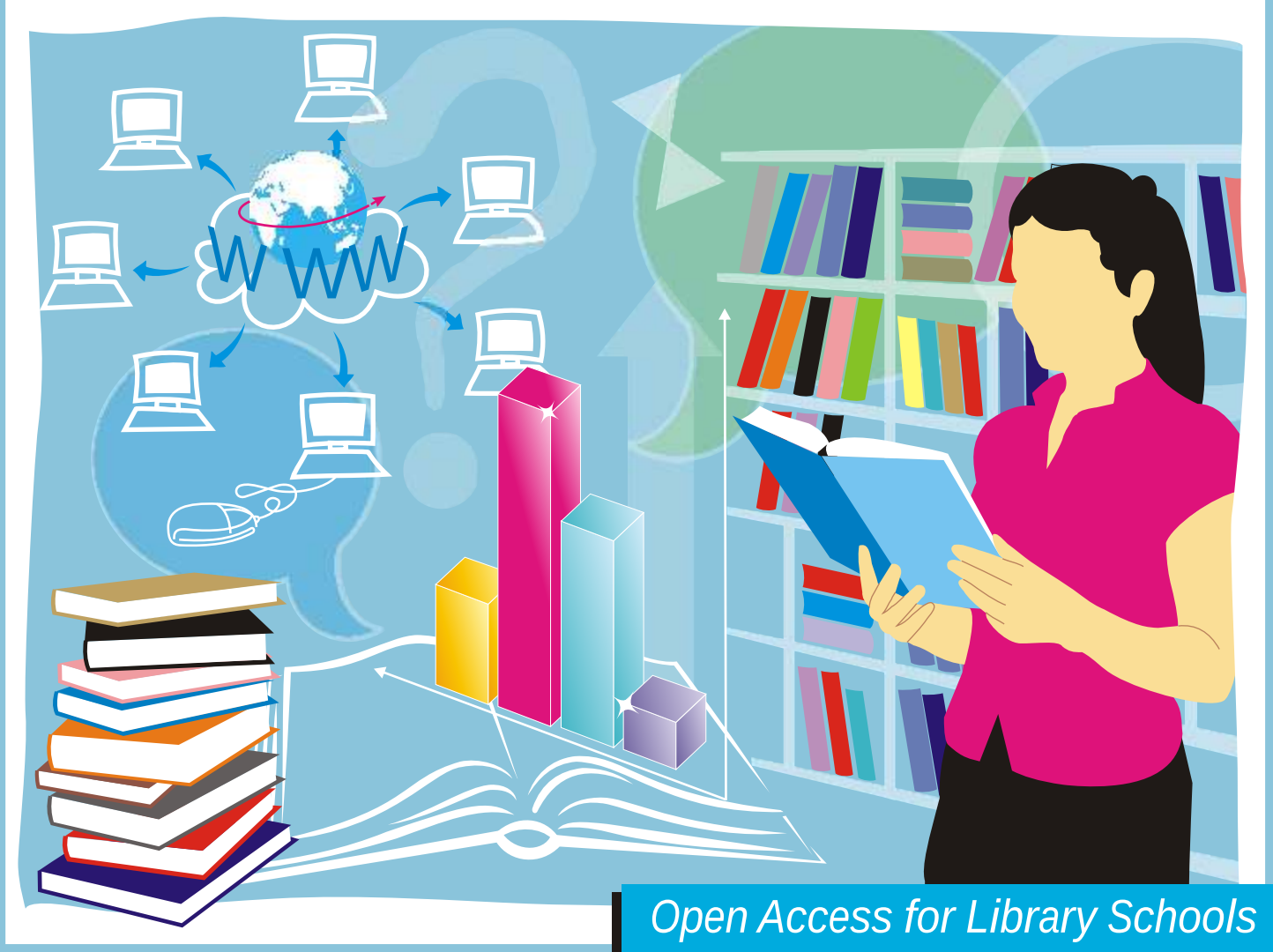
United Nations
Educational, Scientific and
Cultural Organization

3

Resource Optimization



Open Access for Library Schools

United Nations
Educational, Scientific and
Cultural Organization

Resource Optimization

Module

3

Resource Optimization

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

ISBN 978-92-3-100076-8

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

Barnali Roy Choudhury
CEMCA, New Delhi

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

Chief Editor
Sanjaya Mishra
CEMCA, New Delhi

MODULE INTRODUCTION

The concept of open access got momentum since 2000 due to growth in number of scholarly communication, particularly journals, increase in the cost of journals , shrinking budget of libraries and other problems on one hand and the need to access scholarly communications particularly the research out put of public funded research on the other. The access to scholarly communications particularly the journals has been of much concern for a long time.

The open access concept which is a philosophy to achieve the goal of accessing and making available the digital material free of charge which may or may not be free from copyright and licensing restrictions arose out of this necessity. You have been introduced with the concept of open access and open access infrastructure in the previous modules wherein the concepts of scholarly communications, open access and its various forms, issues related with rights management, impact of open access in scholarly communications and technical and management issues have been discussed in the different units of Modules 1 and 2.

This module focuses on resource optimization in general that aims to discuss how the open access environment can be promoted and how the collection development may be facilitated by integrating open access resources with institutional and library resources. At the end of this module, the learner is expected to be able to foster an enabling environment for Open Access, and facilitate collection development by integrating library services.

The module consists of three units. Unit 1 deals with OA mandates and policies; Unit 2 focuses on OA content management; and unit 3 is on harvesting and integration. The Unit 1 which is on open access mandates and policies portraits different policies and mandates at international, national and institutional levels and the related issues. Formulation of Policies/Mandates by the publishers/copy right holders/funding agencies facilitates the wider accessibility of scholarly communications. Through this unit you will be acquainted with sources of OA mandates and policies and analyze the features of some important policies in use. The aim is to prepare you to develop competency to frame a draft OA policy for your institution.

Content management is an important aspect in the context of digital content development, maintenance and access. It is necessary that you be conversant with different aspects of content management such as, it's functional components, the processes by which the content management operates, how the available technologies may be of use etc, particularly in the context of open access resource. An effective content management system should control different workflows right from submission to withdrawal in a participative and collaborative environment. Unit 2 deals with content management in this context of open access and discuss the functional components related to content management; the aspects that need to be critically examined for

different routes of open access; and the cutting edge technologies in OA content management.

The emergence of open access has given rise to development of many distributed repositories following varieties of hardware and software solutions according to the objectives of the repositories. These resulted in problems to the users to access the contents of those repositories individually which may be expensive. To overcome the problems, technological solutions in the form of harvesting have been developed. Unit 3 provides an insight into the harvesting and standards available in the context of open access repositories. The unit focuses on the concept of harvesting, the open standards like OAI/PMH and other harvesting tools, developing harvesting services etc. It is important that open access repositories are integrated with research administrative system and library services. This aspect is also discussed in this Unit.

UNIT 1 OPEN ACCESS MANDATES AND POLICIES

Structure

1.0 INTRODUCTION

The access to scholarly communications particularly the journals is of much concern for a long time. The exponential growth of scholarly literature *vis a vis* their escalating cost and shrinking fund of libraries put severe constraints to their accessibility. The 'Open Access' is a philosophy and the concept arose for solving the problems of accessibility of information available in varieties of forms such as journal articles, books etc. The open access may be defined as a philosophy to achieve the goal of accessing and making available the digital material free of charge which may or may not be free from copyright and licensing restrictions (Ghosh & Das, 2007). Peter Suber[1] defines 'Open Access literature is digital, online, free of charge, and free of most copyright and licensing restrictions'. Formulation of Policies/Mandates by the publishers/copy right holders/funding agencies facilitates the wider accessibility of such communications. Many initiatives have been taken in this regard both at national and international levels adhering to policies formulated by them.

The objective of this unit is to portrait a clear picture of Open Access Mandates/Policies and related issues. The recommendations widely adopted by the open access movement may be summarized as *'deposit immediately, and make open access as soon as legally possible'.* This is an excellent piece of advice for any university or funding agency which is considering adopting a mandatory OA policy.

[1] http://www.earlham.edu/peters/fos/overview.htm

1.1 LEARNING OUTCOMES

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

- Explore the importance of OA mandate and policies;
- Identify sources of OA mandates and policies;
- Analyze the features of some of the important OA policies in use; and
- Frame draft OA policies for your institution.

1.2 POLICIES AND MANDATES FOR OPEN ACCESS

Retaining proper value of the words mentioned in the Berlin Congress and to promote open access, mandate is necessary. The view of Steven V. Hyman, provost of the Harvard University in the context of open access is worth mentioning. "The goal of university research is the creation, dissemination, and preservation of knowledge. At Harvard, where so much of our research is of global significance, we have an essential responsibility to distribute the fruits of our scholarship as widely as possible." Though, it is mentioned in the context of universities, it holds equally good for all agencies generating scholarly communication and to distribute fruits of scholarships as widely as possible, for which developing a policy is of utmost importance. After implementation of open access mandates no special initiatives need to be undertaken to convince researchers to deposit their research outputs soon after publication. It happens naturally under mandatory OA policies

Before a mandatory policy is established, documents dribble in to the repository even many years after the date of publication. Once a mandatory policy is established, the pattern changes dramatically (Sale, 2006). OA experts show in their research works that Mandates/Mandatory Policies adopted by institutions, organizations or funding agencies are populated by contributors quickly and hugely in comparison with the institutions that don't have OA mandate. A graphical representation of increasing parameter of self-archiving is shown in Figure 3.1, which is also re-mentioned later by Richard Poynder in an interview.

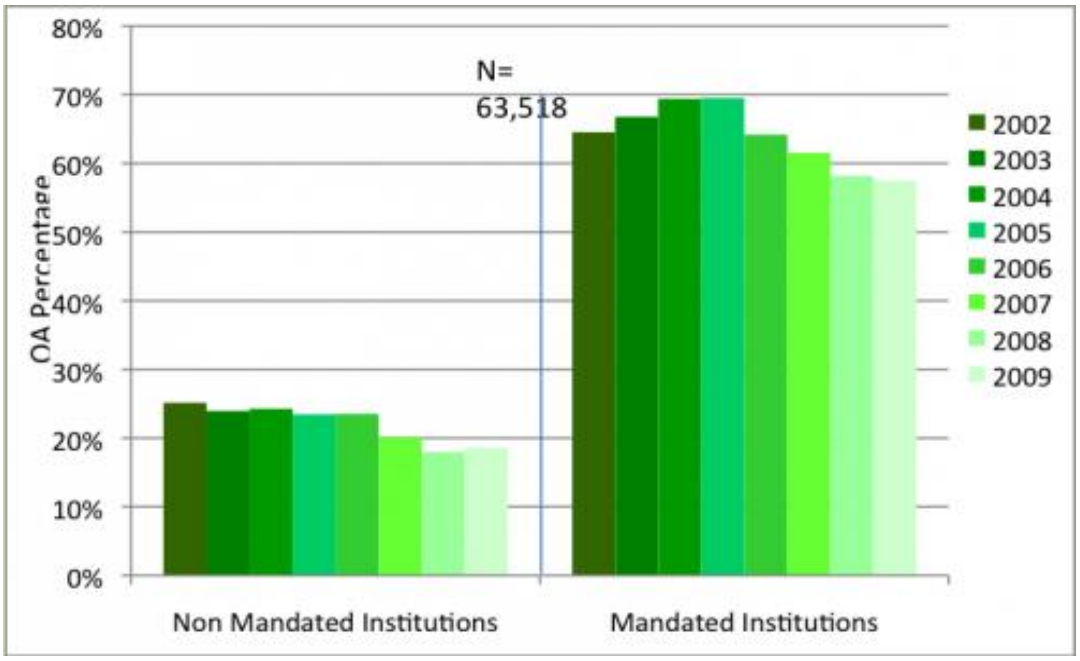


Figure 3.1: Percent of research output that is green OA for institutions where Green OA is or is not mandatory (based on Gargouri et al's 2010 data, as reproduced from Poynder 2011)

This figure shows that open access mandates make article submission in institutional repositories three times more than non-mandated institutions.

Open Access Mandate is a condition/provision that has been taken up by various institutions organizations and funding agencies to make sure the free hand for reusing, remixing, redistribution of scholarly objects. According to Peter Suber, Open Access Mandate is "a condition on a voluntary contract, not an unconditional requirement. It's a reasonable condition as well, since public funders, like NIH, disburse public money in the public interest, and private funders, like the Wellcome Trust, disburse charitable money for charitable purpose." In this context it is mentionable that in his book "Open Access", Peter Suber draws a line between Policy and Mandate. He is much more comfortable to use "Contract" instead of "mandate".

First open access mandate's initiative was taken by the School of Electronic & Computer Science at the University of Southampton, UK in 2002 (Swan, 2012). By this, open access mandate's authors of that school, are bound to deposit their Post-prints in schools repository. In the same year, in 2003, UK parliaments' Science and Technology Committee recommended a funder-based mandate policy in its 2003-2004 report. US House of Representatives voted to set conditions for federal grant recipients. Following the various declarations (Budapest Declaration, Berlin Declaration), and recommendations (viz., World Summit on Information Society, OECD Declaration on Access and to Research Data from Public Funding), " several important research funding bodies have established policies urging their funded researchers to

publish in open access journals"[2]. First institution based mandate/policy was adopted at QUT in Australia. There are now more than three hundred institutions, funding institutions all over the world that have implemented OA policy (Xia et al., 2012) and in 2009-2010 the implementation of open access mandate's rate was higher than ever. The most comprehensive and strong open access mandates of NIH (National Institutes of Health) and the Harvard Faculty of Arts and Sciences were established in the year 2008. *Open*DOAR survey (2006) identified about two thirds of open access repositories did not have publicly stated policies for the permitted re-use of deposited items or for such things as submission of items, long term preservation, etc. In a survey for *Open*DOAR in early 2006, Peter Millington discovered that about two thirds of Open Access repositories did not have publicly stated policies for the permitted re-use of deposited items or for such things as submission of items, long term preservation, etc. This complicates matters for organizations wishing to provide search services, which in turn reduces the visibility and impact of these repositories.

There are two ways of open access - Green and Gold. Open access mandates are applicable for both of these provisions. Till date, there is no exclusive mandate for gold open access. Almost all available mandates are related with green open access.

1.3 TYPES OF POLICIES

There are two types of policies that are prevalent in open access repositories – *voluntary deposit* and *mandatory deposit*.

1.3.1 Voluntary Deposit

This applies to the determination to deposit a research article voluntarily by the author/researcher. Voluntary deposition depends on authors or content creators, who are responsible for scholarly objects. Contributors should be motivated to promote the cause of OA. As Peter Suber argued, "successful policies are implemented through expectations, education, incentives and assistance, not coercion." But there exists no unconditional (must have to make your work open access whether a work is funded or not) Open Access policy or voluntary policy at present except two good policies --Wellcome Trust and NIH (National Institute of Health) (Suber, 2012).

1.3.2 Mandatory Deposit

This applies to the determination to deposit of research articles by the employing institution. Main stake-holder of mandatory deposit is employing institution. There are **three categories** that are identified according to its nature of deposition. These are:

[2] http://openaccess.eprints.org/index.php?/archives/78-guid.html

Immediate Policy

Immediate policy directs authors of the papers,

- to submit their research work (full-text) to repository, which have been accepted for publication in a peer-reviewed journal, immediately after acceptance for publication, if it is any way funded by tax-payers;
- to make its metadata (data about data, like title, author etc.) visible in repository from the time of deposition so that it can prove its existence; but giving respect to authors and publisher interest hold full-text up to embargo period; and
- to make full-text visible after 6- 12 months (recommended) after publication of research paper.

Rights-retention Policy

Simply, by this option, policy makers hold rights to make a research output open access. In this case right is either acquired by policy maker itself by their own policy or by giving grant of waiver to the author/content creator. Second category is identified as "Rights-retention policy with waiver option". In this option, policy makers acquire sufficient rights to make a work open access by giving preference/grants to authors. Harvard University implemented this kind of open access policy.

Loophole Policy

By its nomenclature we can assume that this policy allows making research work open access through the loophole. It means, when author's publisher doesn't permit a work to make open access, this policy finds the loophole (alternative ways) from which deposition as open access is possible. (Suber, 2012).

According to Open DOAR[3], there may be different types of policies as mentioned below relating to individual aspects within overall open access.

- **Metadata Policy** – for information describing items in the repository. Access to metadata; Re-use of metadata.
- **Data Policy** – for full-text and other full data items. Access to full items; Re-use of full items.
- **Content Policy** – for types of document and dataset held. Repository type; Type of material held; Principal languages.
- **Submission Policy** – concerning depositors, quality and copyright.
- **Preservation Policy** Retention period; Functional preservation; File preservation; Withdrawal policy; Withdrawn items; Version control; Closure policy.

[3] http://www.opendoar.org/tools/en/policies.php

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) How many types of policies are prevalent for open access repositories?

……..……..……...

……..……..……...

2) Enumerate the different types of policies relating to individual aspects as suggested by Open DOAR.

……..……..……...

……..……..……...

1.4 ISSUES RELATED TO OPEN ACCESS POLICIES

There are many issues involved in Open access policies. Some of these which need to be looked into are type of open access, terms of deposit, waiver, items to be included in open access and some others. Table 3.1 provides you an idea about the issues involved in OA policies and the coverage of information relating to the individual issues as suggested by Alma Swan.

Table 3.1: Issues involved in OA policies

Sr. No.	Policy Issues of Open Access (OA)	Coverage
1.	Types of Open Access	Policy can cover both 'Green' & 'Gold' route of open access
2.	Items	Articles of journals, Books, Research data
3.	Terms of Deposit	Institutional repositories and Subject specific repositories
4.	Permission to Deposit	Depends on the permission of the copyright holders, authors or publishers
5.	Embargo Period	It may vary from 6 to 12 months.
6.	Waiver	Rights-retention policy provides waiver option. Grants are given as waiver. Authors wish to publish in a particular journal and the publisher will require full copyright to be assigned to the journal.
7.	Compliance	It varies. Institution can control compliance in a most comprehensive way with the help of CRIS (Current Research Information System) and institutional database.
8.	Advocacy	To obtain good result, advocacy programs must encourage potential contributors in terms of awards by show casing the usage and impact statistics.

