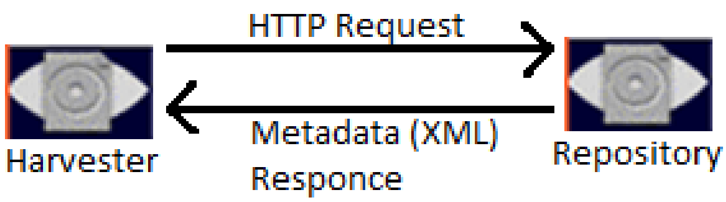
Large-scale Metadata Harvesting – Tools, Techniques and Challenges : A Case Study of National Digital Library (NDL)

Samrat Guha Roy[1], Dr. B. Sutradhar[2], Dr. Partha Pratim Das[3]

Abstract: OAI-PMH enabled open source digital library software like DSpace, EPrints, VuFind, Drupal OAI harvester, PKP harvester had made it possible to harvest massive metadata from different IDR's. IT brought new hope and new opportunities for providing various new services to our library users. This article attempts to explore the tools, techniques and the significant challenges for large-scale metadata harvesting and metadata curation. A recent bibliographic study of Scopus had shown that there is a rapid increase of article publication over the last two decades. "A total of 25,482 publications represent the literary output in different formats, in different subjects, and from various nations" (ul Ajaz Wani & Gul, 2008). All these preprint academic research documents like conference papers, journal article, annual reports, protocols, lecture notes may be already uploaded or needed be upload in various institutional digital repositories (IDR) for long-term digital preservation and reuse. In this study we have harvested the metadata from different such IDRs into a centrally indexed repository for providing a single window search box. Therefore, with this we may dream that day is not far away when we will not need any e-resource subscriptions, as those will be available in our IDR. It will be indeed a great achievement and will be very much helpful to the academic community. However, along with this, a continuous metadata curation is a major intermediate phase, which focuses on the proper mapping of data to metadata. Programmatic curation and manual curations are the two processes done for final curation of the harvested metadata, where both are having their own merits and demerits. This article further focuses on the process workflow of metadata data curation, and the possible challenges need to manage by the librarian for proper indexing of the items.

Key words: OAI-PMH, ORE, Metadata Curation, LRMI, Search Box, Digital Repository, Metadata Harvesting

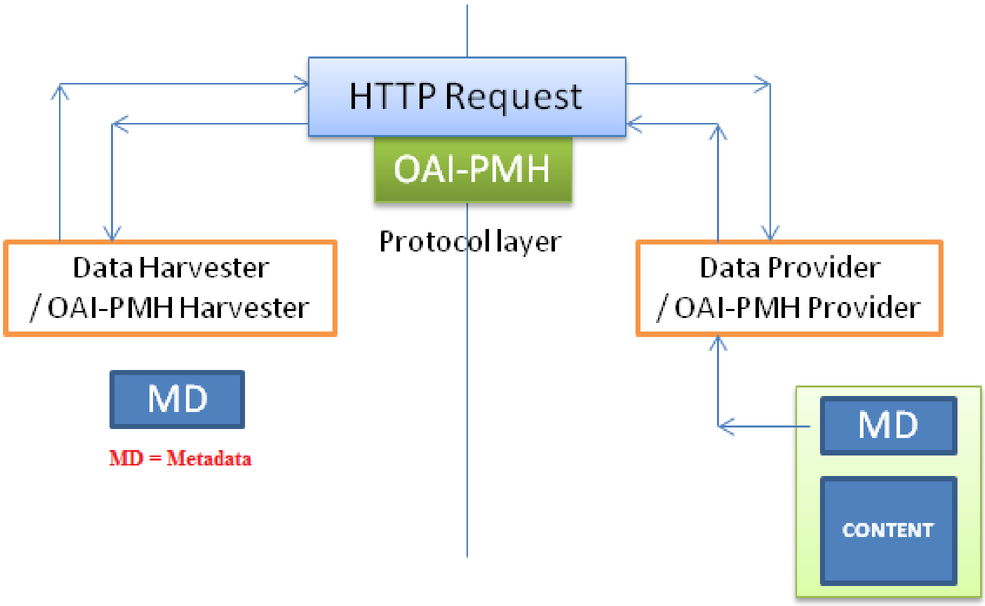
1. **Introduction to Metadata Harvesting:** Metadata defined as "data about data" which provides information about other data. National Information Standards Organization (NISO, 2004) describes three types of metadata (1) Structural metadata, (2) Descriptive metadata and (3) Administrative metadata. Structural metadata is data about the various physical or logical structure of the uploaded item like the controlled vocabulary, thesauri, page layout, file physical format etc. Descriptive metadata is the described information like the title, author, publisher that are always used to locate or discover the item. Administrative metadata help for administering the



