTER (Counting Online Usage of Networked Electronic Resources): Is the mother project for standardization of usage data and statistics.

CRIS-OAR (Current Research Information and Open Access Repositories): aims to support integration of research administration and open access repositories at the institutional level.

DCAM (Dublin Core Metadata Initiative Abstract Model): is a RDF based framework for the components of resource description and how they relate to one another.

DDI (Data Documentation Initiative): is a standard schema for describing data from the social, behavioral, and economics and statistics domains.

Derived indexing: is an indexing technique where terms are extracted from the original documents. It can also be treated as similar to keyword indexing and no controlled vocabulary is consulted.

DRIVER (Digital Repository Infrastructure Vision for European Research): aims to create an infrastructure for open-access repositories in Europe. It provides a set of best practice guidelines (known as DRIVER guidelines) to build pan-European research infrastructure.

DTD (Document Type Definition): is a mechanism for defining metadata in XML languages, and serve as an alternative to W3C XML Schema.

E-GMS: a schema to ensure maximum consistency of metadata across public sector organizations in the UK.

ETD-MS: is a standard developed by NDLTD to deal with metadata associated with both paper and electronic theses and dissertations.

FGDC: a widely-used, schema for digital geospatial data required by the US Federal Government.

FOAF: is a RDF-enabled schema for describing people and intended to be used on the Semantic Web.

Fuzzy search: search technique that can tolerate errors committed during data entry or query input. This technique can detect and correct spelling errors, errors related to OCRing and text compression.

Information mashup: is integration of more than one sources of data on-the-fly through content negotiation technique.

Invisible vocabulary: handles translation between natural language and one controlled vocabulary stored online.

IRR: Information Representation and Retrieval (IRR) system for organization and retrieval of contents of documents.

ISAD(G) (International Standard Archival Description (General)): is a set of general principles for archival description, throughout the archival management process, and applicable to any type of material irrespective of format or media type.

KE-USG (Knowledge Exchange Usage Statistics Guidelines): is an important initiative in aggregating and transferring usage data from OA journals and OA repositories.

Linked data: is a broad term that refers to a framework of four principles (1. Use URIs as names for things, 2. Use HTTP for providing URIs, 3. Provide useful information, using the standards RDF, SPARQL) and 4. Include links to other URIs.) for exposing data on the Semantic Web and making connections between resources.

Lucene: is an open source text retrieval engine that supports full-text search, faceted navigation, provides hit highlighting utility and allows query language as well as textual search.

METS: Meta data Encoding and Transmission Standard

MODS: Meta data Object Description Standard

NEEO (Network of European Economists Online): provides a set of guideline to aggregate item level usage data based on article identifier and user identifier.

OAI-ORE (Open Archives Initiative – Object Reuse and Exchange): is an interoperability standard for compound digital objects,

OAIS (Open Archival Information System): is a "reference model" schema to support preservation of digital information.

OA-RJ (Open Access Repository Junction): is a protocol to support automatic deposition of multi-authored and multi-institutional knowledge objects into multiple repositories.

ONIX (Online Information Exchange): An international standard for representing book industry product information in electronic form.

Ontology: is a formal, explicit specification of a shared conceptualization or simply a model of organized knowledge in a given domain.

OpenAIRE (Open Access Infrastructure Research for Europe): provides guidelines and standards to integrate OA repositories and OA journals.

ORCID (Open Researcher & Contributor ID): is an open international initiative to provide a registry of unique researcher identifiers at global scale.

PersID: supports persistent identification of knowledge objects through an international infrastructure and knowledge base.

PIRUS (Publishers and Institutional Repository Usage Statistics): is a code of practice for managing usage data and is considered as open international standard in OA usage data.

Proximity search: supports finding words are within a specific distance away.

Range search: allows matching documents whose field values are between the lower and upper bound specified by the range query.

RDF (Resource Description Framework): is a standard model for web-based data interchange.

Relevance ranking: means retrieved results are sorted by relevance which is determined by occurrence of the search term in the title or in other metadata;

RSS (Really Simply Syndication): allows users (after subscribing) to receive any new content added by retrieval system, thus avoiding the necessity of continually visiting sites to check for updates.

SCORM (Sharable Content Object Reference Model): is a **specific guidance for applying metadata to learning resources.**

SDMX: Statistical data and metadata exchange

Snowballing approach: is a search strategy that advices searcher to conduct a search first and then modify the search query on the basis of the retrieved results.