(Source: Alma Swan: Policy Guidelines for the Development and Promotion of Open Access)

1.5 IMPORTANCE OF OPEN ACCESS MANDATES

In generic sense, open access policy is required to ensure wide, sustainable access of scholarly outputs/contributions. Universities and other funding agencies like Government, corporate houses, international agencies etc. are the main sources as funders of research works. They are starting to make it part of their mandates to ensure scholars to make their published peer-reviewed research output ("publish or perish") available as Open Access (OA), and increase its visibility, accessibility, accountability and impact of scholarly outputs for any one, from anywhere. When we are talking about OA mandates, we are trying to indicate about green open access. OA mandates are mainly applicable to Green OA. Because of obvious reasons, Gold OA requires no such mandate. Authors are intended to publish their scholarly output in open manner. But, one point is to be noted about "Gold" open access mandate-- if an author is willing to publish his/her research work in gold way, then funder organization or institution needs to provide APC (Article Processing Charge(s)) to authors.

In introduction to this unit, it is stated through a graphical representation that open access mandates increase deposition of items in the forms of article, books and data. The reasons to promote *open access mandates (in specific)* are of importance to different stake holders which are:

- Researchers: At the time of pursuing research work, scholars may consult with most of the articles (metadata and full-text) free of cost to use, re-use, and remix subject with the condition of giving proper credits to content creators. In this case, they can also enjoy data which are not even subscribed by their respected institutions or organizations. After accomplishing research work, researchers may publish their thesis in open manner for wide accessibility and greater visibility with the help of OA services to increase the social and academic impact of their research work. Indeed citations are also increasing due to its easy accessibility.
- Funders: Research works are funded by institutions, organizations or government initiations. OA is a win-win situation for all the stakeholders. Increasing citations will give funders and publisher's greater visibility and rich profile. It will improve scholars' social and academic impact. And it will also help to build a strong culture of sharing of scholarly resources and as a whole leading to the betterment of the society.

1.6 IMPLEMENTING OPEN ACCESS POLICIES

Open Access policy may be adopted at three levels - Institutional Level, Funders' Level and Publishers' Level because of their variant nature of infrastructural needs and usage. This section will expose you with adequate information about policies undertaken by these three levels.

1.6.1 Institutional Policy

The first open access institutional policy was adopted by the School of Electronics & Computer Science[4] at the University of Southampton, United Kingdom, in 2000. This policy claims that authors of the schools under the university have to deposit their post-print articles (final version of their peer reviewed version). After this initiative, Queensland University of Technology, Brisbane (2004) and University of Minho, Portugal adopted open access policy. For the current status of different types of policies, see ROARMAP[5]. Some examples of institutional mandate are:

- **University of Leige:** A good policy is implemented by the University of Liege[6] in Belgium in May 2007, which is an institutional immediate deposit. It expects the authors of papers deposit their articles in institutional repository to maximize the visibility, accessibility, usage and applications of their research work.
- **Harvard Open Access Policy:** Another good term of policy has been adopted by Harvard Faculty of Arts and Sciences. This is the example of rights-retention policy with waiver option. Harvard agreement says "*...Each Faculty member grants to the President and Fellows of Harvard College permission to make available his or her scholarly articles and to exercise the copyright in those articles. In legal terms, the permission granted by each Faculty member is a nonexclusive, irrevocable, paid-up, worldwide license to exercise any and all rights under copyright relating to each of his or her scholarly articles, in any medium, and to authorize others to do the same, provided that the articles are not sold for a profit...".*

1.6.2 Funder Policy

Ten years back in 2002, Budapest Open Access Initiative gave first definitions of the basic concept of open access including green and gold roads/routes. After that many important initiatives have taken place to enrich the open access movement by implementing policy or by taking various initiatives voluntarily or in mandatory option. Major funding bodies supporting open access policies include US NIH (National Institute of Health) and Research Councils, UK. In ROARMAP[7] we see there are 85 funders mandate and 12 proposals are there on process. Some pioneer funders' mandates are enlightening the path of open access. Among them Wellcome Trust mandate is most comprehensive one.

Wellcome Trust: Wellcome Trust (U.K.) is a global charitable foundation dedicated to achieving extraordinary improvements in human and animal health. It ensures that the published scholarly outputs of publicly funded

[4] http://eprints.ecs.soton.ac.uk
[5] http://roarmap.eprints.org
[6] http://orbi.ulg.ac.ba/files/extrait_moniteur_CA.pdf
[7] http://roarmap.eprints.org/view/type/funder=5Fmandate.html

research are made freely available in order to use knowledge in a manner that maximizes health and public benefit. *In its mandate, the Wellcome Trust:*

- *expects authors of research papers to maximize the opportunities to make their results available for free;*
- *requires that all research papers funded in whole or in part by the Wellcome Trust be made available via the UK PubMed Central repository as soon as possible, and in any event within six months of the date of publication;*
- *will provide grant holders with additional funding to cover open access charges, where appropriate, in order to meet the Trust's requirements;*
- *encourages—and where it pays an open access fee, requires—authors and publishers to license research papers in such a way that they may be freely copied and re-used (for example for text and data-mining purposes), provided that such uses are fully attributed;*
- *affirms the principle that it is the intrinsic merit of the work, and not the title of the journal in which an author's work is published, that should be considered in making funding decisions.*

European Commission: Recently European Commission has just implemented its OA policy on 13 December, 2013. The Commission expects researchers: *"Each beneficiary must ensure open access (free of charge, online access for any user) to all peer-reviewed scientific publications relating to its results. [Each beneficiary must] (a) As soon as possible and at the latest on publication, deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript accepted for publication in a repository for scientific publications. Moreover, the beneficiary must aim to deposit at the same time the research data needed to validate the results presented in the deposited scientific publications. [Each beneficiary must] (b) Ensure open access to the deposited publication — via the repository — at the latest: i) on publication, if an electronic version is available for free via the publisher, or (ii) within six months of publication (twelve months for publications in the social sciences and humanities) in any other case"*[8].

1.7 COUNTRIES WITH OA LEGISLATION

National legislations related to OA are presently available only with six countries but many developing countries have started giving serious thought on it. The national OA legislations presently in action are:

Ukraine

Ukraine government supports deposit[9] of publicly funded research outcomes in OA repositories. On June 12, 2009, it included an OA endorsement. Seven OA

[8] http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf

[9] http://zakon.rada.gov.ua/cgi-bin/laws/anot.cgi?nreg=537-16

institutional policies have been adopted in Donetsk National Technical University, Kharkov National Medical University, Sumy State University, Ternopol State Ivan Puluj Technical University, and Ukrainian Academy of Banking of the National Bank of Ukraine, V.N. Karazin Kharkov National University, and Charitable Foundation NaUKMA.

Poland

To help small and medium size enterprises to have access to knowledge and innovations, the government of Poland (the Chancellery of the Prime Minister and the Ministry of Science and Higher Education) is working on a legislation to make the results of publicly funded research open access: deposited in open access repositories and/or published in open access journals. You can also watch an interview[10] with Under-Secretary of State, the Ministry of Science and Higher Education Professor Maciej Banach, conducted by Bozena Bednarek-Michalska, Nicolaus Copernicus University Library in Torun and EIFL-OA country coordinator in Poland during Open Access Week (October 2010) (in Polish language).

Spain

Spain has implemented legislation on open access in three levels- national level, regional level (all "7 Universities' repositories based on Madrid. The harvester is called e-ciencia.) and institutional level (15 Institutional Open Access Policies & Mandates). This national open access law 14/2011, of June 1st, on Science, Technology and Innovation under the Article 37 titled "Open Access Dissemination" "*compels the Spanish researcher to archive in an Open Access repository all the scientific publications made under the National Public R&D funding scheme.* "

Brazil

The new Brazilian Access to Information Law, approved by the Senate and ratified by President Dilma Roussef in November 2011, came into force by 16 May 2012. It is a bold step towards greater transparency and involvement whilst providing a stronger framework to embrace access to information. This law is the fruit of the **advocacy** by leading journalists, NGOs and some members of Congress and Government to gain recognition, as a promoter of transparency and open governments, with one of the co-founders of the Open Government Partnership - OGP.

Argentina

Argentina senate passed the law on open access on November 13th, 2013.

[10] http://vimeo.com/15972149

1.8 TOWARDS AN OA POLICY FRAMEWORK

You already know that there are different OA mandates at different levels – institutional, funders and in some cases national. As a library professional you are mainly concerned with the institutional level OA mandate or OA policy. An institutional OA policy should have following components and respective authority needs to take decision on each of these important issues related to the OA policy of the institute. Table 3.2 identifies the design of institutional OA policy.

Table 3.2: OA institutional policy decisions

Policy	Decision to be taken
Archiving Policy	Mandatory or optional; time; form & format
Collections Organization and Management Policy	Organization & management; categories & sub-categories; browsing services
Contents Policy	Type of material; languages
Copyright and Licensing Policy	Right management; licensing pattern
Data Access Policy	Access to items; access pattern; re-use of items
Embargo Policy	Length of time
Metadata Policy	Access to metadata; re-use of metadata; eligible depositors; authentication; schema used
Multilingual Policy	Incorporation of Indic-script based documents; Browsing & searching of multilingual resources; subject access support system
Preservation Policy	Retention period; file preservation; functional preservation; backup
Quality Control Policy	Eligible reviewer; mechanisms
Submission Policy	Eligible contributors; deposition rules; moderation, workflows
System Management and Administrative Policy	Control & management; responsible person; proper location
User Interface	Unicode-compliant multilingual interfaces; mechanisms for browsing & searching multilingual resources
Version Control Policy	Multiple version control; up-gradation; errata and corrigenda lists
Withdrawn Policy	Reasons for withdrawal or removal

Stuart Shieber and Peter Suber has developed a guide to good practices[11] for university open-access (OA) policies (2012), based on the type of policy adopted at several institutions.

☞ Activity I

Go through the recommendations of these major mandates to prepare an institutional open access policy:

- Harvard University: Faculty of Arts and Sciences[12]
- National Institutes of Health (NIH) http://www.pubmedcentral.gov/
- Research Councils United Kingdom[13] (RCUK)
- US White House Office of Science and Technology Policy[14] (OSTP)
- European Commission[15]
- UNESCO worldwide list of funders' mandate[16]
- MELIBEA[17]

……..……...

……..……...

……..……...

……..……...

……..……...

……..……...

……..……...

……..……...

1.9 LET US SUM UP

The concept of open access to scholarly communications evolved during 1990s to facilitate wider communication of scholarly contributions, feedback and use which resulted in the development of open access repositories. Two ways (routes) of open access have been identified in the literature – Green and Gold.

[11] http://cyber.law.harvard.edu/hoap/Additional_resources#Policies_of_the_kind_recommended_in_the_guide
[12] http://dash.harvard.edu/
[13] http://roarmap.eprints.org/671/1/RCUK%20_Policy_on_Access_to_Research_Outputs.pdf
[14] http://roarmap.eprints.org/773/1/ostp_public_access_memo_2013.pdf
[15] http://ec.europa.eu/programmes/horizon2020/
[16] http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/funding-mandates/
[17] http://www.accesoabierto.net/politicas/?idioma=en

The problem of Open Access Repositories (OAR) is that after establishing the system, we need to appeal to contributors to deposit contents for populating the OA system. This is a global phenomenon and the reason is possibly lack of awareness amongst the authors and complexities of copyrights. Mandatory policies are now widely recognized as the only way to achieve close to 100 percent of contents in institutional repositories. Mandates demand exclusive rights. Open access mandates are not exceptional to this. Open access mandate also demands exclusive rights to publish scholarly outputs of researchers to make greater visibility and accessibility. Formulation of Policies/Mandates by the publishers/copy right holders/funding agencies facilitates the wider accessibility of such communications. Many initiatives have been taken in this regard both at national and international levels adhering to policies formulated by them.

This unit portraits a clear picture of Open Access Mandates/Policies and related issues. The recommendations widely adopted by the open access movement may be summarized as *'deposit immediately, and make open access as soon as legally possible'.* The importance of mandates and policies in the context of open access, the various sources of mandates and policies has been discussed in this unit. The features of some important OA policies in vogue and the initiatives taken at national level by some countries have also been discussed. Draft OA policies framed by experts in the field have also been highlighted, based on which you should be able to frame a draft OA policy for your institution.

1.10 ANSWERS TO CHECK YOUR PROGRESS

1) Two types of policies are prevalent now.
2) The policies relating to individual aspects within overall open access policy, as suggested by Open DOAR are:
   a) Metadata policy
   b) Data policy
   c) Content policy
   d) Submission policy
   e) Preservation policy

UNIT 2 CONTENT MANAGEMENT IN OPEN ACCESS CONTEXT

Structure

2.0 INTRODUCTION

A typical content management system is a computerized system that manages submission, publication, modification and retrieval of digital contents in different forms and formats from a central managerial interface. Advanced content management system also controls different workflows right from submission to withdrawal in a participative and collaborative environment. It is necessary; therefore, that you are conversant with different aspects of content management such as, it's functional components, the processes by which the concept management operates, how the available technologies may be of use etc, in the context of open access resources.

An open access (OA) content management system is essentially Web content management system responsible to create, manage, store and deploy open knowledge objects in the forms of text, embedded graphics, photos, video, audio, and research datasets with an aim to support end user retrieval and participation. OA content management system has additional responsibilities to manage copyright and other legalities, retention of authors' rights, privileges control (who submits/access what), version control, preservation, format management for bit streams, purging control (withdrawal of metadata/items), and embargo control. As discussed in unit 1 on OA policies and mandates, the typical functions of OA content management may be summarized as below:

Contents related functions

- Content archiving: Managing forms, formats, file types of OA content management system;
- Resource optimization: Development, organization and maintenance of OA collection;
- Content coverage: Types of OA objects to be included in the system;

- Submitters, reviewers and other quality control matters;
- Metadata encoding;
- Multi-lingual resource management;

Access and rights related functions

- Copyrights and Licenses: Rights management issues and licensing pattern design;
- Embargo: Mechanism to open up resources as OA after a certain time period (either permission by author or publisher);
- Data access issues: Access and reuse of research datasets;
- Withdrawal of OA objects: Issues and mechanisms;
- Metadata reuse and harvesting issues;

Preservation and maintenance related functions

- Format management: Selection of formats for long-term preservation of OA objects, conversion from one format to another format, backup, restoration etc.;
- Version controlling;
- Archiving;

System and users related functions

- Collection managers and management;
- Privileges and authentication management;
- User interface design;
- Resource integration;
- Integration of communication and interaction tools.

2.1 LEARNING OUTCOMES

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

- Describe the scope and importance of content management in OA;
- Identify the functional components related to content management in OA context;
- Critically examine the processes of content management in Green OA and Gold OA environment;
- Explain principles of sustainable development of OA system; and
- Apply cutting edge technologies in OA content management.

2.2 OA CONTENT MANAGEMENT: AN OVERVIEW

By now, you are able to understand from previous unit of this module and from introduction to this unit that content management in open access (OA) is different from generic content management. Here, focus is concentrated on collection development, resource optimization, rights management, preservation, embargo management and similar other issues. However, typical content management functionalities in OA include – i) Collection development (deposition may be mandatory or optional; time frame of submission; forms and formats for OA objects); ii) Resource optimization (organization and management of deposited OA objects; categories & sub-categories; browsing services); iii) Contents coverage (types of OA resources to be included, languages of materials); iv) Copyright and Licensing policies (rights management; design/selection of licensing patterns); v) Data access (managing access to items; access pattern designing; re-use of items); vi) Embargo management (length of time bar in accessing OA objects); vii) Metadata management (metadata encoding, domain-specific schema selection and implementation; access to metadata; re-use of metadata;) viii) Privilege control (eligible depositors; collection level manager selection; authentication;); ix) Multilingual content management (incorporation of mechanisms to support storing, processing and retrieval of multilingual OA documents); x) Preservation of OA objects (retention period; file formats selection for preservation; functional preservation; backup and restoration); xi) Quality management (eligible reviewer; mechanisms of quality control); xii) Ingest (designing submission procedures, selection of eligible contributors; deposition rules; moderation, workflow for submission); xiii) System administration (policy implementation, maintenance, etc.); xiv) User interface (dissemination of OA contents through intuitive interface); xv) Version control (multiple versions control; up-gradation; errata and corrigenda lists); and xvi) Withdrawal of OA contents (purging, access to purged OA resources, reasons for withdrawal or removal etc.).

An OA content manager needs to take decision in each of the above-mentioned issues related with OA content dissemination. So efficient OA content management needs support from global guidelines and best practices.

2.2.1 Designing a Framework for OA Content Management

This section focuses on designing a framework for OA content management on the basis of following guidelines:

- Data Information Specialists Committee -UK guide[18]
- Open Access Information Resources: How We Evaluate Sites for

[18] http://www.disc-uk.org/docs/guide.pdf

Inclusion[19]

- On line OpenDOAR Policy Tool[20]
- OAIS Reference Model[21]
- TRAC checklist[22]
- Open Access Scholarly Publishers Association, Code of Conduct[23]
- Mullen, L. B. (2010). Open access and its practical impact on the work of academic librarians: collection development, public services, and the library and information science literature. Oxford: Chandos Pub;
- Mullen, L. B. (2011). Open access and collection development in academic libraries; digitization to discovery. IFLA Satellite Conference: Acquisition and Collection Development Section. University of the Virgin Islands;
- DOAJ Selection Criteria[24]

The OAIS (Open Archival Information System) reference model[25] proposed six functional groups for digital content management. This model is quite relevant for OA content management and naturally followed by many established OA service providers like OpenAIRE, DRIVER etc. These six functional entities are interrelated to each other (Figure 3.2) and spans from contributor to customer.