information like how it is created, when created, authorizations, etc. Metadata harvesting (Breeding, 2002) is the process where the “data harvester” collects metadata from “data provider”. The “data providers” are the repository that create and exposes the structured metadata as well as the content to the data harvester. The “data harvester” indexes the harvested metadata into a central data indexer. In the harvesting phenomenon, data harvester gives the HTTP request to fetch information from the data provider using OAI-PMH protocol.

1.1 OAI-PMH: “*The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a protocol developed for harvesting (or collecting) metadata descriptions of records in an archive so that services can be built using metadata from many archives. An implementation of OAI-PMH must support representing metadata in Dublin Core, but may also support additional metadata representations”*

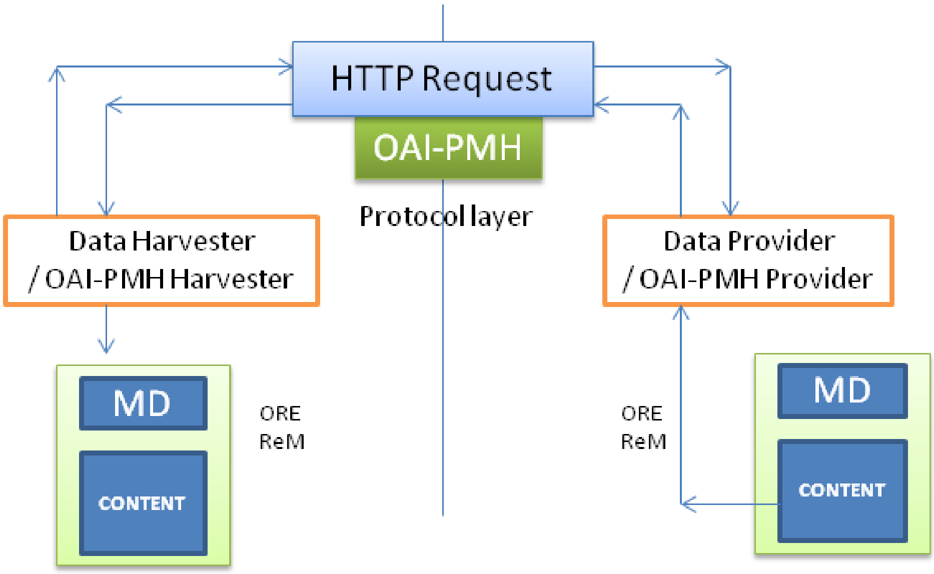
[Source: https://en.wikipedia.org/wiki/Protocol_for_Metadata_Harvesting]



The first version of OAI-PMH introduced in the year January 2001 at a workshop in Washington D.C., and further modifications to the XML standard proposed by W3C. Presently the version, 2.0, released in June 2002 which is based on client–server architecture, in which "harvesters" are the client or “data Harvester”, which sends request information, and the "repositories" are the servers or “data providers” which in turn sends the metadata to the harvester. Data providers send XML metadata in Dublin Core format or other XML format.