Solr: is an open source text retrieval engine and is presently part of the Apache Lucene project. Solr is a standalone enterprise-grade full text search engine with high performance search server.

String indexing: is a special kind of automated indexing where human indexer creates an input string to summarize the content/theme of a document and computer generates index entries from input string on the basis of rules of respective string indexing system.

SURE (Statistics on the Usage of Repositories): aims to coordinate and aggregate usage data from repositories in Netherlands.

SUSHI (Standardized Usage Statistics Harvesting Initiative): is a protocol designed for the transmission and sharing of COUNTER-compliant usage data from OA service providers.

SWAP (Scholarly Works Application Profile): is a DCMI-compliant application profile for the description of scholarly works, developed by UKOLN.

SWORD (Simple Web-service Offering Repository Deposit): is a lightweight protocol to facilitate multiple deposits in OA repositories and OA services.

Term boosting: allows users to control the relevance of a document by boosting its term or phrase terms (e.g. "resource description"^4 "metadata encoding" means preference of phrase one over the second phrase).

Truncation: A search technique that supports retrieval of different forms of a term but all with one part in common.

Vector-space model: is the most promising information retrieval model that supports similarity matching for retrieval on the basis of distance between vectors or degree of vector angle.

VIAF (Virtual Internet Authority File): is an OCLC initiative to aggregate name authority data from 25 national libraries and to make dataset available as Linked Open Data (LOD).

VOA^{3}R: is a user-centered OA retrieval system in the domain of agriculture and aquaculture developed by FAO and integrated with AGROVOC, time-line view and map view.

Web 2.0-enabled retrieval: means integration of Web 2.0 tools to achieve collaborative, interactive and participative OA retrieval system.

Zebra: is open source text retrieval engine developed for indexing and searching highly structured data such as MARC records, and GILS records and presently in use by the most popular open source ILS Koha.

REFERENCES AND FURTHER READING

Aitchinson,J. & Gilchrist A. (1987). *Thesaurus construction: a practical manual*, 2nd ed. London: ASLIB.

Allen, B. L. (1996). Chapter 7, Information tasks: Interacting with information systems, in I*nformation tasks: Toward a user-centered approach to information systems* (pp.188-200). San Diego: Academic Press.

Anido, L. E., et al. (2002). Educational metadata and brokerage for learning resources. *Computers & Education*, 38(4), 351-374.

Baeza-Yates, R.& Ribeiro-Neto, B. (1999). Modern information retrieval. New York: ACM Book Press.

Barton, M. R., & Walker, J. H. (2002). MIT Libraries' DSpace business plan project: final report to the Andrew W. Mellon Foundation. Retrieved February 10, 2010, from http://libraries.mit.edu/dspace-fed-test/implement/mellon.pdf

Bekaert, J., Liu, X., Van de Sompel, H., Lagoze, C., Payette, S., Warner, S., & Van de Sompel, H. (2006). Pathways Core: A Data Model for Cross-Repository Services. *In* Proceedings of the 6th ACM/IEEE Joint Conference on Digital libraries. New York, NY: ACM Press. Retrieved September 23, 2013 from doi:10.1145/1141753.1141863.

Berners-Lee,T. (2006). *Linked Data Design Issues*. (Widely known as "The 4 Rules of the Web") http://www.w3.org/DesignIssues/LinkedData.html

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, *28*(3), 289-299.

Buzzi, M. and Lazzareschi,P. (2008). A comparison between public-domain search engines, Retrieved September 22, 2009 from http://www.cecmg.de/doc/tagung_2007/agenda07/fileadmin/trilog/download/cecmg_2007/Referenten/Tag1/1C2_CBuzzi_Lazzareschi_Acomparison.pdf

Candela, L.; Castelli, D.; Ferro, N.; Ioannidis, Y.; Koutrika, G.; Meghini, C.; Pagano, P.; Ross, S.; Soergel, D.; Agosti, M.; Dobreva, M.; Katifori, V. & Schuldt, H. (2008). *The DELOS Digital Library Reference Model Foundations for Digital Libraries*. DELOS: a Network of Excellence on Digital Libraries.

Choo,W. C., Deltor, B., & Turnbull, D. (2000). Information seeking on the web: An integrated model of browsing and searching. *First Monday*, 5 (2). Available at http://firstmonday.org/article/view/729/638

Chu, H. (1997). Hyperlinks: How well do they represent the intellectual content of digital collections? Proceedings of the 60th Annual Meeting of the American Society for Information Science, 34, 361-369.