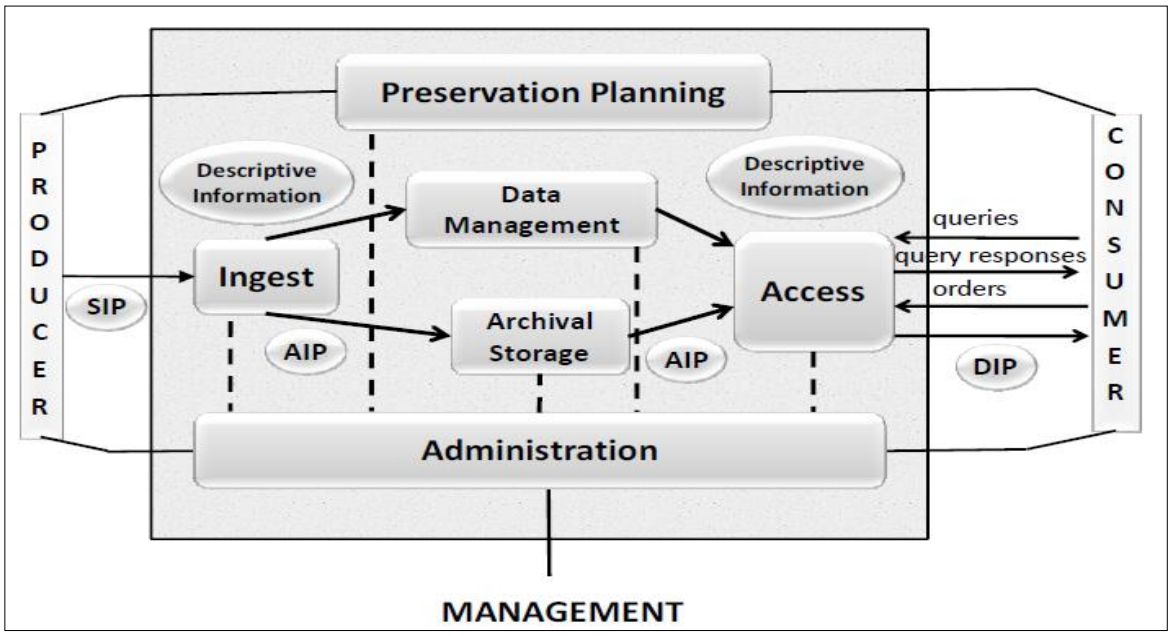


Figure 3.2: Functional entities of OA content management
(Source: OAIS Reference Model, June 2012)

[19] http://www.open. ac.uk/libraryservices/pages/oair/?id=20
[20] http://www.opendoar.org/tools/en/policies.php
[21] http://public.ccsds.org/publications/archive/650x0m2.pdf
[22] http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
[23] http://oaspa.org/membership/ code-of-conduct/
[24] http://www.doaj.org/doaj?func=loadTempl&templ=about#criteria
[25] http://public.ccsds.org/publications/archive/650x0m2.pdf

Resource Optimization

- **Ingest function:** This group acts as a central node for two processes namely SIP (Submission Information Package) & Archival Information Package (AIP) and two functions namely Data Management and Archival Storage. The workflow of the group includes receiving SIPs, checking quality of SIPs, preparing Archival Information Package (AIP) on the basis of formatting and documentation standards, generating Descriptive Information from the AIPs, populating the Archive database, and coordinating connection between Archival Storage and Data Management.
- **Archival Storage function:** The major workflow of this functional group is receiving AIPs from Ingest and appending them to permanent storage. This functional entity also manages storage hierarchy, maintains storage media, supports routine backup activities, allows restoration and disaster recovery capabilities, and provides AIPs to Access functional group.
- **Data Management function:** This group includes tasks to manage Archive database functions including database updates. It also handles queries, provides query responses, and produces reports from these query responses.
- **Administration function:** The broadest functional group of content management that starts from negotiating submission agreements with producers to content retrieval and interacts with all other functional groups continuously. It is also responsible for maintaining archive standards and configuration management of system hardware and software including migrate/update the contents of the Archive.
- **Preservation planning function:** Preservation of OA objects is one of the most important issues in content management. This functional group deals with policy issues related with contents formats, content migration, archival standards, technology environment etc and also performs risk analysis in content migration, templates designing for SIPs and AIPs, and implementation of Administration migration goals.
- **Access function:** It deals with user interface to OA contents. This group includes functions related to receiving users' requests, controlling access to protected resources, executing user’s queries, generating and delivering the responses to users and managing common services to customers.

The Administration functional entity acts as central node for other five functional groups and common services related users. The major information flow inside a content management system is illustrated by OAIS in Figure 3.3.

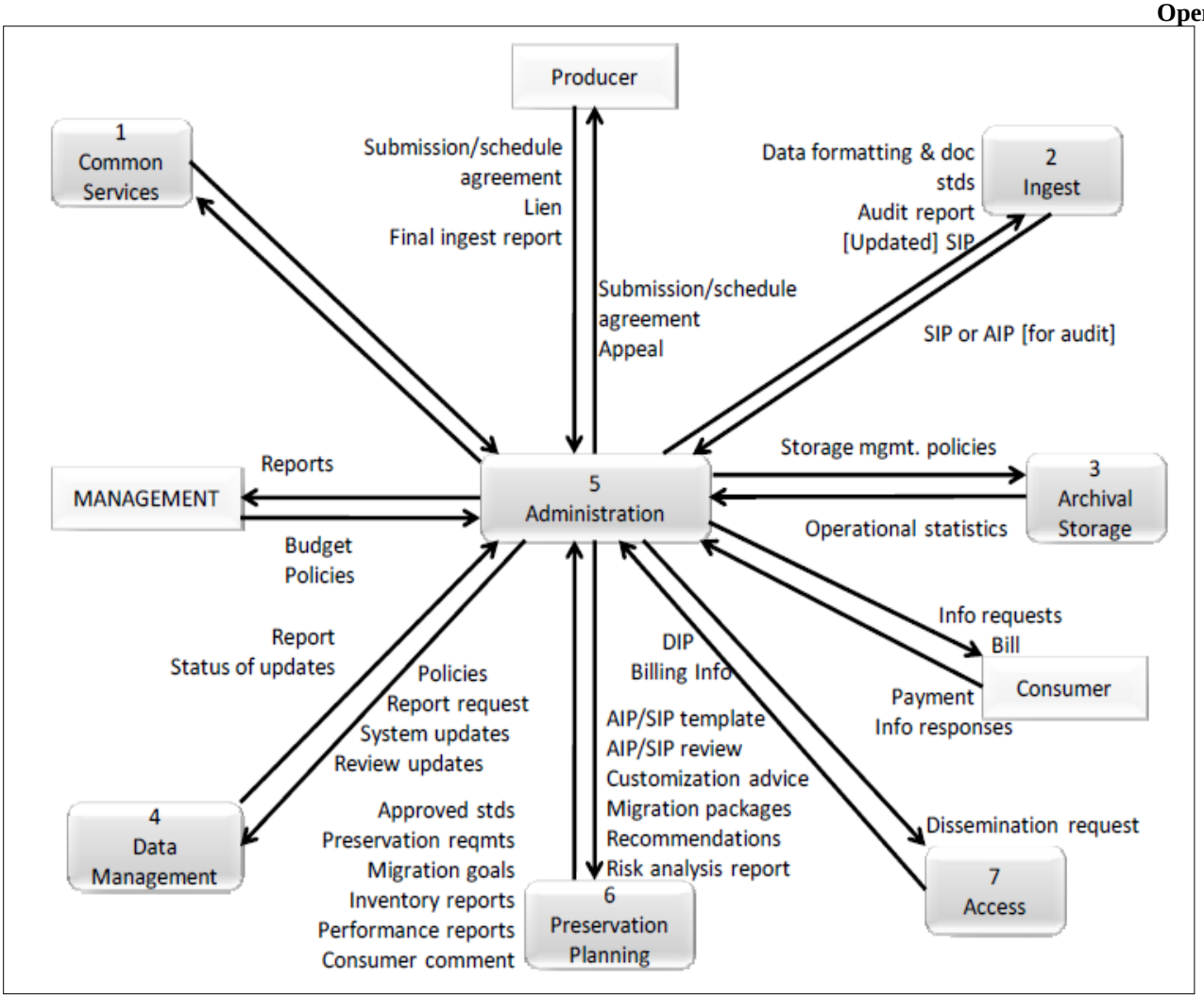


Figure 3.3: Administration of OA content management system

(Source: OAIS *Reference Model, June 2012)*

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) List the major components of an OA content management system.

……..……..……...

……..……..……...

……..……..……...

……..……..……...

2) Discuss major activities related to Ingest in OA.

……...

……...

……...

……...

……...

……...

2.3 OA CONTENT MANAGEMENT: BEST PRACTICES

The DISC-UK Data Share project funded by JISC developed a guide for content management in digital repositories on the basis of best-practice guidelines developed by premier OA service providers. It divides the activities related to content management into seven broad areas and identified factors responsible for efficient dissemination of OA contents. The areas and factors are given below in brief for **your** ready reference (please consult the guidebook[26] for details).

2.3.1 Content Coverage

A repository must be populated with contents and must be managed in an effective and sustainable way. A content coverage road-map will enable this to happen. The types of contents can range from dissertations and articles, to raw research data and data-sets, post-prints (peer-reviewed research articles), book chapters, working papers, theses etc. One of main characteristics of Green OA is the great diversity of contents in repositories. There is no consensus on content types. Different OA repositories have different contents policy (OpenDOAR, 2013; ROAR, 2013). According to OpenDOAR database, nearly about 82% repositories have not defined content policy, and OA repositories mainly contain textual materials (Figure 3.4).

[26] http://www.disc-uk.org/docs/guide.pdf

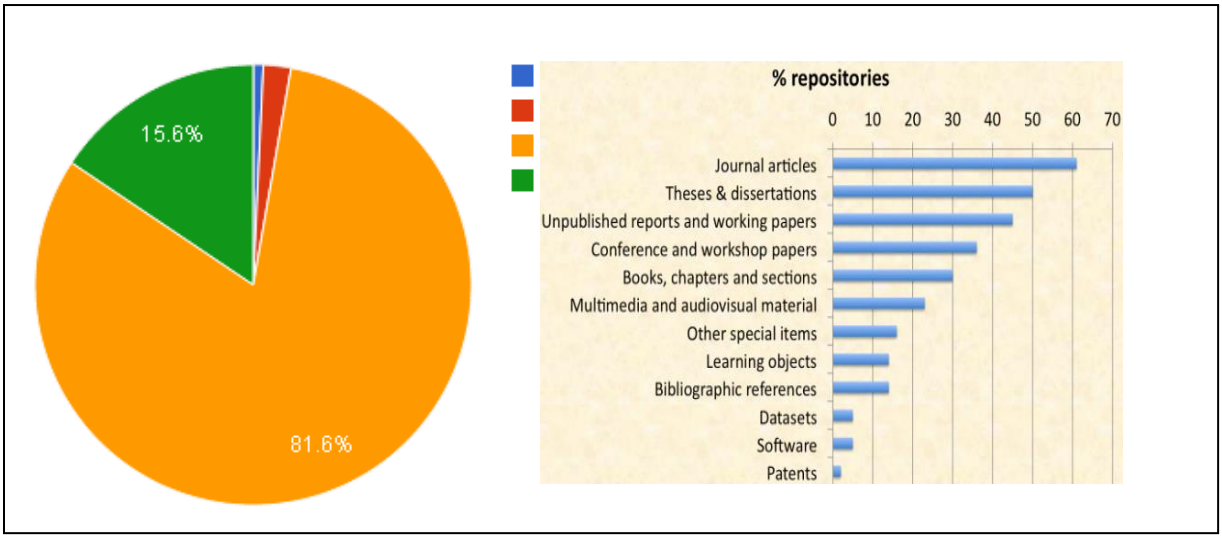


Figure 3.4: Contents in OA repositories
(Source: ROARMAP and *OpenDOAR, December 2013)*

Obviously content management in OA needs to address issues related with:

a) Scope of OA in terms of subjects and languages. The following questions need to be addressed:

- What subject areas will be included or excluded?
- Are there language considerations?
- Will translations be included or required?
- Will text within data files, metadata or other documentation in other languages/English be translated into English/other languages?

b) Kinds of OA research data. Digital research data varied widely – from texts and numbers to audio and video streams. These data would come from:

- Scientific experiments
- Models and simulations (metadata of model and computational data related to model)
- Observations (surveys, censuses, voting records, field recordings, etc.)
- Derived data (processing or combining 'raw' or other data)
- Canonical or reference data (gene sequences, chemical structures etc.)
- Accompanying material

c) Status of the OA research data. It deals with the decisions related to status of data (the research process/life-cycle) to be included in a repository such as:

- 'Raw' or preliminary data?

- Data that is ready for use by designated users;
- Data that is ready for full release;
- Summary/tabular data; and
- 'Derived' data

d) Versions of OA resources. Cross-referencing and version controlling is an important aspect of OA content management. It needs to deal with the following tasks:

- Controlling explicit version numbers as reflected in dataset names;
- Recording version and status of OA resources (draft, interim, final, internal);
- Storing multiple copies of a dataset in different formats;
- Keeping the original copies of data and documentation as deposited;
- Storing supplemental digital objects with the data;
- Recording relationships between items (such as 'supersedes' or 'is superseded by');
- Linking earlier version with later version (identification of the most recent version);
- Ensuring version controlling for different copies of files or materials in different formats; and
- Associating persistent identifier to the latest version

e) Data file formats for OA resources. Selection of the most appropriate file formats for different OA objects and approval of acceptable data file formats from submitters of OA resources is a major area of content management for OA service providers. The conversion of one format into another format is also a mandatory function of OA content management. The following issues need to be addressed:

- Should ASCII files be accompanied by data submitted to system?
- Should spreadsheet files be converted to tab or comma-delimited text?
- Should system accept only file formats that are *de facto* standards?
- Should system allow that specific file formats to be converted into data formats that remain readable and usable?
- Should system accept only 'open' non-proprietary, well-documented file formats wherever possible?
- Should system accept compression formats? (E.g. tar, gzip, zip etc)
- Should system convert proprietary formats to non-proprietary formats?
- Should system create plain text versions of datasets (encoded in either ASCII or Unicode character sets)?
- Should system retain the original bit stream (file) with the item, in addition to its converted formats? and

- Should system accept formats for the purposes of transfer, storage and distribution to users, which do not meet the conditions of long term access?

f) Volume and size limitations for OA resources. Efficient storage space maintenance is another important task of OA content management. It deals with restrictions on the number of files per submission or overall size of the deposited files by contributors. The following factors need to be taken care of:

- Should system restrict OA submission by the number of bytes, or number of separate files, or other conditions?
- Should system use compression software to bundle multiple files (e.g. zip files)?
- Should system apply Storage Area Network (SAN) that supports disk mirroring, backup and restore, archival and retrieval of archived data, data migration from one storage device to another, and the sharing of data among different servers in a network? and
- Should system use Storage Resource Broker (SRB) as a data grid application?

You may also consult following guidelines in managing data formats and data volume:

- PRONOM[27], an on-line information system about data file formats
- Global Digital Format Registry[28] (GDFR)
- JHOVE (JSTOR/Harvard Object Validation Environment[29]
- DROID[30] (Digital Record Object Identification
- Edinburgh Compute Data Facility[31] (ECDF)
- SRB applications in Fedora and Dspace[32]

2.3.2 Content Metadata

Metadata is a crafty area of managing digital archive of any type or size. OA retrieval systems are no exceptions. The Digital Library Foundation (DLF), a coalition of 15 major research libraries, defines three types of metadata which can apply to objects in a digital archive – descriptive metadata, administrative metadata and structural metadata. OA content management system should apply appropriate standards in each of these three areas to ensure adequate description and long term preservation. Descriptive metadata is important for end users to perform retrieval tasks, like searching, browsing, navigating and

[27] http://www.nationalarchives.gov.uk/pronom.
[28] http://www.gdfr.info
[29] http://hul.harvard.edu/jhove
[30] http://droid.sourceforge.net/wiki/ index.php/Introduction
[31] http://www.ecdf.ed.ac.uk
[32] http://www.itee.uq.edu.au/~eresearch/projects/dart/outcomes/FedoraDB.php

collocating OA resources. Administrative metadata is used by OA content managers for maintaining the OA collection, and Structural metadata is generally used by software (at the interface) to compile individual digital objects into more meaningful units. You may refer to Unit 1 of Module 4 for a detail discussion on resource description through metadata applications. However, from the content management point of view following factors need to be considered:

1. Access to metadata
    a) Should system allow anyone to access the metadata free of charge?
    b) Should system restrict access to some or all of the metadata?
2. Reuse of metadata
    a) Should system allow metadata be reused in another medium without prior permission, provided there is a link to the original metadata and/or the repository is mentioned?
    b) Should system allow reusing the metadata for commercial purposes?
    c) Should system ask for formal permission for metadata reuse?
    d) Should system allow metadata harvesting of dataset descriptions by other institutions on the basis of OAI/PMH or OAI/ORE?
    e) Should system determine level of metadata reuse (dataset descriptions or full descriptive metadata)?
3. Metadata types and sources
    a) What descriptive metadata elements should be in use for describing the intellectual content of the object?
    b) What administrative metadata elements should be included to allow a repository to manage the object (scan format, storage format, technical metadata, copyright and licensing information, preservation metadata)?
    c) What structural metadata should be adopted that help to ensure ties aggregation of digital objects to make up logical units?
4. Metadata schemas

    No single metadata element set can satisfy the functional requirements of different types of resources, organizational requirements 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 but other resource types like learning objects, ETDs (Electronic Thesis and Dissertations), research datasets etc are coming in a big way. Therefore, content managers may need to put in place additional metadata schemas to support the Ingest, management, and use of data in OA collections. For an illustrative list of popular domain-specific metadata schemas, section 4.1.5 of Unit 1, Module 4 may be referred to.