1.2 OAI-ORE: “*OAI-ORE defines standards for the description and exchange of aggregations of Web resources. The OAI-ORE specification implements the ORE Model, which introduces the Resource Map (ReM) that makes it possible to associate an identity* *with aggregations of resources and make assertions about their structure and semantics".*

[Source: https://en.wikipedia.org/wiki/Open_Archives_Initiative_Object_Reuse_and_Exchange]



The major objective of OAI-ORE is to provide the content along with the proper metadata schema this enables us to reuse the object and further preservation.

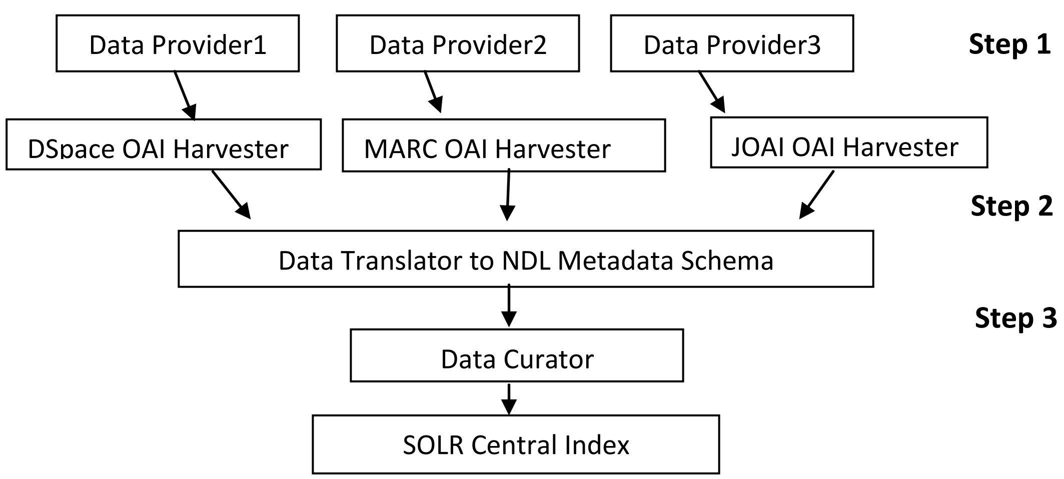
2. **Literature Review:** As according to the studies made by (Arlitsch & O'Brien, 2012) said that "Google Scholar has difficulty indexing the contents of institutional repositories" as most repositories uses Dublin core metadata information which cannot be used for bibliographic citation. The various search engines are also unable to search content from the digital repositories as the web-bots, web-crawlers are not enabled in the IDR by default. Therefore, the major part of academic related data remains hidden from the search engines. In Indian context, there are many digital repositories, which are not available in public domain, they all are available in the institute LAN, and again they remain invisible to the search engines. The search engine does not read contents from images, scripts, applets, video/audio formats or software so a huge content remains invisible. Therefore, it has suggested in this paper that all these images, software's, multimedia contents should have proper metadata along with before uploaded into any repository. A central indexer will harvest metadata (Zeng, Lee, & Hayes, 2009) from various IDRS i.e. Digital repository and this eventually becomes a single window search facility for the end users. During the study, we found that the digital repositories must be visible to web crawlers otherwise; they should be OAI-PMH compliant. Then also another problem still exits to the search engine as it does not classify the crawled information and does not displays the crawled information's academic usage like the content will be useful to a research scholar or it is useful to a k12 student. Henceforth

digital achieving with proper metadata will provide better result that is more accurate to the end users.

“Google” do not have much detailed coverage of Dublin core metadata or learning resource metadata schemas. (Yang, 2016) made a study on “Search Engines Notice” where he defined that many search engines does not notices the metadata and the content of the DSpace repositories. Zhang Xiaolin (2009) made a good study about Chinese Digital Library Project where it is cited that for building an effective and efficient digital library “structured metadata schema” needs to be defined. A proper structure will resolve the accessibility, interoperability, and sustainability issues of digital library.