Chu, H. (2009). *Information representation and retrieval in digital age*. New Jersey: Information Today Inc.

Chu, H. and Rosenthal, M.(1996). *Search engines for the Web: a comparative study and evaluation methodology*. Proceedings of the 59th Annual meeting of ASIS, 33, 125-135, Retrieved May 10, 2009, from http://www.asis.org/annual-96/ElectronicProceeding/chu.html

COAR. (2012). The *current state of open access repository interoperability*. Retrieved November 11, 2013 from http://coar-repositories.org

Cooper, W. S. (1973).On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, *24*(2), 87-100.

Cox, K. (1992). Information retrieval by browsing. In Chin-Chi Chen (Ed.), *NIT'92: Proceedings of the fifth international conference on new information technology* (pp. 69-79). West Newton, MA: Micro Use Information.

Crow, Raym (2002). *The case for institutional repositories: a SPARC position paper*. Retrieved September 10, 2013, from http://www.arl.org/sparc/IR/IR_Final_Release_102.pdf.

DINI (2003). Electronic Publishing in Higher Education: How to design OAI interfaces – Recommendations. Retrieved December 22, 2013, from http://nbn-resolving.de/urn:nbn:de:kobv:11-10046033

Fenichel, C.H. and Hogan, T.H. (1981). *Online searching: a primer*. Marlton: Learned Information.

Fenner, M. (2011). Author identifier overview. LIBREAS. Library Ideas, 18. Retrieved from http://libreas.eu/ausgabe18/texte/03fenner.htm

Foskett, A.C. (1996). *The subject approach to information*. London: Library Association Publishing.

Friesen, N. (2002). *E-learning Standardisation: An Overview*. Retrieved November 2, 2010, from http://www.cancore.ca/pdfs/e-learning_standardization

Fugmann, R. (1993). Subject analysis and indexing: theoretical foundation and practical advise. Frankfurt: Indeks Verlag.

Geraci, A (1991). IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries. IEEE Press.

Ghosh,S.B. and Satpathi,J.N. Ed.(1998). *Subject indexing systems: concepts, methods and techniques*. Calcutta: IASLIC.

Gibbon, D. (2004). Archiving language resource objects in XML. *Proceedings of EMELD*. Retrieved December 02, 2013, from http://emeld.net/workshop/2004/gibbon-paper.pdf

Graaf, M. V., & Eijndhoven, K. V. (2008). *The European repository landscape*. Amsterdam: Amsterdam University Press.

Grossman, D.A. and Frider, O. (1998). *Information retrieval: Algorithms and heuristics*. Boston: Kluwer Academic Publishers.

Guha, B. (1983). *Documentation and information: Services, techniques and systems*, 2nd ed. Calcutta: World Press. Pp.36-37.

Hadge, G. (2001). Metadata made simpler. Bethesda: NISO Press. Retrieved May 05, 2003, from http://www.niso.org/news/Metadata-simpler.pdf

Harman, D.(1995).Overview of the fourth text retrieval conference (TREC-4). *In The Fourth Text REtrieval Conference (TREC-4)* (No. 4, pp. 1-24).

Harnad, S. (2005). Fast-forward on the green road to open access: the case against mixing up green and gold, *Ariadne,* no. 42 (2005), Retrieved November 8, 2013 from http://www.ariadne.ac.uk/issue42/harnad/.

Harter,S. P., & Hert, C. A. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. *Annual Review of Information Science and Technology (ARIST)*, *32*, 3-94.

Herner, S. (1970). Browsing. *In* Allen Kent, Horald Lancour, and William Z. Narsi (Eds.), *Encyclopedia of Library and Information Science* (Vol. 3, pp. 408-415). New York: Marcel Dekker.

Hirwade, M., & Hirwade, A. (2006). Metadata harvesting services in India. *Library Herald,* 44(4), 275-282.

IEEE. (2013). *LOM- overview*. Retrieved September 2, 2011, from http://www.cen-ltso.net/main.aspx?put=211

IFLA. (2002). *Digital libraries and metadata resources*. Retrieved December 28, 2013, from http://www.ifla.org/files/assets/preservation-and-conservation/publications/digitization-projects-guidelines-app.pdf

IGNOU (2005). Information processing and retrieval, Ed.by S.B.Ghosh. MLIS study material, MLII-102, pts. I &II, New Delhi: IGNOU.