You may consult following guidelines in managing OA metadata:

- UK Metadata Guidelines for Open Access Repositories (2013) in its document entitled "Phase 1: Core Metadata (Version 0.9)"[33]
- OpenAIRE Guidelines (OpenAIRE project[34])
- Vocabularies[35] for OA (V40A): An initiative of JISC/UKOLN to develop vocabulary control devices, category lists and authority files for OA resources
- RIOXX: Developing Repository Metadata Guidelines[36], an initiative to define a standard set of bibliographic metadata for UK Institutional Repositories
- Linked Content Coalition[37], an initiative to develop rights managements metadata for OA resources
- NISO Specification for Open Access Metadata and Indicators[38], a NISO initiative to develop standard metadata set specifically meant for OA resources

2.3.3 Content Ingest

Submission of metadata and objects into OA system is technically called *Ingest*. Most of the repository management software includes Ingestion process as a module of the system. OAIS reference model includes Ingest as functional entity. As prescribed by this model, OA content management helps ingestion through services and functions that accept Submission Information Packages from contributors, prepares Archival Information Packages for storage, and ensures that Archival Information Packages and their supporting Descriptive Information become established within the OA system. However major issues related with OA content ingestion are –

- *Eligible depositors*
  a) Should system restrict eligibility by status? If yes, who are eligible for deposition - e-people (registered members), academic staff, registered students, employees of the institution, department, subject community or delegated agents, data producers or their representatives ('self deposit') or only repository staff?
  b) Should system restrict eligibility by content (such as, may only deposit their own work);

[33] http://docs.rioxx.net/guidelines/UK_Metadata_Guidelines_v_1.0.pdf
[34] http://www.openaire.eu
[35] http://www.jisc.ac.uk/aboutus/howjiscworks/ unit2committees/workinggroups/ palsmetadatagroup/v4oa.aspx
[36] http://docs.rioxx.net/guidelines/UK_Metadata_Guidelines_v_1.0.pdf
[37] http://www.linkedcontentcoalition.org
[38] http://www.niso.org/apps/group_public/download.php/9845/Open%20Access%20Metadata%20-%20Work%20Item%20for%20ballot.pdf

c) Must enter descriptive metadata for deposited items; limited to depositing datasets as defined by the repository; may only deposit data of a certain type or subject)?

d) Should system provide a confirmation of receipt to the depositor for submitted item?

- *Moderation by repository*

a) Should content manager review items (for - eligibility of authors/depositors; relevance to the scope of the repository; valid formats; exclusion of spam)?

b) Should system check to ensure that data integrity has been fully maintained during the transfer process?

c) Should system check metadata records for accuracy?

d) Should system implement Digital Object Identifiers (DOIs) or another persistent identifier, such as the Handle system?

- *Data quality requirements*

Responsibility: Generally contributors are responsible for the quality of the digital research data. OA content management system is responsible for the storage quality and data availability. OA system accepts no responsibility for mistakes, omissions, or legal infringements for the deposited objects. OA system may provide licenses to depositors to cover the range of requirements for reuse of the data.

Quality assessment: Sometimes OA system may evaluate data quality for content inclusion on the basis of following parameters:

a) Are the research data based on work performed by the data producer?

b) Does the data producer have a record of academic merit?

c) Was data collection or digitization carried out in accordance with prevailing criteria in the research discipline?

d) Are the research data useful for certain types of research and suitable for reuse?

- *Confidentiality and disclosure*

This area of OA content management is guided by DANS (Data Archiving and Networked Services, The Netherlands). DANS provides Data Seal of Approval that contains guidelines for applying and checking quality aspects of the creation, storage and (re)use of digital research data in the social sciences and humanities. These guidelines serve as a basis for granting a “data seal of approval” (DANS, 2008).

- *Embargo status*

OA content management system should provide agreements about the embargo that include length of embargo and condition that ends embargo on an OA object. The following issues need to be addressed:

a) Should system allow embargo status and length of embargo is determined by OA content manager or by contributors?

b) Should system allow a mechanism where the metadata is publicly accessible but the data are embargoed or restricted in some way?

c) Should system allow to automatically releasing the data on the end date of the embargo or should system manually manage embargo?

- *Rights and ownership*

OA content management must enter into license agreement with the depositor upon submission of OA resource through an in-built or click-through Depositor Agreement. The agreements should at least have three parts – ***rights of the OA system (Repository), rights of contributors (Depositor) and copyrights.***

Repository rights: The issues to be considered for repository rights are

a) Can repository change file format suitable for long-term preservation or otherwise?

b) Is the repository free to change the original submitted material for preservation?

c) Can the repository translate, copy or re-arrange datasets to ensure their future preservation and accessibility, and keep copies of datasets for security and back-up?

d) Can the repository migrate datasets to another repository?

e) Can the repository incorporate metadata or documentation into public access catalogues for the datasets it holds?

f) Will the repository be under any obligation to reproduce, transmit, broadcast or display a dataset in the same format or software as that in which it was originally created?

g) While every care will be taken to preserve the dataset, will the repository be liable for loss or damage to the dataset?

Depositors' rights: The OA content management system should take into consideration the issues like

a) Do depositors retain the right to deposit the item elsewhere in its present or future version(s)?

b) Can depositor place embargo on items submitted to OA system?

c) Can depositor withdraw items from OA system?

d) Can depositor edit metadata of submitted objects?

Copyrights: An OA content management system should ensure following issues (illustrative not comprehensive) related with Intellectual Property Rights (IPR):

a) Content of deposited dataset does not breach any law and does not infringe the copyright of any other person;

b) Any copyright violations are entirely the responsibility of the authors/depositors; In case of copyright violation the relevant item will be removed immediately from OA system;

c) System shall not take legal action on a depositor's behalf in the event of breach of intellectual property rights or any other right in the material deposited;

d) Depositors retain all moral rights to the work including the right to be acknowledged

You may consult following sources for ready reference on the above topic:

- Edinburgh DataShare repository[39]
- Open Data Foundation[40]
- Open Knowledge Foundation[41]

2.3.4 Content Access and Reuse

OA content management system sometimes requires restrictions on use and reuse of OA resources for example, registration in systems to access OA resources, signing a license in downloading OA resources, acknowledgement in adopting and adapting OA resources etc. In most of the cases following three types licenses are in use - *Creative Commons*[42]; *Science Commons*[43]; and *Open Data Commons*[44]. The following aspects relating to access reuse of data and tracking users are to be kept in view:

1. Access to data objects: Following managerial aspects of access to OA need to be considered:
   a) What should be the level of access to OA - at the institutional/departmental level, user registration level, or at the dataset level?
   b) Should there be a fit-to-all access tag or should datasets be individually tagged with different rights, permissions, and/or conditions?

[39] http://datalib.ed.ac.uk/DataShare/Depositor-Agreement.pdf
[40] http://www.opendatafoundation.org/
[41] http://www.okfn.org/
[42] http://creativecommons.org
[43] http://www.dcc.ac.uk/resource/legal-watch/science-commons/
[44] http:// www.opendatacommons.org

c) Should system need to confirm users' acceptance of the terms and conditions of access?

d) What should be the data access method(s) in the system - link to download entire data files? Batch mode access to data? Query-based access to contents?

e) Should users allow to comment or rating OA objects or submit reviews?

f) Should system be integrated with visualization and mapping applications or tools?

g) Should system adopt collaborative, participative and interactive architecture?

2. Reuse of data objects: The main consideration of OA content management in reuse of data is whether or not the user is required to agree to the terms of an on-line Terms of Use statement? The other important factors are:

a) Whether or not the reuse of OA contents (including datasets) be limited?

b) What are the possible limitations (if any) - limitation to non-commercial usages, prohibition to modify data, or other constraints on their redistribution or modification.

c) Whether or not OA system can lift restriction (if any) on a case-by-case basis?

d) What attribution(s) of CC license(s) be adopted by OA system?

e) What is/are the condition(s) that allow redistribution of OA contents at the user end?

f) Will users of the data be required or requested to cite the dataset/s? If yes, what should be the minimum bibliographic data elements?

g) Will there be any restriction on making copies of the data and accompanying materials?

h) Will OA system allow harvesting of full-text or metadata for citation analysis?

3. Tracking users and use statistics: Recoding or tracking user behavior in OA system is useful for planning and improving the system as a whole and at the same time the issue is controversial in nature. Therefore OA content management must be judicious in decision taking. The considerations may be concentrated on:

a) Should OA system track use patterns of individual users through log analysis?

b) What granularity level is required to allow the identification of individual users and their usage pattern?

c) Should OA system adopt policy to determine that to whom and to what extent OA statistics be exposed?

2.3.5 Content Preservation

Content preservation is extremely important to support continuous OA services. As per the guidelines following four factors are important:

1. Retention period: OA system should have managerial policies for the following issues in relation to retention period:
   a) Whether or not OA contents be retained indefinitely?
   b) What should the minimum period of retention?
   c) Should all items will be retained for the lifetime of the repository or retention periods be set for individual items?
2. Functional preservation: Functional preservation solely depends on File Format standard selection to get rid of rapid technical obsolescence of content bit streams. OA system should have mechanisms to ensure usability of OA contents through specific file format support.
3. File preservation: As you know already from previous sections, selection of file format for OA contents and mechanism to convert one file format into another are the two ways to ensure functional preservation. An OA content management should consider following factors in this direction:
   a) Whether or not OA system support various file formats?
   b) What file formats should be adopted for different types of bit streams?
   c) What is the plan and processes for migrations or transformation at the time of need?
   d) Whether or not OA system should support encryption or compression for archival files?
   e) What are the plans and procedures for back-up and restoration of OA contents?
   f) What should be the policy, plan and process for file format migration?
4. Fixity and authenticity: Fixity means a checking on integrity and authenticity of the digital objects. OA content management system must have fixity mechanisms to validate the authenticity of information extracted from a digital object. Fixity mechanisms (such as checksums, message digests, and digital signatures) are used to verify content level integrity during submission, downloading and file transfer. Fixity may be determined at various levels such as - at the points of creation, accession, ingest, transformation dissemination

☞ **Activity I**

Check Fixity issues at PARADIGM PROJECT. (2007) .Metadata for Authenticity: Hash Functions and Digital Signatures. Universities of Oxford and Manchester. Available from: http://www.paradigm.ac.uk/workbook/metadata/authenticity.html

……..……..……...

……..……..……...

2.3.6 Content Withdrawal

Sometime an OA system needs to withdraw contents from a production system. This requires managerial considerations for the following factors:

a) Whether or not items be removed from the repository?

b) What conditions repository should choose to remove items?

c) What are the reasons for withdrawal by repository (copyright violation, legal requirements and proven violations, national security, falsified research, confidentiality concerns etc.)?

d) Should items be removed at the request of the depositor?

e) What should be the terms of the withdrawn items - withdrawn items are deleted entirely from the database; withdrawn items are not deleted, but are removed from display; 'tombstone' citations made available to avoid broken links?

f) What to do with the metadata for withdrawn items; metadata of withdrawn items will / will not be searchable?

2.3.7 Sustainable Development

Confederation of Open Access Repositories (COAR), an active OA promotion agency (with membership of over 100 institutions worldwide from 35 countries and 4 continents) has mission to enhance the visibility through global networks of Open Access repositories. COAR published a guide in June 2013 entitled *Incentives, Integration, and Mediation: Sustainable Practices for Populating Repositories*[45]. This guide advocated eight measures in sustainable OA content management to achieve goals of an OA system. The measures are as follows:

- *Advocacy*: It means promotion of open access at institutional level even for those institutions which have OA policies and mandate;
- *Institutional Mandates*: It means that an institute may make it mandatory for faculty and affiliated researchers to deposit peer-reviewed, scholarly articles published by authors into their institution's open access repository;
- *Metrics*: It is generally observed that usage statistics supplied by

[45] https://www.coar-repositories.org/files/Sustainable-best-practices_final.pdf

repository services can act as a strong incentive for researchers to contribute into OA repositories;

- *Recruitment and Deposit Services*: Content recruitment services like rights checking, and depositing on behalf of authors can be an effective way of populating repositories;
- *Researcher Biographies*: Integration of faculty members / researcher biographies with OA repositories (in order to link the citations with full text content in the repository) can be a successful strategy for populating the repository;
- *Research Information Systems*: Integration of research monitoring system with institutional repository (such as CRIS and DSpace integration (see Unit 2 of Module 4 for details) can be useful for OA content management system;
- *Publisher Agreements*: Orientation services on publisher policies (for example, use of tools like SHERPA/RoMEo in terms of whether, when, and what version authors are allowed to be deposited as OA) can reduce confusion at the contributor's level; and
- *Direct Deposit*: Integration of direct deposit service, which transfers articles directly from the publisher into the institutional repository, may be very useful for OA content management (such as integrating DSpace with OJS via SWORD protocol).

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) What is the need of OA content management system?

……..……..……...

……..……..……...

……..……..……...

4) How OA content management may achieve sustainability in OA development?

……..……..……...

……..……..……...

……..……..……...

2.4 CONTENT MANAGEMENT IN GREEN OA

One of the most important elements of Green OA development is selecting the software system that best satisfies the needs of the institution. These needs will be driven by each institution's content policies and by the various administrative and technical procedures required to implement those policies. Open Society Institute developed a guide for selecting institutional digital repository (IDR) software[46] against some framed parameters. This guide includes the IDR software that satisfies three criteria:

- Available via an Open Source license — that is, they are available for free and can be freely modified, upgraded, and redistributed;
- Comply with the latest version of the Open Archives Initiative metadata harvesting protocols —this OAI compliance helps ensure that each implementation can participate in a global network of interoperable research repositories; and
- Currently released and publicly available — several new systems are currently being developed.

The next important task in Green OA development is building content management workflow on the basis of organizational "Communities" — natural sub-units of an institution that have distinctive information management needs. **Communities** are defined to be the schools, departments, labs, and centers of the institute. Each community can adapt the system to meet its particular needs and manage the submission process itself. Communities can be divided into **sub-communities**, which can be further sub-divided and **Collections** are part of a community or sub-community. **Items** are the actual resources that are uploaded into the IDR. Each item may belong to one collection. Each item contains bit streams that are the computer files which make up the IDR resource.

2.4.1 Selection of Software

A total of seven repository software has been identified from the open source domain that satisfies the above criteria as framed by Open Society Institute. These are:

ARNO[47]

The ARNO project—Academic Research in the Netherlands Online—has developed software to support the implementation of institutional repositories and link them to distributed repositories worldwide (as well as to the Dutch national information infrastructure). The project is funded by IWI (Dutch acronym for: Innovation in Scientific Information Supply). Project participants are the University of Amsterdam, Tilburg University and the University of Twente. The ARNO system was released for public use in December 2003. It is designed to provide a flexible tool for creating,

[46] http://www.budapestopenaccessinitiative.org/pdf/OSI_Guide_to_IR_Software_v3.pdf
[47] http://arno.uvt.nl/~arno/arnodist/

managing, and exposing OAI-compliant archives and repositories. The system supports the centralized creation and administration of repository content, as well as end-user submission. While ARNO offers considerable flexibility as a content management tool, it does not provide a self-contained, "off-the-shelf" institutional repository system. To be able to offer these services ARNO implementers need to deploy other, third party software (e.g. iPort, i-Tor).

CDSWare[48]

CDSWare is maintained and made publicly available by CERN and supports electronic preprint servers, online library catalogs, and other web-based document depository systems. CERN uses CDSware to manage over 450 collections of data, comprising over 620,000 bibliographic records and 250,000 full-text documents, including preprints, journal articles, books, and photographs. CDSware was built to handle very large repositories holding disparate types of materials, including multimedia content catalogs, museum object descriptions, confidential and public sets of documents, etc. Each release is tested live under the rigors of the CERN environment before being publicly released.

DSpace[49]

DSpace is designed by MIT in collaboration with the Hewlett-Packard Company between March 2000 and November 2002. Version 1.1.1 of the software was released in August 2003. The system is running as a production service at MIT, and a federation comprising large research institutions is in development for adopters worldwide. DSpace architecture supports the participation of the schools, departments, research centers, and other units typical of a large research institution. As the requirements of these communities might vary, DSpace allows the workflow and other policy-related aspects of the system be customized to serve the content, authorization, and intellectual property issues of each. Supporting this type of distributed content administration, coupled with integrated tools to support digital preservation planning, makes DSpace well suited to the realities of managing a repository in a large institutional setting.

Eprints[50]

The University of Southampton developed the Eprints software for managing large institute oriented digital archive for scholarly objects. The first version of the system was publicly released in late 2000. The project was originally sponsored by CogPrints, but is now supported by JISC as part of the Open Citation Project and by NSF. Eprints worldwide installed base affords an extensive support network for new implementations. The size of the installed base for Eprints suggests that an institution can get it up and running relatively quickly and with a minimum of technical expertise.

[48] http://cdsware.cern.ch
[49] http://www.dspace.org/
[50] http://software.eprints.org/

Fedora[51]

The Fedora digital object repository management system is based on the Flexible Extensible Digital Object and Repository Architecture (Fedora). The system is designed to be a foundation upon which full-featured institutional repositories and other interoperable web-based digital libraries can be built. Jointly developed by the University of Virginia and Cornell University, the system implements the Fedora architecture, adding utilities that facilitate repository management. The current version of the software provides a repository that can handle one million objects efficiently. The system's interface comprises three web-based services: A management API that defines an interface for administering the repository, including operations necessary for clients to create and maintain digital objects; An access API that facilitates the discovery and dissemination of objects in the repository; and A streamlined version of the access system implemented as an HTTP-enabled web service.

i-TOR[52]

i-Tor—Tools and technologies for Open Repositories—was developed by the Innovative Technology-Applied (IT-A) section of Netherlands Institute for Scientific Information Services (Dutch acronym: NIWI). NIWI calls i-TOR "a web technology by which various types of information can be presented through a web interface," irrespective of where the data is stored or the format in which it is stored. i-Tor aims to implement a "data independent" repository, where the content and the user-interface function as two independent parts of the system. In essence, i-Tor acts as both an OAI service provider, able to harvest OAI compatible repositories and other databases, and an OAI data provider.

MyCoRe[53]

MyCoRe grew out of the MILESS Project of the University of Essen. The MyCoRe system is now being developed by a consortium of universities to provide a core bundle of software tools to support digital libraries and archiving solutions (or Content Repositories, thus "CoRe"). The bundle is designed to be configurable and adaptable to local requirements (hence, the "My"), without the need for local programming efforts. The core contains all the functionality that would be required in a repository implementation, including distributed search over geographically dispersed repositories, OAI functionality, audio/video streaming support, file management, online metadata editors etc.

These seven open source IDR software may be compared by following the framework developed by Open Society Institute[54]. The basic parameters are:

- Standard system features

[51] http://www.fedora.info/
[52] http://www.i-tor.org/en/toon
[53] http://www.mycore.de/engl/index.html
[54] http://www.budapestopenaccessinitiative.org/pdf/OSI_Guide_to_IR_Software_v3.pdf

Resource Optimization

- Hardware, Companion software and Database
- Client, Staff requirements and Installation base
- Repository administration: Installation and Update
- User management
- Content submission management
- Content management
- User interface and Search features

UNESCO recently (2014) released an institutional repository software comparison[55], and you may further like to study the same.