In this paper, we have proposed that when archiving digitised documents that are having enriched metadata information’s libraries should embrace and harness collaborative and crowed sourcing metadata approaches as rightly said by (Deng & Reese, 2009) in their studies of "Customized mapping and metadata transfer from DSpace to OCLC to improve ETD work flow". It has said that the metadata interoperability does not depend in developing a set of standards on top of existing ones rather we should extend the existing metadata schemas. It is along these lines that a conceptual metadata framework aimed at contributing towards the semantic interoperability of disparate digital libraries as suggested in this paper.

3. **Research Methodology:** The present study is carried out in four steps. First step is to find out various OAI-PMH (Zavalina, 2014) complaint software (tools) that may be used as data harvester, second by searching digital repositories available in India from DOAR/ROAR as well as by visiting various institutional web sites, which will be our data providers. Therefore, harvester harvest and indexes the metadata into central indexer and third step is to curate the metadata. The metadata curation task is accomplished in two ways one manual curation and secondly is the programmatic curation. The main objective of this study is to find out the various tools techniques and challenges faced while harvesting large-scale metadata contents from various digital repositories (Indian context) and to integrate into one indexer. Finally, the ingestion is performed.



3.1 The various OAI-PMH complaint software (Tools): There are many open source software's available today, which are OAI complaint, and they may be used for metadata harvesting.

Sl. No	Software Name	OAI Data Harvester	OAI Data Provider	Output Data Format	Challenges Found
1	DSpace	Y	Y	AIP / XML	More than 8000 data harvest is an issue
2	EPrints	Y	Y	XML	Multiple Record Harvest Occurs
3	Greenstone	Y	Y	XML	Lower versions don't support OAI-PMH
4	MARC Edit	Y	N	Marc / XML	Marc Tag mapping to DC
5	PKP Harvester	Y	Y	XML	XML Parsing
6	Drupal - OAI	Y	Y	XML	XML Parsing
7	VuFind - OAI	Y	Y	XML	XML Parsing
8	JOAI	Y	Y	Xml	XML Parsing

According to the process in this study that the data providers are first harvested using the above mentioned "data harvester". Then the data is translated to the new NDL metadata schema. These are further curated using programs or at sometimes curated manually. Finally, the data is ingested into SORL index that are queried by the user in single user search box.

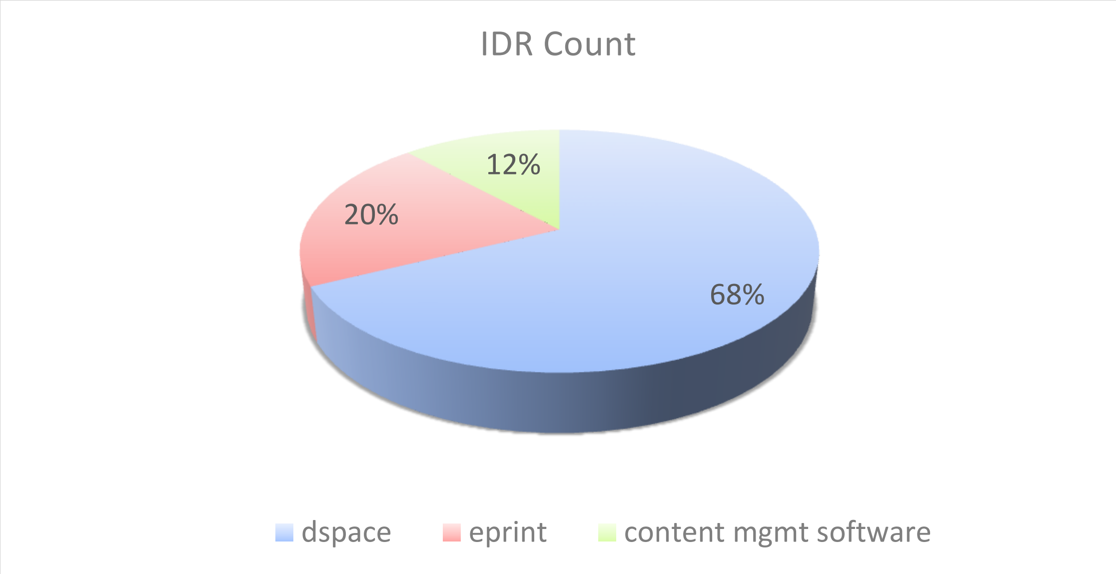
3.2 Institutional Digital Repositories in India (Techniques): We have collected at first all the Indian repositories listed in DOAR then checked for active repositories that are harvestable. We have also searched the various institute's/universities web site and collected the repository url. The repository where we have found the OAI is not properly indexed there we have contacted the concerned administrator and guided them for proper SOLR and OAI index. So at the end of 8^{th} month we have harvested around 68 digital repositories and the total content volume is 5,45,8856 (five lakhs and forty-five thousand plus). Table below lists all the harvested IDRs and harvested content volumes. (Houssos et al., 2011) rightly cited in TDPL 2011 conference about the Europeana that provides digital content access services across Europe's cultural organisations (that is, libraries, museums, archives and audio/visual archives).