IMS. (2003). *IMS Global Consortium Home Page*. Retrieved November 2, 2010, from http://www.imsglobal.org

Jacsó, Péter (2004). Thoughts about federated searching. *Information Today*, Vol. 21, Issue 9.

Keen, E.M.(1971). Evaluation parameters. *In* Gerard Salton (Ed.). *The SMART retrieval system: experiments in automatic document processing* (pp. 74-111). Englewood, NJ: Prentice Hall.

Kent, A., Berry, M., Leuhrs, F.U. & Perry, J.W. (1955). Machhae literature searching VIII. Operational criteria for designing uffonnation retrieval systems. *American Documentation*, 6, (2), 93-101.

Koll, M. (2000). Track 3: information retrieval. *Bulletin of the American Society for Information Science and Technology*, *26*(2), 16-18.

Korfhage, Robert R. (1997). *Information storage and retrieval*. John Wiley & Sons, New York, Chichester.

Kowalski, Gerald. (1997). Information retrieval systems: Theory and implementation. Boston: Kluwer Academic Publishers.

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2005). An architecture for complex objects and their relationships. *International Journal on Digital Libraries*, 6(2), 124-238. Retrieved October 11, 2013 from doi:10.1007/s00799-005-0130-3.

Lagoze,C. and Sompel, H.V. (2003). The making of the open archives initiative protocol for metadata harvesting. *Library Hi Tech*, Vol. 21, No. 2 (2003): 118-128.

Lalmas, Mounia, and Anastasios Tombros. (2007). Evaluating XML retrieval effectiveness at INEX. *SIGIR Forum* 41(1):40–57. DOI: doi.acm.org/10.1145/1273221.1273225. 216, 526, 531.

Lancaster, F. W. (1968). Evaluation of the MEDLARS demand search service. Bethesda: U.S. Dept. of Health, Education, and Welfare, Public Health Service, Available at http://collections.nlm.nih.gov/catalog/nlm:nlmuid-0147241-bk

Lancaster, F.W. and Warner, A.J. (1999). *Information retrieval today*. Arlington, VA: Information Resources Press.

Large,A, Tedd, L, and Hartley, R.J. (1999). Information seeking in the online age: principles and practice. London: Bowker-Saur.

Lesk, M. E. (1995). The seven ages of information retrieval. UDT occasional paper. *IFLA Universal Dataflow and Telecommunications (UDT): occasional papers.*

Lesk, Michael. (2004). *Understanding digital libraries*, 2nd edition. Morgan Kaufmann. xxii, 526.

Luhn, H.P. (1958). Review of information retrieval methods. In C.K. Schultz (Ed.), H.P. Luhn: Pioneer of information science, pp. 140-144, New York: Spartan Books.

Manning, C.D., Raghavan, P. and Schutze, H. (2009). *An introduction to information retrieval*. Cambridge: Cambridge University Press.

Marchionini, G. (1995). *Information seeking in electronic environments*. New York: Cambridge University Press.

Marchionini, G. (1999). Chapter 6, Browsing strategies. *In Information seeking in electronic environments* (pp.100-138). NY: Cambridge University Press.

Meadow, Charles T., Donald H. Kraft, and Bert R. Boyce. (1999). *Text information retrieval systems*. Academic Press.

Middleton, C and Baeza-Yates, R. (2007). A comparison of open source search engines, Retrieved July 12, 2009 from http://wrg.upf.edu/WG/dctos/Middleton-Baeza.pdf

Mukhopadhyay, P. (2012). Maching with mashup: application of information mashup for developing open library system. Challenges in Library Management System (CLMS 2012): Proceedings of the national seminar held on 24-25 Feb. 2012 (pp. 39–47). Retrieved October 11,

2013 from http://www.iacs.res.in/conferences/clms/Chakenges_Library_ Management_System.pdf.

Mukhopadhyay, Parthasarathi (2011). *Information retrieval: emerging trends*. Ph.D. Coursework in LIS - Block 4: Unit 2, IGNOU, New Delhi, 2011.

Mukhopadhyay, Parthasarathi (2011). *Search tools and techniques*. Ph.D. Coursework in LIS Block 4: Unit 3, IGNOU, New Delhi, 2011.