Each of these basic parameters is again divided into number of factors. Let's examine content management features of Green OA software under three most important issues related with OA content management.

Content Management: Metadata

OSI identified six factors under this group:

- Metadata schema
- Support for extended metadata
- Metadata harvesting
- Addition/Deletion of metadata fields
- Set default values for metadata
- Support Unicode character set for metadata

Table 3.3: Contents Management (Metadata)

Parameters / Checkpoints	**Score (1= full support; 0.5= partial support; 0= no support)**					
	CDSware	**Dspace**	**Eprint**	**Fedora**	**GSDL**	**OPUS**
Metadata schema	0.5 (Standard MARC21)	0.5 (Qualified Dublin core)	0.5 (Dublin Core)	0.5 (Dublin Core)	1 (Dublin core, Qualified Dublin core, AGLS,GILS)	0.5 (Qualified Dublin core)
Support for extended metadata	0	1	0	1	1	0
Supports metadata harvesting	1	1	1	1	1	0
Add or delete metadata fields	1	1	1	1	1	1
Set default values for metadata	1	1	0 (Not assigned)	0	0	1
Support Unicode character set for metadata	1	1	1	1	1	0
Total Score	**4.5**	**5.5**	**3.5**	**4.5**	**5**	**2.5**

55 http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/news/institutional_repository_software.pdf

Content Management: Preservation

Preservation is another important issue of OA content management. It is also based on five parameters as prescribed by OSI guide. Generally Green OA software follows three models of OA content preservation namely California Digital Library (**CDL**), IDR of MIT library (**MIT**) and Harvard Digital Repository Services (**HDR**). The parameters are:

a) Models followed (CDR, HDR, and MIT)

b) Format support

c) Approved file format function

d) File format ingested

e) Submitted items can comprise multiple files

Table 3.4: Contents Management (Preservation)

Parameters / Checkpoints	**Score (1= full support; 0.5= partial support; 0= no support)**					
	CDSware	**Dspace**	**Eprint**	**Fedora**	**GSDL**	**OPUS**
Models followed (CDR, HDR, MIT)	1 (C)	1 (M)	1 (C)	1 (M)	1 (H)	1 (H)
Format support	0	1	1	1	1	0
Approved file format function	0	0	0	1	1	0
File format ingested	0.5 (3rd party tool)	1	1	1	0	1
Submitted items can comprise multiple files	1	1	1	1	0	1
Total Score	**2.5**	**4**	**4**	**5**	**3**	**3**

Content Management: Content Export- Import

This particular section has been examined on the basis of the three points mentioned below:

- Upload compressed files
- Volume import for objects
- Volume import for metadata

Resource Optimization

Table 3.5: Contents Management (Contents Export-Import)

Parameters/ Checkpoints	Score (1= full support; 0.5= partial support; 0= no support)					
	CDSware	**Dspace**	**Eprint**	**Fedora**	**GSDL**	**OPUS**
Upload compressed files	1	0.5 (do not uncompress)	1	1	1	1
Volume import for objects	1	1	1	1	1	0.5 (require script modification)
Volume import for metadata	1	1	1	1	1	0.5 (require script modification)
Total Score	**3**	**2.5**	**3**	**3**	**3**	**2**

2.4.2 Content Management Workflow

Most of the above mentioned Green OA software follows OAIS reference model and organizes OA contents into three layers, each of which consists of a number of components. For example, DSpace is based on three layers as follows (Figure 3.5):

- **Storage layer**: responsible for physical storage of metadata and content;
- **Business logic layer**: deals with managing the content of the archive, users of the archive (e-people), authorization, and workflow; and
- **Application layer**: containing components that communicate with the networked world outside of the individual repository software installation, for example the Web user interface and the modules for metadata harvesting service.

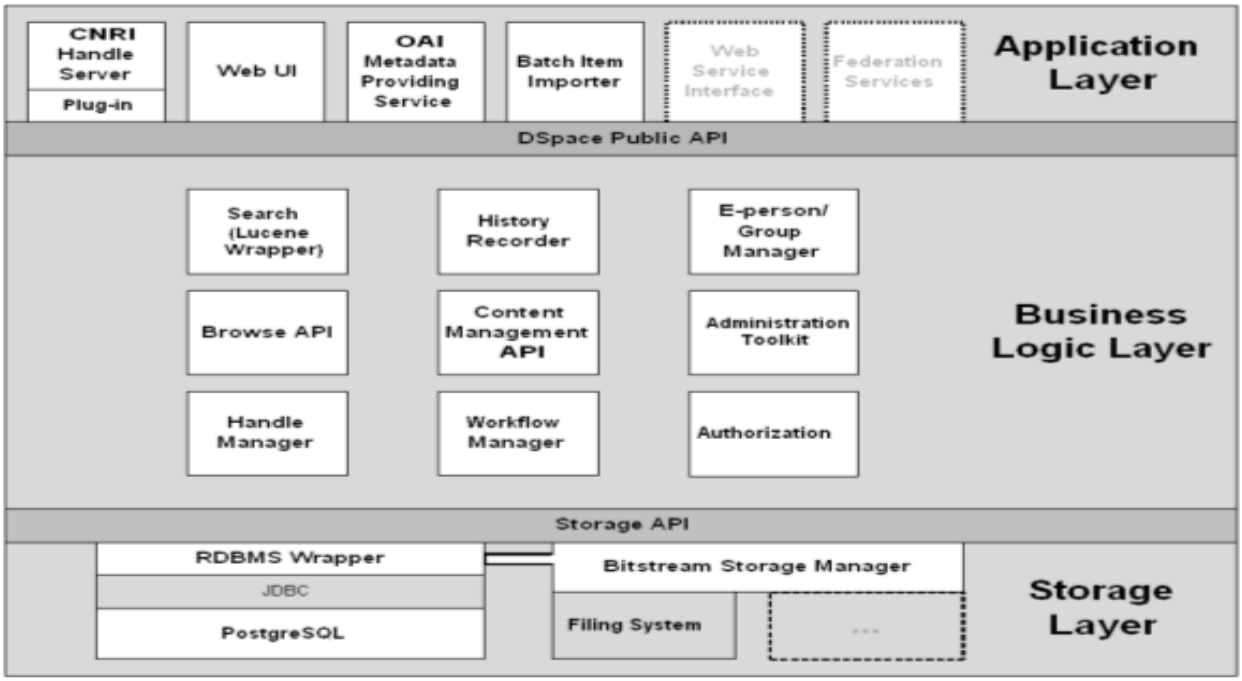


Figure 3.5: Architecture of DSpace

Metadata management

Most of the Green OA software use qualified Dublin Core metadata standard for describing items intellectually (specifically, the Libraries Working Group Application Profile, see unit 1 of Module 4 for details). Only three fields are mandatory: title, language, and submission date. All other fields are optional. Content management deal with/come across metadata in the following modules:

a) Administration modules: Dublin core registry, administrative metadata- default values, mail alert to subscribers;

b) Submission modules: descriptive metadata;

c) Harvesting – OAI-PMH using the DC elements (unqualified); and

d) Search result display: brief and full metadata.

Workflow management

After installation and initial configuration of software, a series of related questions are required to be solved.

a) Who is allowed to deposit items?

b) What type of items will they deposit?

c) Who else needs to review, enhance, or approve the submission?

d) To what collections can they deposit material?

e) Who can see the items once deposited?

All of these issues are part of content management system- contributors, end users and support staff, and are then modeled in a workflow for each collection to enforce their decisions. Generally OA content management starts with defining **e-people** who have **roles** in the **workflow** of a particular **Community** in the context of a given **collection**. Individuals from the Community are registered with OA system, and then assigned to appropriate roles. This workflow can be represented schematically as in Figure 3.6.

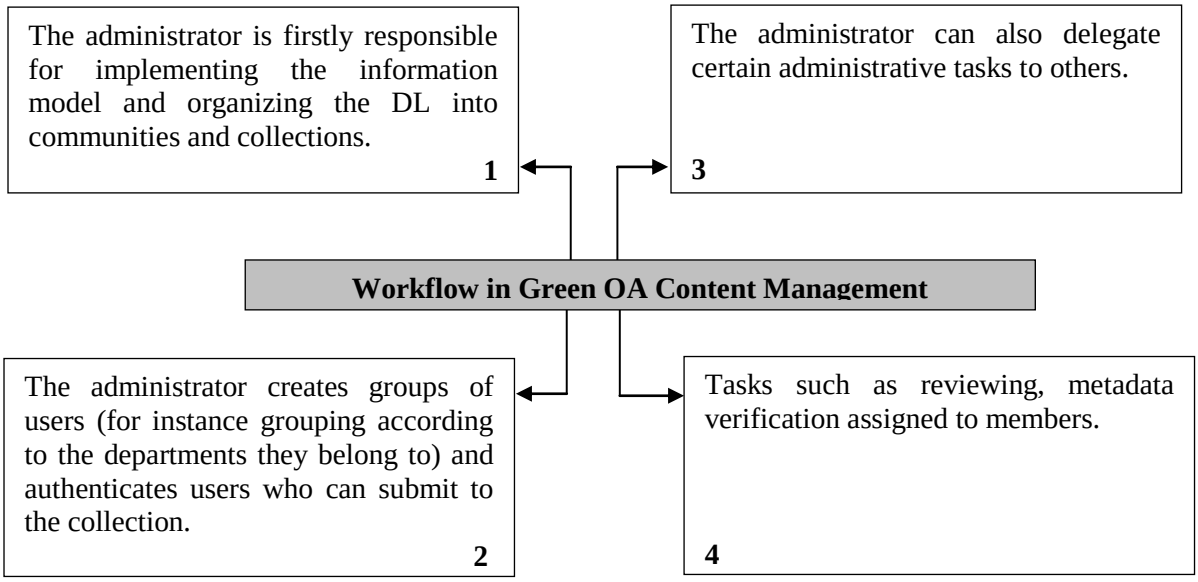


Figure 3.6: Schematic representation of work-flow in OA repository

- Users can register themselves as members. Members can subscribe to entire collections or sub-collections depending on their interests. Mail alerts are sent to the members of each collection whenever a resource is added to that collection.
- Members authenticated by administrator as ‘submitters’ can submit resources to the collection.
- The submitters are required to furnish metadata, basically Dublin Core data, for the resource they are submitting to the IDR. Resources with a multiple files (such as website) can also be submitted.
- The submitters are required to agree with the terms and condition of licensing set forth by the OA system before their submission can be passed on for review process. License information for every resource is stored.
- Content management supports many popular data formats and has the provision for registering new bit stream formats.

Content Identifier

One goal of OA content management is to ensure persistent access to resources so that it is possible to find and retrieve deposited items far into the future. In particular, it is considered crucial that citations to archived material, whether found in printed articles or online, remain valid for long periods. To achieve this goal, DSpace implemented CNRI handles as the persistent identifier associated with each item. The Handle System covers assignment, management, and resolution of these persistent identifiers (or "handles"). Although CNRI has not registered with the IETF for an official namespace, handles are compliant with the IETF's Uniform Resource Name (URN) specification.

User Interface

User interfaces of web-scale repository software generally include different interfaces like:

- One for submitters and others involved in the submission process;
- One for end-users looking for information; and
- One for system administrators.

The end-user or public interface supports search and retrieval of items by browsing or searching the metadata. Once an item is located in the system, retrieval is accomplished by clicking a link that causes the archived material to be downloaded to the user's web browser.

Search and Retrieval

The end user can browse, search and access the collections using the hierarchies and also the alphabetic bar menu. Almost all open source OA

repository software use open source Text Retrieval Engines for content retrieval. OA content management system is known to users for its features related with OA contents retrieval.

Interoperability

As per IEEE, interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged. An OA content management system should support all interoperability areas as identified by COAR (Confederation of Open Access Repositories):

- *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.
- *Content level interoperability*: This refers to the facilities of multiple-deposit process where author submits document in one place and automatically contents transfer from one system to another.
- *Network level interoperability*: This supports development of national and regional repository networks on the basis of metadata harvesting.
- *Statistics and usage data level interoperability*: It supports aggregation and exchange of usage information from different repositories and information systems (like CiteSeer).
- *Identifier level interoperability*: It refers consistency in identification and naming of authors, items, location of items, institutions, funding agencies, grants etc in organizing open access resources.
- *Object level interoperability*: This refers exchange of compound digital objects on the basis of standards for exchange of web resource aggregations.
- *Semantic level of interoperability*: This refers to meaningful exchange of data at machine-level.

(You may refer to Unit 2 of Module 4 for details of interoperability issues related with OA content management).

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) Why an OA content management system should support interoperability?

……..……..……...

……..……..……...

……..……..……...

6) What is workflow in OA content management system?

……..……..……...

……..……..……...

……..……..……...

2.5 CONTENT MANAGEMENT IN GOLD OA

Most of OA journals use Open Journal System (OJS), the open source software for managing and publishing scholarly journals online. OJS is a highly flexible editor-operated journal management and publishing system developed by the University of British Columbia under Public Knowledge Project[56]. The content management system of Gold OA includes following facets:

Site Administration: It **i**ncludes procedures related with installation, configuration and creation of OA journals (administrator can create as many individual journals as are required, and oversee the administration of each journal site that is created.)

Site Management: It involves configuring site-level settings and creating new journals to be hosted within a single site. Journal sites are entirely independent, with the exception of user accounts;

Administrative Functions: It includes tasks related with systems management, Clearing Data Caches, Clearing Template Cache, User management and Role determination.

Journal Management: It deals with all aspects of Journal Management, in consultation with the Editors, including setting up and configuring the journal system, enrolling users in the various roles needed to run the journal, setting up the various sections of the journal, and many other managerial tasks. It involves following procedures:

- Management of Journal Pages
- Users Management
- Roles Management

Submission Process Management: An OA journal provides authors with the ability to upload their submission directly to the journal website. Authors may have the option, on registering as an Author, of also registering as a Reviewer (to be called upon to undertake peer reviews of other submissions) and/or a Reader (to be notified of the Table of Contents new issue of the journal). The submission process includes following content management functions:

- Author Guidelines

[56] http://pkp.sfu.ca

- Submission Requirements
- Indexing and Metadata
- Supplementary Files

Editorial Process: This step consists of a review, typically a blind peer review, followed by a section editor's decision to accept or decline the submission. If accepted in the review stage of the editorial process, the submission then goes through the editing stage which consists of copy editing, layout and proof reading.

Editor's Role

- Submissions (Unassigned, In Review, In Editing, Archives)
- Submission Summary (Submission Management, Submission Metadata)
- Ensuring a Blind Peer Review

Section Editor's Role

- Submissions (In Review, In Editing, Archives)
- Review (Review Version, Peer Review, Editor Decision)
- Editing (Copy editing, Layout, Proof reading)

Reviewer's Role (Submissions, Review)

Author's Role (Submissions, Review, Peer Review, Editor Decision)

- Editing (Copyediting, Proof reading)

Copyeditor's Role (Submissions, Copy editing)

Layout Editor's Role (Submissions, Layout, Proof reading)

Proof reader's Role (Submissions, Proof reading)

Publishing Process: It includes steps related with creating issues and/or volumes for the journal, scheduling submissions to those issues, organizing their Table of Contents, and then finally publishing the issue. The journal can be published in a number of formats. The steps are given below:

- Create Issue
- Schedule Submissions
- Table of Contents
- Publish

Journal Web Site: It includes facilities to improve reading experiences of users. These features are commonly used by online journals and familiarity with them in the case of this journal will facilitate getting the full value out of

other online journals. The procedures include developing following sections of an OA journal:

- Home
- About
- User Home
- Register & Profile
- Current & Archives
- Search & Browse
- Reading Tools

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) Mention the salient points of publication process related to OA journal?

……..……..……...

……..……..……...

……..……..……...

8) What is Role management in OA journals?

……..……..……...

……..……..……...

……..……..……...

2.6 INTEGRATION OF OPEN CONTENTS AND LIBRARY RESOURCES

Web 2.0 tools are increasingly applied in OA content dissemination services all over the world for achieving interactive, collaborative and participative architecture in content retrieval. The software tools and services, which are making the dream of interactive OA retrieval a reality, may be categorized into four major groups (not entirely mutually exclusive):

The Read/Write Web Component: Tools that are leveraging read/write Web include blogs, RSS (Really Simple Syndication or Rich Site Summary), online storage and sharing tools (such as MySpace, Facebook, YouTube, Podcasts) etc.

Social Networking Component: Social networking component includes tools that support community communication and interaction in digital environment. Tools such as instant messaging, discussion forum, event listing (chronological and upcoming), Flickr, Jumpcut etc. are enhancing online socialization through community oriented communication and interaction (Birdsall, 2005).

Collective Intelligence Support Component: Wikis are currently most popular tools for collaborative knowledge sharing, and the best-known example is Wikipedia (http://en.wikipedia.com/wiki/). Other tools such as LibraryThing, PaperBackSwap Second Life, Digg, Technorati, Folksonomy, Social bookmarking, Amazon services are also facilitating the collective wisdom movement in the next generation Web.

Information Mashups Component: Information mashups tools allow remixing of data, technologies or services from different online sources to create new hybrid services through lightweight application programming interface (API).

An OA content management system must take into account the use of Web 2. 0 tools for better user services and management of OA system. Data 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, Facebook, and YouTube) and collaborative tools (like Blog, Flicker, Podcasting). In a total of 1,412 accessible repositories (in 1977 total listed repositories as listed in openDOAR), 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.

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) Categorize Web 2.0 tools as per the intrinsic attributes.

……..……..……...

……..……..……...

……..……..……...

10) Discuss the use of anyone of Web 2.0 tools in OA content management.

……..……..……...

……..……..……...

……..……..……...

2.7 LET US SUM UP

This unit is an attempt to let you know the issues related with OA content management. At the outset this unit mentions the four major areas of OA content management to provide you a panoramic view of complexities of the domain. The Overview section is based on the OAIS reference model for content management and its extension to OA content management. OAIS reference model comprehensively indicates major components of content management as well as procedures related with these identified components. The section on Best practices related with OA content management forms the base of this unit. It explores almost all the managerial issues related with OA contents such as Content Coverage, Content Metadata, Content Ingest, Content Access and Reuse, Content Preservation, Content Withdrawal and Sustainable Development. This unit also provides major workflows related with content management in Green OA, Gold OA and mentions tools for facilitating interactive OA content management.