For the study purpose total 5,45,886 metadata that is being harvested from various digital repositories of India; all these data have gone through the translator and curation stages. During translation LRMI mapping is done using various types of java codes. As well as when we have found that DSpace did not harvested more than 8000 records we have taken the csv format of the data in bulk format and created the AIP package using SAF (Simple Archive Format) tool used for bulk data import into DSpace [source: https://github.com/lib-uoguelph-ca/dspace-csv-archive].

SL. NO	Name of Source	Collections
1	Indian Academy of Sciences	88596
2	KrishiKosh - Indian National Agricultural Research System	49654
3	IISc - Institutional Repository	40139
4	Inflibnet - Shodhganga	36313
5	West Bengal Public Library Network	30972
6	Jadavpur University	30437
7	Osmania University Digital Library	24471
8	IIT Bombay	16744
9	Gokhale Institute of Politics and Economics	16648
10	ICRISAT - Institutional Repository	13427
11	IIT Roorkee - Thesis	13191
12	Manipal University	12813
13	IIM Ahmedabad	10954
14	CSIR - Indian Institute of Chemical Technology	10368
15	Central Marine Fisheries Research Institute	10122
16	University of Mysore	10109
17	Aligarh Muslim University	8835
18	ICRISAT - Open Access Repository	8184
19	CUSAT - Institutional Repository	8108
20	Bharathidasan University	7837
21	Indian Institute of Astrophysics	6520
22	CSIR - National Metallurgical Laboratory	6052
23	CSIR - National Aerospace Laboratories	5786
24	Directory of Open Access Journals	5504
25	IIT Delhi	5256
26	ISI Kolkata	5167
27	CSIR - National Institute of Oceanography	4679
28	Raman Research Institute	4609
29	CUSAT - Thesis	4059
30	NIT Rourkela - Thesis	3230
31	NCERT	3166
32	IUCAA- Pune	3067
33	CSIR - Central Electrochemical Research Institute	2551
34	CSIR - Central Glass and Ceramic Research Institute	2517
35	IISc - Thesis	2372
36	NIT Rourkela - Institutional Repository	2288
37	MoES - Indian National Centre for Ocean Information Services	2232
38	SreeChitraTirunal Institute for Medical Sciences & Technology	2062
39	VidyaPrasarak Mandal	1939
40	Inflibnet - Shodhgangotri	1930
41	BirbalSahni Institute of Paleobotany	1777
42	IIT Kharagpur	1705
43	CSIR - National Physical Laboratory	1563
44	Inflibnet - Inflibnet's Institutional Repository	1470
45	Society For Natural Language Technology Research	1172
46	IIT Gandhinagar	1157
47	University of Kashmir	1009
48	IIT Hyderabad	1000
49	Tamil Nadu Agricultural University	996
50	CSIR - Open Access Repository of Indian Thesis	984
51	Swami Vivekananda Yoga AnusandhanaSamsthana	949
52	S.N. Bose National Centre for Basic Sciences	946
53	IISER Bhopal	944
54	Aryabhatta Research Institute of Observational Sciences	806