Murtha, B. (Ed.). (2002*). Introduction to metadata: Pathways to digital information version 2.0*. Retrieved April 04, 2003, from http://www.getty.edu/research/ institute/standards/intrometadata.html

Nolan, C. W., & Costanza, J. (2006). Promoting and archiving student work through an institutional repository: Trinity University, LASR, and the Digital Commons. Serials Review, 32 (2), 92-98. http://dx.doi.org/10.1016/j.serrev.2006.03.009

OpenDOAR (2013). *Recorded metadata policies.* Retrieved December 28, 2013, from http://www.opendoar.org/onechart.php

Paisley, W.J. (1968). Information needs and uses. *Annual Review of Information Science and Technology*, 3, 1-30.

Pinfield, S., Gardner, M., & MacColl, J. (2002). Setting up an institutional e-print archive. *Ariadne*, 31. Retrieved from May 12, 2009, from http://www.ariadne.ac.uk/issue31/eprint-archives/

Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of information science*, *20*(2), 108-118.

Roy Choudhury, B. & Mukhopadhyay, P. (2012). Organising open access scholarly objects in LIS: a domain-specific harvesting approach. Information and Knowledge Dissemination: Present Status and Future Direction (IKD 2011) (pp. 344–354).

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2012). Open access repositories in Asia: from SAARC to Asian tigers.*Library Philosophy and Practice (e-journal)*. (Paper 808.).

Rusbridge, Chris, and William J. Nixon (2010). *Setting up an institutional ePrints archive—what is involved*?Unpublished paper, UKOLN Meeting.Retrieved July 12, 2010 from http://www.lib.gla.ac.uk/eprintsglasgow.html.

Salton, G. (1981). The SMART environment for retrieval system evaluation–advantages and problem areas. *Information Retrieval Experiment*, 316-329.

Salton, G. (1992). The state of retrieval system evaluation. *Information Processing & Management*, *28*(4), 441-449.

Sarkar, P., & Mukhopadhyay, P. (2010). Designing Single-Window Search Service for Electronic Theses and Dissertations through Harvesting. *Annals of Library and Information Studies*, 57(4), 354-364. Retrieved

December 22, 2013 from nopr.niscair.res.in/bitstream/123456789 /1053/4/ALIS 57(4) 356-364.pdf

Seeley, Y. (2007). Full-Text Search with Lucene. *ApacheCon, May*, 2. Retrieved January 12, 2014 from http://docs.huihoo.com/apache/apachecon/eu2007/lucene_intro.pdf

Shafi, S.M., Gul, S. and Shah, T. A. (2012). Web 2.0 interactivity in open access repositories. *The Electronic Library*, Vol. 31 No. 6, 2013, pp. 703-712

Singh, S., Pandita, N., & Dash, S. S. (2008). Opportunities and challenges of establishing open access repositories: a case study of OpenMED@NIC. *Trends and Strategic Issues for Librarians in Global Information Society: ICCSR Sponsored Seminar* (March 18-19, 2008, Chandigarh) (pp. 98-104). Chandigarh: Panjab University

Singh, V. (2009). A comparison of open source search engines, Retrieved September 22, 2009 from http://zooie.wordpress.com/2009/07/06/a-comparison-of-open-source-search-engines-and-indexing-twitter/

Smith, E.S. (1989). On the shoulders of giants: From Boole to Shannon to Taube: the origins and development of computerized information from the mid-19th century to the present. *Information Technology and Libraries*, 12(2), 217-226.

Sparck Jones, K. (2000). Further reflections on TREC. *Information Processing and Management*, 36(1), 37-85.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, *37*(5), 331-340.

Swets, J. A. (1963). Information retrieval systems. *Science*, *141*(3577), 245-250

Swets, J. A. (1969). Effectiveness of information retrieval methods. *American Documentation*, *20*(1), 72-89.

UKOLN. (2002). *Metadata page*. Retrieved March 13, 2003, from http://www. ukoln.ac.uk/metadata/intro.html

Wielinga B.J. (2001). From thesaurus to ontology. *In* Proceedings of the International Conference on Knowledge Capture, pp. 194-201, New York: ACM Press.

World Wide Web Consortium. (2003*). Metadata and resource description*. Retrieved March 31, 2003, from http://www.w3.org/Metadataoverview.pdf

Yang, Q.S. and Hofmann, M.A. (2011). Next generation or current generation? a study of the OPACs of 260 academic libraries in the USA and Canada. *Library Hi Tech*, Vol. 29 No. 2, pp. 266-300.

This module has been jointly prepared by UNESCO and The Commonwealth Educational Media Centre for Asia (CEMCA), New Delhi.