2.8 ANSWERS TO CHECK YOUR PROGRESS

1) The components of OA management systems are

   a) Ingest

   b) Archival storage

   c) Data management

   d) Administration

   e) Preservation planning

   f) Access

2) The activities of Ingest are to act as central node for two processes namely SIP (Submission Information Package) & Archival Information Package (AIP) and two functions namely Data Management and Archival Storage. The workflow of the group includes receiving SIPs, checking quality of SIPs, preparing Archival Information Package (AIP) on the basis of formatting and documentation standards, generating Descriptive Information from the AIPs, populating the Archive database, and coordinating connection between Archival Storage and Data Management.

3) The OA content management system is needed for accessing, managing and disseminating contents.

   Various guidelines have been formulated which requires a number of issues to be sorted out for developing a system. Some of these are content coverage, content metadata, content Ingest, content access and reuse, content preservation, content withdrawal and sustainability in development.

4) The sustainability to achieve the goals of OA System can be achieved by following 8 guidelines as suggested by ROAR. These are- Advocacy, Institutional mandates, Metrics, Recruitment and deposit service, Researcher biographies, Research information system, Publishers' agreement and Direct deposit
5) Because interoperability provides the ability of two or more systems or components to exchange information and to use the information that has been exchanged. An OA content management system should support all interoperability areas as identified by COAR (Confederation of Open Access Repositories).
6) Contributors, end users and support staff are part of content management system and are modeled in a workflow for each collection to enforce their decisions. Generally OA content management starts with defining e-people who have roles in the workflow of a particular Community in the context of a given collection. Individuals from the Community are registered with OA system, and then assigned to appropriate roles.
7) The publication process of OA journal involves varieties of tasks such as journal management, subscription process management, and editorial process publishing process.
8) The role management is a part of overall journal management process. Role management defines the task /role of various groups of people such as, journal manager, editor, section editor, copy editor, layout editor, proof reader, reviewer, author, reader etc.
9) Based on the intrinsic characteristics, Web2.0 can be categorized as Read/Write Web component, Social Networking component, Collective Intelligence Support component, and Information Mush up component.
10) All categories of Web 2.0 tools have different functions and utilities. For example, Information mashups tools allow remixing of data, technologies or services from different online sources to create new hybrid services through lightweight application programming interface (API).

UNIT 3 HARVESTING AND INTEGRATION

Structure

3.0 INTRODUCTION

You have already learned about the concept of open access in the previous units. The emergence of open access particularly since 2000 has given rise to development of many distributed repositories following varieties of hardware and software solutions according to the objectives of the repositories. These resulted in problems to the users to access the contents of those repositories individually which may be expensive. To overcome the problems, technological solutions in the form of harvesting have been developed. This unit provides you an insight into the harvesting and standards available in the context of open access repositories.

In the present system of publication, scholarly output is obscuring its institutional origins. The reason for it is quite simple - much of the intellectual output and value of an institution's intellectual property is diffused through thousands of scholarly journals and other forms around the world. An institutional repository consolidates the intellectual product created by organizations'/ universities' researchers, making it easier to demonstrate its scientific, social and financial value. These emerging knowledge entities are contributing greatly in developing Open Access Knowledge Movement. Open Access Knowledge System is based on a set of principles and methodologies related to the production and distribution of knowledge objects with the philosophy of openness. Knowledge objects include Data (scientific, technical, historical, geographic or otherwise), Contents (such as journal papers, reports, patents, books, and other artifacts) and General information (including information services). Open access knowledge system can be considered as a superset of open data, open content, open access publishing and open learning resources. It is powered by open source software and open standards. Open

access publishing is the publication of material in such a way that it is available to all potential users without financial or other barriers. An open access publisher is a publisher, in some cases it may be distributors, producing/distributing such materials. Many types of materials can be published in this manner: scholarly journals (known as open access journals), magazines and newsletters, e-text or other e-books (whether scholarly, literary, or recreational), music, fine arts, or any product of intellectual activity (Lagoze & Sompel, 2003).

As a whole the situation is not quite friendly for OA (open access) users. For example, Directory of Open Access Repositories (OpenDOAR) lists a total of 2606 repositories in 2014 on different subjects. These are following varieties of hardware and software solutions according to the objectives of the repositories and local requirements. These resulted in problems to the users to access the contents of those repositories individually which may be expensive. To overcome the problems, technological solutions in the form of harvesting have been developed. It may be applied to both Gold OA and Green OA but presently most of the harvesting services are related with Green path of OA.

To handle all these forms and formats of documents centrally we need a worthwhile technology so that we can manage this vast world of knowledge in a centralized system. In this context harvesting mechanism may help us. So we should have a proper and clear concept of the term Harvester.

3.1 LEARNING OUTCOMES

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

- Explore the concept of harvesting;
- Differentiate between federated search service and centralized search service
- Utilize open standards - Open Archives Initiatives/Protocol for Metadata harvesting(OAI/PMH) and open source harvesting tools;
- Develop harvesting services as may be required for your purpose; and
- Understand the process of integration with existing search services and research administrative system.

3.2 INSTITUTIONAL REPOSITORIES

The term institutional repositories refer to a digital archive that is available online. The purpose of institutional repositories is to collect, preserve and disseminate digital copies of the intellectual output of a research institution for global visibility of the institution's scholarly research.

The field of library and information science is progressing rapidly in digital endeavor. This domain is trying to spread its task of organization, navigation

and dissemination of knowledge (especially in the form of scholarly communication) in favor of high information inflation. But, it is difficult to manage all these sporadic information. The reason being quite simple - much of the intellectual output and value of an institution's intellectual property is diffused through thousands of scholarly communications around the world. An institutional repository concentrates the intellectual product created by researchers, making it easier to demonstrate its scientific, social and financial value. These emerging knowledge entities are contributing greatly in developing open access knowledge movement. Institutional Digital Repositories (IDRs) are digital collections that organize, preserve, and make accessible the intellectual output of a single institution or a group of related institutions (Crow, 2002).

3.3 NEED FOR SINGLE WINDOW SEARCH INTERFACE

The repositories may be different in their coverage, software usage, nature of contents and most importantly in retrieval techniques and tools. As a result, it is difficult for end users search comprehensively these repositories that provide scholarly materials freely. This situation calls for the development of a single window search service covering all the repositories in a given domain of knowledge. Of course the repositories need to be compatible with interoperability standard(s) for building a search service on the basis of harvested metadata. These single windows search services (based on resource metadata) are advantageous to scholars and others as it brings them closer to uniform access interface for scholarly information bearing objects and cultural resources (Cole, 2003).

Directory of Open Access Repositories (OpenDOAR) shows that there are around 2606 open access repositories as on 2014 and Registry of Open Access Repositories (ROAR) lists a total of 3612 institutional repositories. However, in LIS domain most of the IDRs are cross-institutional i.e. these repositories allow submission of scholarly materials globally. Presently, 1685 open access repositories are OAI/PMH compatible in OpenDOAR among 2616 repositories (as on dated 19th March, 2014). Open DOAR repository includes 1554 multi-disciplinary subject fields in 2163 institutional repositories, 283 Disciplinary, 95 aggregating and 74 Governmental repositories.

New services arise when the conditions are favorable. The emergence of the low-cost hardware, standardized software tools, open source software, open interoperability standards, low-cost communication devices, cheap storage devices, and distributed information system (open access journals, open access repositories, subject gateways) provided ideal ingredients for the development of a localized single window search interface for open access repositories (Chudnov, 1999). These single window search services can harvest metadata from different repositories that are open and compatible with interoperability

standards. In such a system users can perform search centrally, display metadata of a resource from local server and retrieve full-text resource from the original server (may be anywhere in the world). As these services can be tuned to harvest metadata selectively, it may help in reducing cross-disciplinary semantic drift during search and retrieval.

3.4 HARVESTERS

"A harvester is a client application that issues OAI-PMH requests. A harvester is operated by a service provider as a means of collecting metadata from repositories"[57]. Harvesting is an automated, regular process of collecting metadata descriptions from different sources to create useful aggregations of metadata and related services.

The concept of harvesting has been developed in the context of processing of metadata. The concept refers to a technique of extracting metadata from individual repositories and collecting it in a central catalogue. Different metadata schema/standards, such as Dublin Core, Text Encoding Initiatives (TEI), Metadata Encoding and Transmission Standards (METS) and many others have been developed. Though every standard has been developed to suit a particular need, nevertheless, these can be used by others too. It becomes difficult for user groups to get access to the literature of different repositories following different standards. This resulted in the development of harvesting technology to facilitate search by the user groups.

Metadata harvesting refers specifically to the gathering together metadata from a number of distributed repositories (e.g. eprint archives) into a combined data store.

3.5 INTEROPERABILITY AND CROSSWALK: STANDARDS AND TOOLS

Interoperability means the provision of exchanging data without minimal loss of content functionality of multiple systems (with different hardware & software platform and data structure interface).

A crosswalk is mapping of the elements, semantics and syntax from one metadata schema to those of another so that metadata created by one community can be used by another group that employs a different metadata standard. By Crosswalk, it is possible to use metadata created by one community by another group that may employ different metadata standards. It is useful for virtual collections where resources are drawn from varieties of sources and expected to act as a whole.

Interoperability and crosswalk ensures exchange of bibliographic data among heterogeneous systems across the globe. Automated and digital library systems

[57] http://www.openarchives.org/OAI/openarchivesprotocol.html

are now supporting various standards and protocols like Z39.50, OAI/PMH, METS (Metadata Encoding and Transmission Standard), MARC-XML, SRU (Search/Retrieval via URL protocol) and SRW (Search /Retrieve Web service protocol) to achieve interoperability. In other words, interoperability is the ability of systems, services and organizations to work together seamlessly toward common or diverse goals. In the technical arena it is supported by open standards for communication between systems and for description of resources and collections, among others. Interoperability is considered here primarily in the context of resource discovery and access. The domain of LIS services uses extensively two interoperability standards – Z 39.50 and OAI/PMH. These two interoperability standards are different in nature. OAI/PMH deals with metadata harvesting whereas Z 39.50 is a protocol for distributed search services. There are some similarities of distributed search services and centralized search services in case of distributed object type, bibliographic world view and object presentation through data provider. Similarly these two interoperability protocols are different in some aspect of searching and semantic mapping. Z39.50 search is basically distributed search whereas OAI based searching is basically centralized searching. In case of Z39.50 protocol search is done by data provider and in OAI protocol search is done by service provider. Semantic mapping is done at the time of searching in Z39.50 protocol and in OAI it is done after metadata delivery.

Now, let's explore the concept of harvesting in order to fetch metadata from OAI-compliant repositories. With the better use of harvesting mechanism we can create centralized search service to consolidate OA contents deposited in OARs. OAI-PMH is based on HTTP, XML and supports unqualified Dublin Core. So implementing such kind of low barrier protocol is quite effortless at the service provider end. OAI is supporting distributed network information services. In that case any organization can create their own harvesting system to fetch data from distributed service providers across the world and may facilitate search service centrally by processing in local server. Let's see how OAI-PMH mechanism works upon.

3.6 OAI-PMH MECHANISM

Harvester is a client application which is operated by a service provider to collect metadata from repositories. Repositories are accessible by networked infrastructure by the means of 6 OAI-PMH requests (popularly called OAI verbs) that act as content negotiation mechanism between data providers (holder of metadata) and service provider (gatherer of metadata). So there are two classes of participants in the OAI-PMH framework (Lagoze & Sompel, 2003) - in one side there is service provider powered by harvester or harvesting software and on the other side there is data provider backed by repository. The OAI/PMH is a light-weight standard protocol for harvesting metadata records from 'data providers' to 'service providers'. It provides some rules to harvest the metadata of a repository not the full contents. The contents

should be retrieved form source repository. Figure 3.7 shows that how a request is given by a service provider to the data provider.

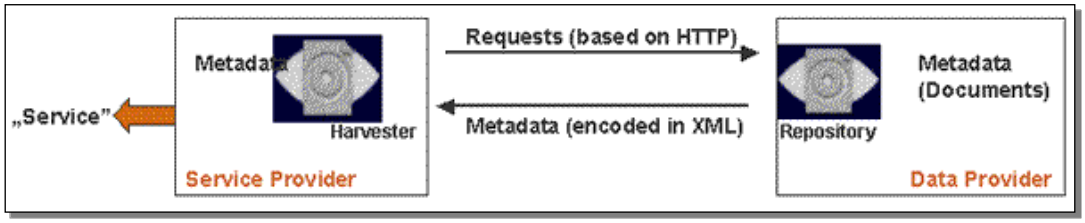


Figure 3.7: Two classes of OAI-PMH providers

- **Service Providers** use metadata harvested via the OAI-PMH as a basis for building value-added services; and
- **Data Providers** administer systems that support the OAI-PMH as a means of exposing metadata.

Any one of the following harvesters can be used for harvesting metadata from data providers to service providers using six OAI verbs (Sutradhar, 2013):

- Arc[58]
- Citebase[59]
- CYVLADES[60]
- DP9[61]
- DLESE OAI Software[62]
- OaI Repository Explorer[63]
- OAIster[64]
- OASIC[65]
- OAIHarvester[66]
- MeInd[67]
- METALIS[68]
- My OAI[69]
- Perseus[70]
- Public Knowledge Project-Open Archives Harvester[71]

[58] http://arc.cs.odu.edu/
[59] http://citebase.eprints.org/cgi-bin/search
[60] http://www.ercim.org/cyclades/
[61] http://arc.cs.odu.edu:8080/dp9/index.jsp
[62] http://dlese.org/oai/index.jsp
[63] http://re.sc.uct.ac.za/
[64] http://oaister.imdl.umich.edu/0/oaister
[65] http://oasic.ccsd.cnrs/fr/
[66] http://www.oclc.org/research/software/oai/harvester.htm
[67] http://www.meind.de/
[68] http://metalis.cilea.it/
[69] http://www.nestrl.org/
[70] http://www.perseus.tufts.edu/cgi--bin/vor
[71] http://pkp.ubc.ca/harvester

Resource Optimization

Repositories are always managed by data provider that makes OAR open to harvesting. OAI-PMH distinguishes between three distinct entities viz., Resource, Item and Record. Service provider send request by using HTTP protocol and Data Provider responds in XML syntax. Request epitomes are issued as GET or POST methods over HTTP protocol. In this mechanism a service provider may fetch OAI-PMH compliant documents from different data providers and data provider may also act as aggregators. It may be mentioned that a repository can act as service provider and data provider at the same time as well as only service provider or data provider.

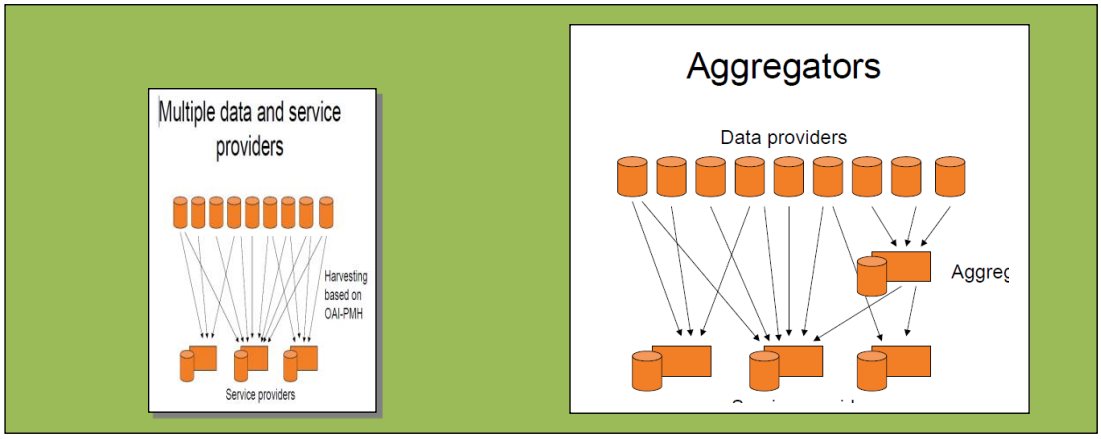


Figure 3.8: OAI-PMH data providers and service provider

Figure 3.8: Shows the functions of service providers, data providers and aggregators. At the time of harvesting, Service provider sends queries to the data provider in term of six OAI-PMH request/verb as shown in Figure 3.9.

Identify (return general information about the archive and its policies);

ListSets (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);

ListRecords (retrieves metadata records for multiple items); and

GetRecord (returns the metadata for a single item in the form of an OAI record).

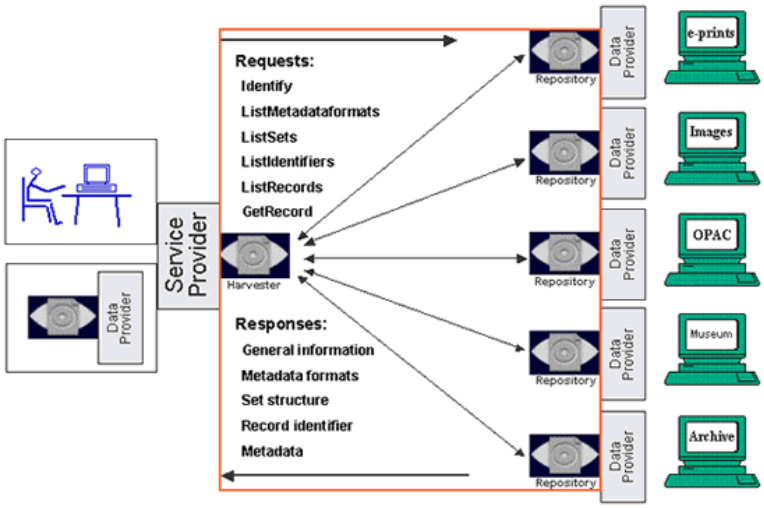


Figure 3.9: OAI-PMH Structure Model[72]

[72] http://www.oaforum.org/tutorial/english/page3.htm

Figure 3.10 shows you a model representation of request or verb at the time of fetching data from a repository.