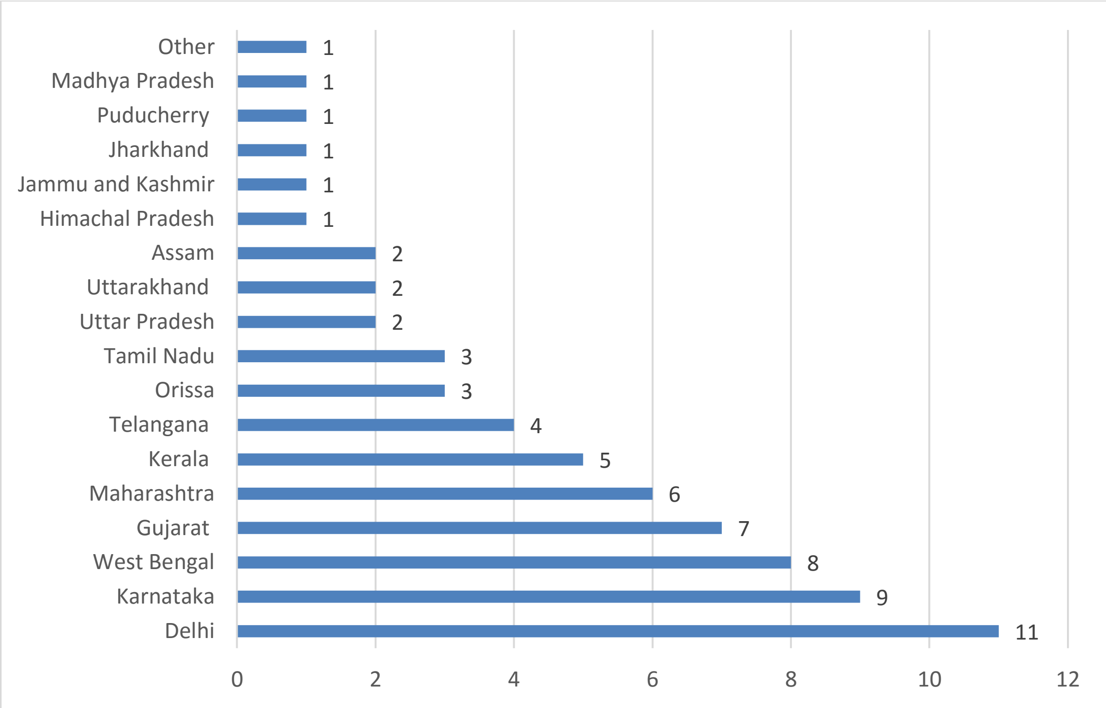
SL. NO	Name of Source	Collections
	(ARIES)	
55	Madras Diabetes Research Foundation	800
56	Pondicherry University	751
57	ICAR - Indian Institute of Spices Research	725
58	IIT Bhubaneswar	609
59	Maharaja Sayajirao University of Baroda	543
60	IACS Kolkata	537
61	IIT Guwahati	513
62	Indian Institute of Geomagnatism	496
63	Chitkara University	339
64	Indira Gandhi Institute of Development Research (IGIDR)	334
65	Indraprastha Institute of Information Technology Delhi	264
66	National Institute of Immunology	228
67	PanditDeendayal Petroleum University	192
68	DRDO - Institutional Repository	169
		545886

The pie chart below shows that the majority of the digital repository software present in India are is DSpace 68% followed by EPrints 20% and other CMS 12%. This



is also helpful for the implementation of LRMI metadata schema in NDL as most DSpace are having the same software platform with different versions.

We have also analyzed to find the state that has a more number of IDRs present in public domain. The result showed Delhi has the height number of IDR 11 followed by Karnataka as 9 and west Bengal 8. This data is highly useful for us to focus for the regions that need more attentions for the development of IDR like Madhya Pradesh, Puducherry etc.



3.3 Metadata Translation and Curation: The various metadata (Zeng et al., 2009) elements of LRMI schema that is being used for this study are lrmi.educationalUse, lrmi.timeRequired, lrmi.typicalAgeRange, lrmi.interactivityType, lrmi.learningResourceType, lrmi.useRightsUrl, lrmi.isBasedOnUrl,lrmi.educationalRole, lrmi.educationalAlignment.educationalFramework, lrmi.educationalAlignment.educational.pedagogicObjective, lrmi.educationalAlignment.educational.difficultyLevel

During the metadata harvesting process the Dublin core elements gets harvested after that this data information are programmatically translated to Dublin core metadata as well as the various LRMI metadata elements are populated. This is done for more proper search result for the end users.

4 Data Analysis

4.1 Harvesting Challenges: There are several difficulties in harvesting metadata. A metadata harvester needs to be administered from time to time for proper data harvesting as well as it is needed to Stop, Start or Restart/Refresh at regular interval. While large collections download many times data becomes corrupted and does not parse well. This is the very crucial part of the large scale metadata harvesting as quality (Park & Tosaka, 2010) is affected if any part of the data is corrupt. It is also found that invalid data crashes or stops the harvesting parser too. It is important, therefore, to have access to the raw data in cases of poor metadata harvesting.

Some of the major harvesting challenges found during the study are mentioned below:

- Untitled Metadata: It is found while harvesting from many EPrints repository (data provider) using DSpace as data harvester we get “Untitled” string in “Title” metadata. For example, University of Mysore, CSRI – NPL etc. solution to this type of case is to re-index the data provide and then harvest again. If then also the problem is not resolved, then it is better to use another tool for data harvesting.
- Junked Unicode Character: It is found while harvesting using marc edit most of the Unicode Latin words comes as junked character codes. This type of error is resolved during metadata programmatic curation.
- Incomplete Harvest: DSpace harvester stops after harvesting 8000 records. So when the collection size is more than 20,000, it becomes a challenge to harvest all. Solution is to stop all other harvesting threads and refresh the particular collection harvesting point it will harvest gradually.
- Connection Time out: Is occurs when the data provide server is not active on internet.
- Multiple Record Harvest: EPrints repositories preserves the items based upon subject classification keywords. It also provides the data based upon the subject keyword handle id. Hence while harvesting subject wise may lead to harvest the same item multiple time.
- OAI Index Error: “No Record Found” error is displayed in DSpace while giving the “ListSets” command. Solution to this kind of error is to re-index the SOLR and OAI indexes. [command:]

4.2 Metadata Curation Challenges: Curating large-scale harvested metadata is always a challenging task. The various crosswalk (Khoo et al., 2015) programs are used for Dublin core metadata curation along with the LRMI metadata. It may be done in two ways 1) programmatic curation 2) manual curation. Programmatic curation needs more logic and advanced dictionary mapping whereas manual curation is done manually by “subject matter experts” SME which are time consuming but more accurate. Bulk data modifications are done using programmatic process but codes needed to be written more precisely as a wrong logic will make a huge modification within the data.

- Many Author Names are written as Dr. S. K. Ghosh and it should be mentioned as Ghosh, S. K. [lastname, firstname]
- Html code needs to be replaced like ‘&’ should be ‘&’
- Latin Unicode character should have proper encoding

5 Conclusion: In our case study, we found that the large-scale metadata harvesting could be easily accomplished by using various OAI complaint software tools like DSpace, EPrints, PKP Harvester etc. However, after that translating / curating the metadata into Dublin core metadata schema and LRMI definitely needs more precision and accuracy.