Figure 3.10: OAI-PMH request model[73]

The Open Archives Initiative Metadata Harvesting Protocol (OAI/PMH) supports interoperability and sharing of metadata across an array of institutions. 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. Open source OAI harvesting tools provide opportunities to make the difficult job an easy one. As mentioned earlier, there is an array of open source harvester software (compatible with OAI/PMH V.2).

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[74]). Please see Table 3.6 in section 3.7 for a comparison of open source harvesting software.

3.7 DESIGNING HARVESTING FRAMEWORK

Design and development of harvesting framework requires an array of steps, strategies and planning. The three major components of such a framework design are – i) development of software architecture; ii) selection, installation and configuration of harvesting tool; and iii) selection of repositories and collection of essential attributes for harvesting (title, resource URL, base URL, mail id of administrator etc.). The prototype harvesting framework may be developed by you at your library based on open source software and open

[73] ibid
[74] https://diva.cdlib.org/projects/harvesting_crawling/recall_crawl/spider_eval.pdf

standards. It uses Linux as operating system, Apache as Web server, MySQL as RDBMS i.e. LAMP architecture as base, and PKP version 2.X as harvesting tool.

Table 3.6: Comparison of open source harvesting software

Parameters/Criteria	Metadata Harvesting Software			
	Arc	***DLESE***	**OAICat**	**PKP**
1. OS related				
1.1 Windows				
1.2 Unices				
1.3 Platform independent	Y	Y	Y	Y
1.4 Others				
2. Software architecture				
2.1 LAMP based				Y
2.2 Java based	Y	Y	Y	
2.3 Others				
3. Protocol related				
3.1 Santa Fee				
3.2 OAI/PMH ver 1.0	Y	Y	Y	Y
3.3 OAI/PMH ver 2.0	Y	Y	Y	Y
3.4 Others				
4. Harvesting process related				
4.1 Data provider-Service provider	Y	Y		Y
4.2 Aggregator	Y			
4.3 Others				
5. Harvesting administration				
5.1 Metadata schemas support	Multiple	Multiple	ETD-MS, DC	Multiple
5.2 User registration/creation	Y			Y
5.3 Independent archive manager				
5.4 Site submission by users				Y
5.5 Theme selection				
5.6 Layout design interface		Y		Y
5.7 Language interface selection				
5.8 Crosswalk creation				Y
5.9 Plug-in management		Y		Y
6. Retrieval related				
6. A ***Browsing facility***				
6.A.1 Browsing by metadata elements		Y		Y
6.A.2 Sorting by metadata elements	Y			Y
6.B ***Searching***				
6.B.1 Simple search	Y	Y	Y	Y
6.B.2 Advanced search	Y	Y	Y	Y
6.B.3 Field-level search	Y	Y		Y
6.B.4 Search operators support				Y
6.B.5 Control for display of results		Y		
6.B.6 Web 2.0 features	Y			Y

The requirements of PKP harvester are as follows:

- PHP >= 4.2.x (including PHP 5.x); Microsoft IIS requires PHP 5.xMySQL >= 3.23.23 (including MySQL 4.x/5.x)
- Apache >= 1.3.2x or >= 2.0.4x or 2.0.5x /Microsoft IIS 5.x or 6.x
- Operating system: Any OS that supports the above software, including Linux, BSD, Solaris, Mac OS X, Windows (preferably NT based Windows flavors)

As a whole, the use of open source software in developing domain specific harvesting system depends on a structured methodology. The steps related with the creation of the harvesting framework may be divided into three major groups (Mukhopadhyay, 2010).

Group I: LAMP related activities

PKP harvester 2.X is based on AMP architecture. Naturally, you have to install Apache, MySQL and PHP prior to installing PKP harvester. Although there is no hard and fast rule, the installation sequence of this manual follows the order below:

Apache (The Apache httpd server is a powerful, flexible, HTTP/1.1 compliant open source Web server)

- Installation of Apache;
- Testing of Apache; and
- Apache Configuration and Control.

PHP (PHP is an open source server side scripting language)

- Installation of PHP
- Configuration of PHP

MySQL (MySQL, the most popular Open Source SQL database is developed, distributed and supported by MySQL AB)

- Installation of MySQL
- Initialization of MySQL Server
- Creation of database, user and manage permission

Testing of AMP Links through Scripts

- Testing PHP-Apache Link
- Testing PHP-MySQL Link

Group II: Harvester related activities

Selection

The first task is to select appropriate harvesting software. A preliminary study identifies a total of four open source (as a library professional you already know the advantages of using open source software) harvesting software on the

Resource Optimization

basis of their user bases. The final choice of software may be based on the selection framework given in the table below. The framework is based on six major parameters – Platform of OS; Architecture of harvesting software; Protocol support; Harvesting processes; Administration of harvesting processes; and Retrieval features.

Installation

This activity includes two major tasks – i) installation of PKP harvester and ii) configuration of PKP harvester. The installation process of PKP harvester is quite straight forward. It requires two sets of information – a) login name and password for the administrator and b) database details (name of the Mysql database, user of database and password of the database user). The configuration processes are divided into three groups – a) site management (configuration of site specific details, language, crosswalk, plug-ins and reading tools); b) Archives (creation of archives, managing created archives); and c) other administrative functions (layout, customization etc.).

Group III: Repository related activities

The most important task of the administrator is to setup archive(s) for metadata harvesting. You can start with a selective numbers of OAI/PMH compatible open access repositories. The intrinsic attributes of these repositories are given in Table 3.7.

Table 3.7: Attributes of some open access repositories

Name of open access repositories	DLIST Digital Library of Information Science & Technology	LDL Librarians Digital Library	ELIS – Eprints on LIS	ERPAePRINTS Electronic Resource Preservation and Access Network ePRINTS service
Sponsoring Institute	School of Information Resource & Library Science, University of Arizona(UA), US	Documentation Research and Training Centre (DRTC), Indian Institute, Bangalore,. India.	AePIC(Advanced e-Publishing Infrustructure), CILEA(Consorzio Interuniversitario Lombardo Per I' Elaborazione Automatica), Italy	Electronic Resource Preservation and Access Network (ERPANET), United Kingdom.
No of records	693 items (2009-03-13)	490 items (2013-12-17)	15819 items (2014-01-07)	86 items (2014-01-07)
Software in use	EPrints	Dspace	EPrints	EPrints
URL of the repository	http://dlist.sir.arizona.edu/	https://drtc.isibang.ac.in	http://eprints.rclis.org	http://eprints.erpanet.org/
OAI/PMH base URL	http://dlist.sir.arizona.edu/perl/oai2	http://drtc.isibang.ac.in/oai/request	http://eprints.rclis.org/perl/oai2	http://eprints.erpanet.org/perl/oai2
Document type	Articles; Reference; Conference; Theses; Unpublished books; Learning Objects; Specials.	Articles; Conferences; Theses; Multimedia.	Articles; References; Conferences; Theses; Unpublished books; Learning Objects; Specials.	Articles; Conferences; Theses; Unpublished books.
Language	English	English, Hindi, Kannada	English, Italian, Spanish	English

Let us say we are going to develop a harvesting service for above-listed repositories in library and information science. The BASE URL or OAI/PMH URL we may collect from OpenDOAR. Registry of Open Access Repositories[75] lists around 15 LIS specific repositories which allow us to search & list open access repositories by subject, country and content type. After selecting suitable repositories (you want to fetch from), give OAI-PMH base URL and fetch respective repository. To fetch a repository, there are few steps relating to Open Harvesting Software which is based on PKP software.

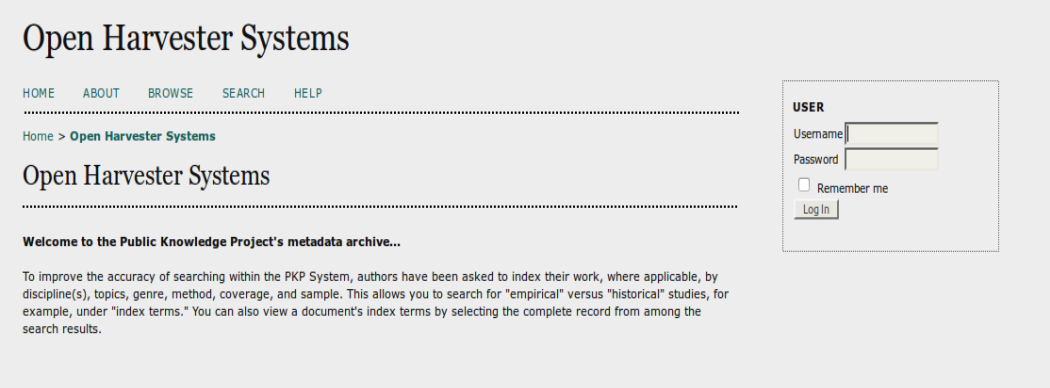


Figure 3.11: Open Harvesting System, UniLIS

It will facilitate you to access administrative home page of UniLIS (example in Figure 3.11) from where you can add archive or manage your archive. You can also do site management and administrative functionalities from this home page (Fig. 3.12).

Figure 3.12: Site administration page of UniLIS

[75] http://www.roar.eprints.org

Resource Optimization

Let's start with *add archive*. Our main motto is to create a domain specific institutional repository. After clicking on "Add Archive" link, you will get this interface (Figure 3.13).

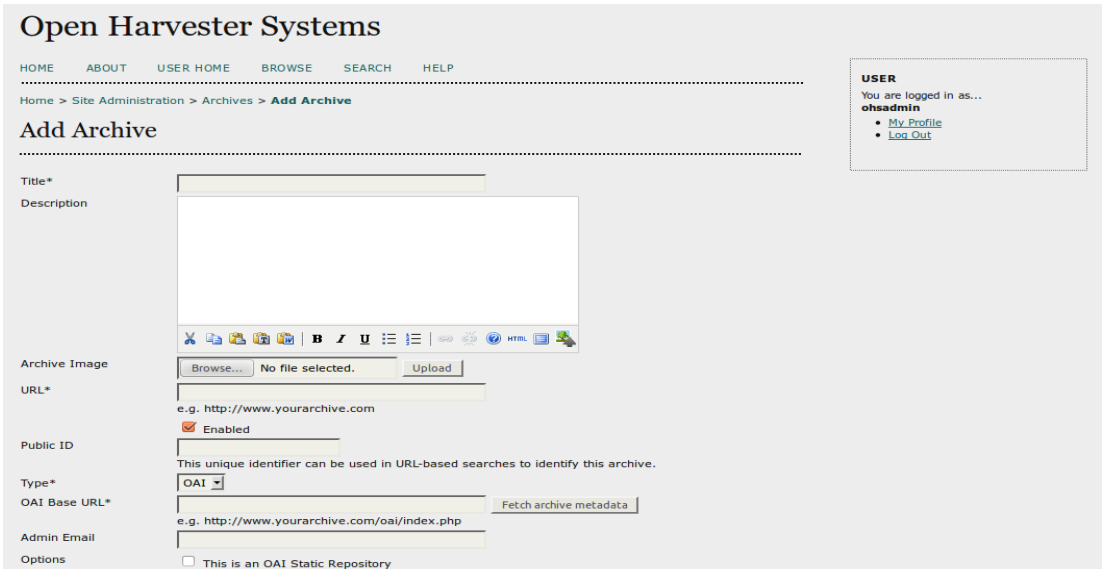


Figure 3.13: Add Archive Interface

Here, you have to fill required fields to fetch a repository. The basic fields are:

- Name of open access repositories;
- Sponsoring Institute ;
- No of records;
- Software in use ;
- URL of the repository;
- OAI/PMH base URL ;
- Document type; and
- Language

On the basis of these given data of a particular repository, harvester will fetch data in its own archive. Figure 3.14 shows how it looks like after fetching data from repository. For example we fetched E-LIS and QUT ePrints repositories in UniLIS due to its good number of valuable records.

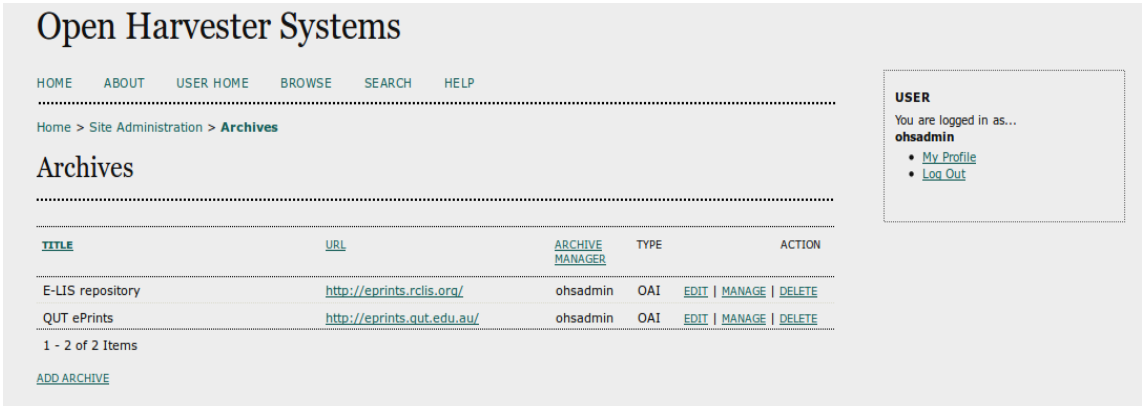


Figure 3.14: Archives

From this interface we can edit, manage or even delete this repository (Figure 3.15). Let's think positive and manage this repository by clicking on manage link. At the moment, you are clicking on manage link by which you will get this interface where you have the provision to "update metadata index" for "All sets". It will take time depending on the speed of the network (a few minutes to a few hours) in order to fetch full repository. It is possible to harvest in selective order like "by collection", "by date range" etc.

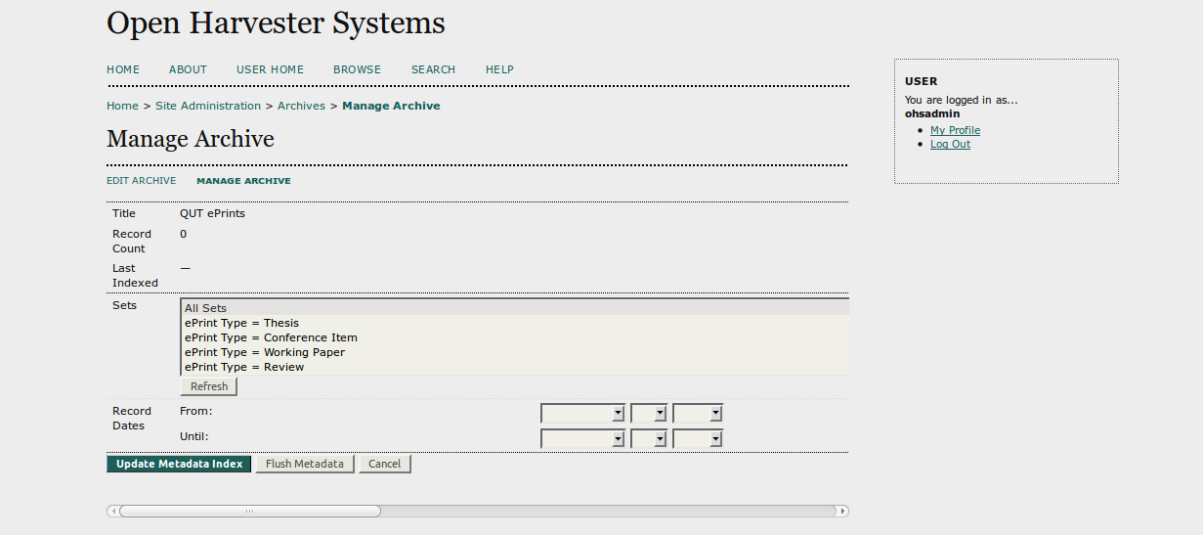


Figure 3.15: Manage Archive interface

After updating, the harvester will show the listed repositories including number of records that are fetched in its archive (Figure 3.16). Repositories are now browsable and searchable by end-users. One can browse a particular repository by clicking on it.

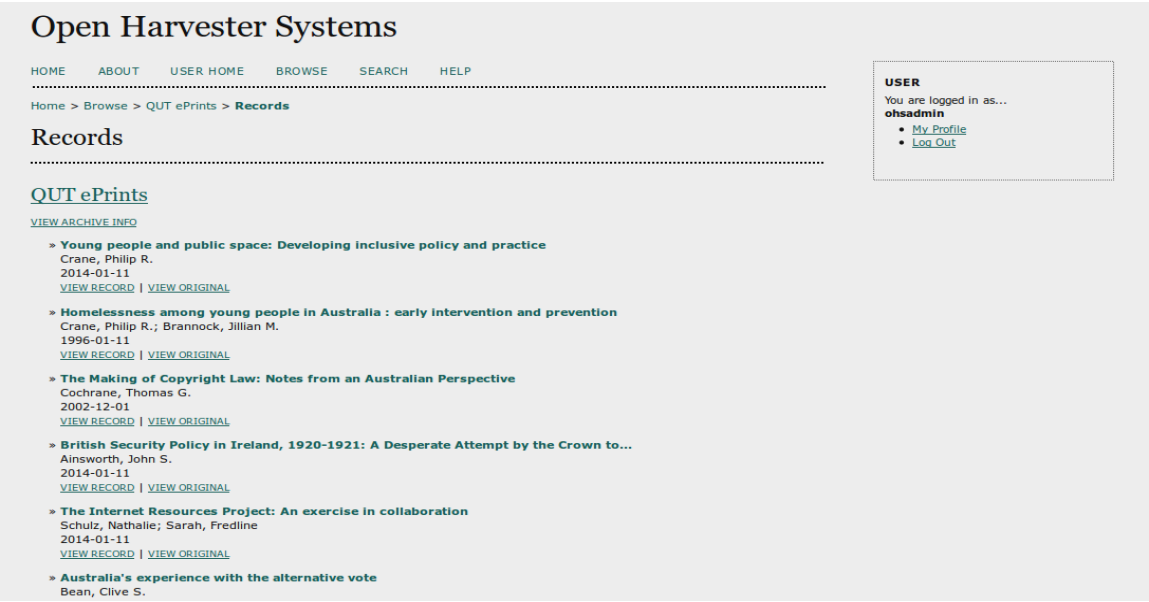


Figure 3.16: Records of QUT ePrint

Resource Optimization

There is also search option from where you can search by repository, contributor, coverage, date etc. And from here any one can see metadata related information centrally by clicking on "View Record" (archive information) and can access it in global platform by clicking on "View Original" (from repository information) (Figure 3.17).