However, after metadata curation process every harvested item will have proper descriptions about learning resource type, which eventually helps the end user to get results that are more relevant. Based on this metadata, we were able to do beyond basic searches and browsing. Initial user experience testing of the learning resource metadata and recommendation ranking gave us promising results. In this paper, we created a prototype of a harvested metadata model that is part of a work in progress project for harvesting and exposing educational objects to the end users. The proposed model was able to accommodate different metadata schemas harvested from these repositories, and annotations implemented successfully.

6 References

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech*, *30*(1), 60–81. https://doi.org/10.1108/07378831211213210

Breeding, M. (2002). Understanding the protocol for metadata harvesting of the open archives initiative. *Computers in Libraries*, *22*(8), 24–29. https://doi.org/10.1108/07378830310479776

Deng, S., & Reese, T. (2009). Customized mapping and metadata transfer from DSpace to OCLC to improve ETD work flow. *New Library World*, *110*(5/6), 249–254. https://doi.org/10.1108/03074800910954271

Houssos, N., Stamatis, K., Banos, V., Kapidakis, S., Garoufallou, E., & Koulouris, A. (2011). Implementing enhanced OAI-PMH requirements for Europeana. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6966 LNCS, pp. 396–407). https://doi.org/10.1007/978-3-642-24469-8_40

Khoo, M. J., Ahn, J.-W., Binding, C., Jones, H. J., Lin, X., Massam, D., & Tudhope, D. (2015). Augmenting Dublin Core digital library metadata with Dewey Decimal Classification. *Journal of Documentation*, *71*(5), 976–998. https://doi.org/10.1108/JD-07-2014-0103

NISO. (2004). Understanding Metadata. *National Information Standards*, (MD:NISO Press), 20. https://doi.org/10.1017/S0003055403000534

Park, J.-R., & Tosaka, Y. (2010). Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly*, *48*(October), 696–715. https://doi.org/10.1080/01639374.2010.508711

ul Ajaz Wani, M., & Gul, S. (2008). Growth and Development of Scholarly Literature: An Analysis of SCOPUS. *Library Philosophy & Practice*, *10*(2), 1–8.

Yang, L. (2016). Making Search Engines Notice: An Exploratory Study on Discoverability of DSpace Metadata and PDF Files. *Journal of Web Librarianship*, *2909*(May), 1–14. https://doi.org/10.1080/19322909.2016.1172539

Zavalina, O. L. (2014). Complementarity in Subject Metadata in Large-Scale Digital Libraries:

A Comparative Analysis. *Cataloging & Classification Quarterly*, *52*(January 2015), 77–89. https://doi.org/10.1080/01639374.2013.848316

Zeng, M. L., Lee, J., & Hayes, A. F. (2009). Metadata decisions for digital libraries: a survey report. *Journal of Library Metadata*, *9*(3/4), 173–193. https://doi.org/10.1080/19386380903405074

7 About Authors:

7.1 Mr Samrat Guha Roy is currently working as Asst. Librarian at Central Library IIT Kharagpur. He is having a wide experience in data harvesting using various open source software. mailto: samrat@library.iitkgp.ernet.in

7.2 Dr. B. Sutradhar is Librarian at Central Library IIT Kharagpur and Co-PI NDL project India. He is the main driving force behind the project for the creation and implementation of digital repositories in India so NDL harvester may harvest metadata. mailto: bsutra@library.iitkgp.ernet.in

7.3 Dr. Partha Pratim Das is a Professor at the Department of Computer Science and Engineering, IIT Kharagpur. He is the Joint-PI of the NDL project and Head, RM School of Engg. Entrepreneurship, IIT Kharagpur. He has over 35 years' professional experience between Industry and Academia. His interests include Digital Geometry, Image Processing, Object-Oriented Systems, Software Engineering, and Embedded Systems.
mailto: ppd@cse.iitkgp.ernet.in