Record Details

- My Profile
- Log Out

Young people and public space: Developing inclusive policy and practice

QUT ePrints

VIEW ARCHIVE INFO

FIELD	VALUE
Title	Young people and public space: Developing inclusive policy and practice
Creator	Crane, Philip R.
Subject	160512 Social Policy 160810 Urban Sociology and Community Studies 160702 Counselling Welfare and Community Services Youth Young people Public space Inclusion
Description	Issues about young people's use of public and community spaces are now commonly raised in many countries. As urban space becomes more intensely used and the patterns of use of various types of space changes so a range of tensions have emerged for a range of parties including local government, shopping centre management, youth services and young people themselves. (This article is based on a paper delivered at the International Conference on Young People and Social Exclusion, University of Strathclyde, Glasgow, 10 September 1999.)
Publisher	The Centre for Youth Work Studies
Date	2000
Type	Journal Article
Format	application/pdf
Relation	http://eprints.qut.edu.au/3/1/Crane_syij_article.pdf Crane, Philip R. (2000) Young people and public space: Developing inclusive policy and practice. Scottish Youth Issues Journal, 1, pp. 105-124.
Identifier	http://eprints.qut.edu.au/3/
Rights	Copyright 2000 (please consult author)

Figure 3.17: View Records Interface

So far harvesting works well with Green OA or OARs that include pre/post print versions of journal articles, theses, reports, learning objects, slide presentations etc. It supports localized searching of metadata elements in two modes – simple and advance. Users can limit search in a single repository or a group of repositories (by default a given search session includes all the available repositories). Search can be filtered by DC metadata elements like title, author, date range, language etc. The latest version of PKP can harvest metadata from DCMES (Simple DC metadata), MARC 21 Bibliographic Format and ETD-MS (metadata for electronic theses and dissertations).

Harvesting is a modest beginning of a new era of localized search services that harvest metadata from different OAI/PMH compatible open access resources such as, open access journals, open access repositories and open access ETDs etc. Even we can use one single search interface for different services of an institution. In next section, we will explore the integration between these two different services in a single local search interface. Many major OA services like BASE search engine, OAIster etc. are developed on the basis of harvesting technology.

3.8 INTEGRATION OF OPEN ACCESS WITH REPOSITORIES

This section is divided into two sub-sections. First one is intended to describe, integration with existing search services. Second sub section will be concentrated on integration with research administrative systems.

3.8.1 Integration with Existing Search Services

An institution can facilitate different kinds of search services- e-journals search, external database search, library OPAC search, IR search etc. which are most comprehensive and common facilities provided by institutions. All these services in a typical library setup are available through different user interfaces. It is less than an ideal situation as far retrieval efficiencies are concerned. To avoid this situation, institution may facilitate a single window search interface by using discovery tools or by using custom search engines. The way by which an institute can provide single window search interface may be referred to as integration.

Almost each and every institution provides more than two services to their end-users. At the same time, they may provide information relating to end-users' search query from their internal database (includes Library OPAC, Institutional repository, e-journal database) and from external databases (e.g. open access database like BASE or subscribed database like Scopus). It becomes much awkward and monotonous to jump from one search interface to another to retrieve any information regarding any search query. In order to consolidate this search, a mechanism must be there to search the entire external and internal databases through a single window search interface. It is quite simple to do this by using tools like Custom Search Engine, Discovery tools etc. Before creed to mechanisms we should be little bit acquainted with these two terms.

Custom Search Engine (CSE)

Custom search services are facilities to use established search engines to organize and index selected Web entities though API (Application Program Interface). This may be applied to index OA journals and OA repositories. Each such custom search service has URL and search only resources included in it. Finally, the custom search services may be integrated with local repository or local user interface to provide a single-window facility to search local OA repository as well as global OA repositories. The major search services like Google, Yahoo etc. provide custom search mechanisms to empower the users for search. These facilities allow searchers to index a set of selected websites. Each custom search interface provides URL (Uniform Resource Locator) to ensure global access.

The process of integration is straight forward and includes steps mentioned below:

- ***Development of CSE on a given area:*** You may try developing a CSE by utilizing Google API[76] for open access LIS journals.
- ***Widget generation:*** Widget is an application, or a component of an interface, that enables a user to perform a function or access a service. Google CSE allows users to generate Widget in the form HTML and Java code. You may generate Widget for your CSE by simply clicking the appropriate button.
- ***Widget integration:*** The next step is to integrate Widget for your CSE in existing library systems (like Web-OAPC) or OAR search interface.
- ***Testing and Debugging:*** Finally, after testing and debugging (if necessary), the product is ready for end users.

Discovery Tools

Discovery tools, powered by federated search mechanisms, allow users to perform concurrent searching in the library catalogue (metadata level), journal articles (full-text level), electronic theses and dissertations, consortia databases, public web, open access repositories, union catalogues etc. through a single-search interface with a set of feature-rich tools to support users. In simple words, a web-scale discovery services allow users to search local and remote databases, open access and commercial knowledge from a pre-harvested single central index. The unified interface is a big boost for users as they no longer need to choose a specific search tool to begin their search. These tools are available commercially (e.g. EBSCO Discovery Service) and also as open source products (such as VuFind, SOPAC, Blacklight, OpenBib etc).

3.8.2 Integration with Research Administrative Systems

Researchers register their work by giving metadata like title of a research work, name of researcher etc. under an institutional system. It may be under university system or under any funder agency. So institutions should maintain all records for official as well as academic purposes. In the Research Administrative System, at the time of submission of research thesis, it will automatically be uploaded in institutional repository also. In this context we may have a short look on CRIS/OAR Interoperability Project. Current Research Information and Open Access Repositories (CRIS/OAR) transfers metadata of publications automatically from research information system to institutional repository with option (from authors) to integrate full-text resources. It aims to achieve grand unification of research administration needs and OA repositories.

[76] http://www.google.com/cse/

CRIS/OAR Interoperability Project[77]

This project started in January 2009 and outcomes were presented in October 2010. The latest release of open source repository management software Dspace supports CRIS/OAR. This allows Dspace repository population by automatically transferring metadata-only records from the CRIS to the OAR and asking authors to attach the appropriate full-text versions of the works to the records in the repository.

Suggested Activities

1. Use Google Custom Search (https://www.google.com/cse/) and develop a search engine to search open access contents across a specific discipline. Register open access journals (see http://www.doaj.org/) and open access repositories (see http://www.opendoar.org/index.html and http://roar.eprints.org/) as targets.

2. Use any open source harvesting software (e.g. PKP Harvester) to create a federated search interface for any five OAI-PMH compliant IRs in an area (discipline) of your choice.

3.9 LET US SUM UP

The open access movement supported by the development of a number of open source software has resulted in the development of a number of institutional repositories, following varieties of hardware and software solutions according to the objectives of the repositories. These repositories may also be different in their coverage, software usage, nature of contents and most importantly in retrieval techniques and tools. As a result, it is difficult for end users search comprehensively these repositories that provide scholarly materials freely. This situation necessitates the development of a single window search service covering all the repositories in a given domain of knowledge. These single windows search services (based on resource metadata) are advantageous to scholars and others as it brings them closer to uniform access interface for scholarly information bearing objects and cultural resources. To overcome the problems, technological solutions in the form of harvesting have been developed.

Harvesting refers to a technique of extracting metadata from individual repositories and collecting it in a central catalogue. Metadata harvesting refers specifically to the gathering together metadata from a number of distributed repositories (e.g. eprint archives) into a combined data store. This unit explains the concept of harvesting and harvesters.

[77] https://infoshare.dtv.dk/twiki/bin/view/KeCrisOar/WebHome

Resource Optimization

The provisions of exchanging data without minimal loss of content functionality of multiple systems (with different hardware & software platform and data structure interface) are achieved through the technology of interoperability. Through Croswalk, it is possible to use metadata created by one community by another group that may employ different metadata standards. It is useful for virtual collections where resources are drawn from varieties of sources and expected to act as a whole. Interoperability and crosswalk ensures exchange of bibliographic data among heterogeneous systems across the globe. Various tools and standards to achieve interoperability have been discussed in this unit. The OAI/PMH standards are extensively used in the domain of library and information services. It is necessary that you should be aware of how harvesters work. This aspect has been dealt with under the heading Harvester mechanism. The framework for Harvester designing has been discussed in detail in the section 3.7.

In any open access environment two aspects – integration of open access repository with existing search services and integration with research administrative systems are important. The way by which an institute can provide single window search interface may be referred to as integration. Almost each and every institution provides more than two services to their end-users. At the same time, they may provide information relating to end-users' search query from their internal database (includes Library OPAC, institutional repository, e-journal database) and from external databases (e.g. open access database like BASE or subscribed database like Scopus). To consolidate the search, a mechanism must be there to search the entire external and internal databases through a single window search interface. This can be done by using tools like Custom Search Engine, Discovery tools etc.

Researchers register their work by giving metadata like title of a research work, name of researcher etc. under an institutional system. The institutions (university or funding agency) should maintain all records for official as well as academic purposes. In the Integration with Research Administrative System, at the time of submission of research thesis, it will automatically be uploaded in institutional repository also. Researchers register their work by giving metadata like title of a research work, name of researcher etc. under an institutional system. It may be under university system or under any funder agency. So institutions should maintain all records for official as well as academic purposes. In the Research Administrative System, at the time of submission of research thesis, it will automatically be uploaded in institutional repository also. The unit concludes with a description of CRIS/OAR Interoperability Project.

KEYWORDS

Author Addenda: A contract between author and publisher to retain rights of an article for his/her creation.

Content Preservation: Important to support continuous OA services. Retention period, Functional preservation, File Preservation and Fixity and Authenticity are considered as main important factor to preserve content.

Direct Deposit: Integration of direct deposit service, which transfers articles directly from the publisher into the institutional repository, may be very useful for OA content management (such as integration DSpace with OJS via Sword protocol).

Distributed Repositories: Repositories which are a collection of resources that can be accessed to retrieve information are available at different locations.

Gold Open Access: Publishing in an open access journal

Green Open Access: Self archiving of articles in an open access repositories

Information Mashup: Term "Mashup" is retrieved from the idea of consolidating data from two or more sources and presenting it with a new look. It is a web application.

Ingest: Submission of metadata and objects into OA system is technically called Ingest.

Interoperability: Compatibility of two or more computer systems so that they can exchange data and information and can use the exchanged data and information without any kind of manipulation or loss.

Mandate: An official order or commission to do something (as per Oxford Dictionaries).

Metadata: Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve, use or manage information resources. It is the key to ensuring that resources will survive and continue to be accessible in future. Some of the examples are Dublin Core, TEI, METS etc.

OA Mandate: Open Access mandates is a condition/provision that has been taken by various institutions organizations and funder agencies to make sure the free accessibility for reusing, remixing, redistribution of scholarly communication.

OAI Verbs/Requests: The six OAI verbs are: Identify, ListMetadataFormats, ListSets, GetRecord, ListIdentifier, ListRecords

OAI/PMH: Open Archive Initiatives/ Protocol for Metadata Harvesting is a framework for aggregating metadata from multiple data providers.

OAIS: Open Archival Information System is developed by the Consultative Committee for Space Data Systems and adopted by ISO as International Standard ISO 14721:2002. This reference model is used as a framework for the development of preservation archives for digital materials.

OSI: Open Society Institute.

Post-print: Post-print is a peer-reviewed journal article which has been published or in process to be published. But an author copy is there with revisions having been made.

Pre-print: Pre-Print is state of a paper that is, just exact before publishing.

REFERENCES AND FURTHER READING

Arts and Humanities Data Service (2004). *AHDS Guides to Good Practice*. London: AHDS. Retrieved from: http://ahds.ac.uk/creating/guides/index.htm.

Carr, L., White, W., Miles, S. and Mortimer, B. (2008). Institutional Repository Checklist for Serving Institutional Management, Version 0.2. OR2008, Third International Conference on Open Repositories, Southampton 1-4 April 2008. Southampton: University of Southampton. Retrieved from: http://pubs.or08.ecs.soton.ac.uk/138/1/IRChecklist.pdf.

Chudnov, Daniel (1999). Open source library systems: Getting started. Retrieved June 22, 2004, from www.oss4lib.org/readings/oss4lib-gettingstarted.php

Cole, T.W. (2003). Using OAI: Innovations in the Sharing of Information. *Library Hi Tech*, 21(2): 115-117.

Crow, R. (2002). The Case for Institutional Repositories: A SPARC Position Paper. (Washington, DC: Scholarly Publishing & Academic Resources Coalition). Retrieved January 12, 2009, from http://www.arl.org/sparc/IR/IR_Final_Release_102.pdf.

DANS (2008). *Data Seal of Approval*. The Hague: Data Archiving and Networked Services. Retrieved from: http://www.datasealofapproval.org/

Das, Anup Kumar (2008). *Open Access to Knowledge and Information: Scholarly Literature and Digital Library Initiatives - the South Asian Scenario*. New Delhi: UNESCO.

Das, Anup Kumar (2009). Open Access to Research Literature in India: Contemporary Scenario. *ISSI Newsletter*, 5(1), 9-14.

Das, Anup Kumar; Dutta, C. & Sen, BK (2007). *ETD Policies, Strategies and Initiatives in India: A Critical Appraisal*. Presented in ETD 2007 Symposium, Uppsala, Sweden. http://eprints.rclis.org/9944/

Digital Curation Center (2008). The DCC Curation Life Cycle Model. Edinburgh: DCC. Retrieved from: http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf.

Digital Preservation Coalition (2008). Preservation Management of Digital Materials: The Handbook (section 5.5). York: DPC. Retrieved from: http://www.dpconline.org/graphics/medfor/recommendations.html

Dulong de Rosnay, M. (2008). Check Your Data Freedom: Taxonomy to Assess Life Science Database Openness. London: Nature Proceedings. Retrieved from: http://dx.doi.org/10.1038/npre.2008.2083.

Gargouri, Y., et al. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLOS One*, 5 (10). E 13636. http://www.plosone.org/article/info:doi/10.1371/journal.pone.0013636

Ghosh, S.B. and Das, Anup Kumar (2007). Open Access and Institutional Repositories – A Developing Country Perspective: A Case Study of India. *IFLA Journal*, 33 (3), 229-250.

Harnad, S. (2005). Impact Analysis in the Open Access Era. http://openaccess.eprints.org/index.php?/archives/2005/10/10.html

Knowledge Exchange (2009). Open Access-What Are the Economic Benefits? A Comparison of the United Kingdom, Netherlands and Denmark. http://www.knowledge-exchange.info/Default.aspx?ID=316.

Lagoze, C. and and Sompel, H.V. (2003). The Making of the Open Archives Initiative Protocol for Metadata Harvesting. *Library Hi Tech*, 21(2): 118-128.

Macdonald, S. (2008). Data Visualisation Tools: Part 1 - Numeric Data in a Web 2.0 Environment. DISC-UK, 2008. Retrieved from http://www.disc-uk.org/docs/Numeric_data_mashup.pdf

Massachusetts Institute of Technology Libraries (2009). Data Management and Publishing. Cambridge, MA: MIT Libraries. Retrieved from http://libraries.mit.edu/guides/subjects/data-management/

Millington, Peter. (2006). Moving forward with DOAR Directory: 8th International Conference on Current Research in Information Systems, Berge, 11-13 May 2006.

Mukhopadhyay, P. (2006). Five Laws and Ten Commandments: The Open Road of Library Automation in India. In: *Proceedings of the National Seminar on Open Source Movement – Asian Perspective*. Kolkata: IASLIC. Retrieved from https://drtc.isibang.ac.in/handle/1849/409

NISO (2004). Understanding Metadata. Bethesda, USA, NISO Press. ISBN: 1880124629. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Online Computer Library Center (2007). Trustworthy Repositories Audit & Certification: Criteria and Checklist. Version 1.0. Dublin, OH: OCLC. Retrieved from http://www.crl.edu/PDF/trac.pdf.

Open Knowledge Foundation (2009). Comprehensive Knowledge Archive Network. Cambridge: Open Knowledge Foundation. Retrieved from: http://ckan.net/).

Purdue University Libraries (2007). D2C2, Distributed Data Curation Center. Purdue University Libraries, West Lafayette, IN, USA. Retrieved from: http://d2c2.lib.purdue.edu/).

Repositories Support Project (2008). *Repository Policy Framework*, Briefing Paper. Bristol: JISC. Retrieved from http://www.rsp.ac.uk/pubs/briefingpapers-docs/repoadmin-policyv2.pdf.

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).

Sale, A (2006). The Acquisition of Open Access Research Articles. *First Monday*, 11(9). Retrieved from http://firstmonday.org/issues/issue11_10/sale/index.html

Sarkar, P. & Mukhopadhyay, P. (2010). Designing single-window search service for electronic theses and dissertations through metadata harvesting. *Annals of library and information studies*, 57(4), 356–364.

Sherpa, (2012). OpenDOAR Policies Tool. Nottingham: University of Nottingham. Retrieved from http://www.opendoar.org/tools/en/policies.php.

Suber, Peter. (2012). *Open Access*. Cambridge: The MIT Press.

Sutradhar, B (2013). Design and development of interoperable institutional digital repositories among Indian Institutes of Technology in India: a prototype. Conference Paper, International Conference on digital Libraries, New Delhi, 581-599.

Swan, A. (2012). *Policy Guidelines for the Development and Promotion of Open Access*. Paris: UNESCO.

UK Data Archive (2011). *Managing and Sharing Data: Best Practice for Researchers*. Essex: University of Essex. Retrieved from http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Witt, M. (2008). Institutional Repositories and Research Data Curation in a Distributed Environment. *Library Trends*, 57(2). Retrieved from http://muse.jhu.edu/journals/library_trends/v057/57.2.witt.html

This module has been jointly prepared by UNESCO and The Commonwealth Educational Media Centre for Asia (CEMCA), New Delhi.