Chapter 1

1.1 Introduction

It is a well established fact that information resources are the foundation of effective research, teaching and learning. To enable and to continue supporting instructional, research and service missions, the libraries have to enhance and enrich access to well organized, professionally managed, coherent, and comprehensive collections of scholarly information in all depth and breadth needed to realize the goals of the organization. It also requires enhancing and improving upon the availability of timely relevant and personalized services to make effective use of these vast and complex information resources as a national asset. Further, it is the responsibility of libraries to ensure persistent access to the high quality digital materials that result from teaching, learning and research.

Late in 21st century, the consensus of contributors such as faculty, researchers, practitioners, decision makers, students, publishers and librarians in the context of Open Access Initiative have set Performance Indicators (PIs) for a free access to the scholarly communications and to build a critical mass of scholarly content in a sustainable manner. Of course, the management of this critical mass of scholarly content, efficiently and effectively, has been the responsibility of the librarians supported by publishers and aggregators, which ultimately strengthen, promote and give a visibility to the researched content of the country. While making the scholarly content accessible free of cost and free of other barriers, it has demonstrated that this scholarly content is ‘public good’, which has been built based on the evidence and proactive response from the scholars through their contributors, funding agencies, government, apex bodies of the research pool and professional association to accelerate the growth and development. There are many quantitative and qualitative metrics that promote the overall effect or impact of this as public good and scholarly content pool (1).
The present study concentrates on demonstrating national consolidation of scholarly content as well other data types such as Research Reports, Supplementary data, Clinical trials etc in Health Sciences as National Repository, derived by harvesting metadata from institutional repositories and other archives using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). The Institutional Repositories serve as the input point for collection development along with the regular collection of the library for an expanding collection. The national consolidation of scholarly literature depends on the decisions of institution’s workflow, apply at different stages, including the selection, acquisition, processing, weeding, retention, preservation, and relegation and discarding of library materials.

1.2 Definition of the concepts

1.2.1 Institutional Repository

An institutional repository is a digital archive of the intellectual product created by the faculty, research staff, and students of an institution and accessible to end users both within and outside of the institution, with few if any barriers to access. In other words, the content of an institutional repository should have the followings (2):

- Institutionally defined;
- Scholarly;
- Cumulative and perpetual; and
- Open and interoperable

1.2.2 Data Provider

Data providers maintain one or more repository that supports OAI-PMH as a means of exposing metadata. This is the starting point for the harvesting or aggregated system. It means that contents and its associated metadata which resides in the repository needs to be harvested. Before metadata to be harvested from the digital repository, there is a need to understand how users may able to establish digital repository.
1.2.3 National Digital Repository System (NDRS) or Service Provider

NDRS is basically an aggregation of digital repositories, where it relieve repository managers from the maintenance of direct services, end users can be directed straight to the required resource from an aggregator, offer an enhanced metadata to the end users, facilitate preservation through appropriate metadata provision and/or content package and maintains the control of over the content while releasing metadata. It also provides control and personalization of content access through RSS feed, provides the exposure for content providers, preservation and metadata enhancement capabilities to support the long term storage and access to the content.

NDRS offer added-value services of their own to enhance aggregated metadata, amalgamate collections as the basis for analysis such as text and data mining, brokering role to facilitate access, a single point of information for statistics about access and downloads of data, to multiple source of research and other materials to aid discovery and possible exploitation through the building of value added services on top (3).

1.2.4 Metadata

Metadata is defined as ‘data about data’. For digital objects, metadata is needed to support their functionality over the years, as the object itself does not give enough information. Metadata is of vital importance to undertake proper digital preservation. Carefully collected metadata enables the future rendering of the object, creates trustworthiness as it will prove the authenticity of the object, and helps the future user to identify the object over the years. To meet these requirements, several metadata sets have been developed and standardized. Dublin Core Metadata Element Sets (DCMES), Metadata Object Description Schema
(MODS), Metadata Encoding and Transmission Standard (METS) are some examples of metadata standards used in the creation of institutional repositories (4).

1.2.5 Interoperability

Interoperability is the ability of two or more systems to exchange information and to use the information that has been exchanged. There are different kinds of different kinds of interoperability discussed such as technical interoperability, semantic interoperability, Political/Human Interoperability, Inter-community interoperability, Legal Interoperability, international interoperability. However, this thesis is more concerned with technical interoperability and search interoperability. Technical interoperability is all the hardware and software components of networks and information system can physically communicate and transfer information successfully whereas search interoperability is as 'the ability to perform a search over diverse set of metadata records and obtain meaningful results' (5).

1.2.6 Open Access Initiative-Protocol for Metadata Harvesting (OAI-PMH)

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting that can be used by a variety of communities who are engaged in publishing content on the Web (6).

1.2.7 Depositor

Any author of a paper and Institution repositories are the depositors for NDRS. The terms author, depositor and participant are used interchangeably in this article.
1.2.8 Items

Discrete manifestations of intellectual creation as described for the term "Institutional Repository" above. Typically, an item is equivalent to a research paper, technical report, or similar object.

1.2.9 Participation

Allowing one's intellectual creation to be deposited and made available through a disciplinary or institutional repository. This is assumed to be a conscious choice for all depositors.

1.3 Health Information Landscape

Information in developing countries is, though available in abundant and authors or organization/institutions are willing to share, yet they are not accessible on the Internet due to lack of proper Institutional repositories in biomedical institutions, which has been called as information drought (there is enough cloud but no rain). Therefore, there are gross inequalities of health information in the developing and developed world.

The Internet is bringing together people and publishers in identifying the information needs and mobilizing all relevant published and unpublished information and data like epidemiological data, clinical trials, drug trails and case conferences, Atypical case reports...
(unpublished) and dissertation and thesis. There is abundant information addressed to health workers and the public (community), not covered on Internet. In biomedical field, authentication of information and creating context is most important. Therefore, there is a need for new alliances among subject experts, librarians, publishers and distributors of information for creating relevant packages of knowledge or adapting for use, wherein librarians help in filtering the huge flow of information and selecting topics of interest to the local users and present them in an appropriate form. Librarians are to help local publishers to identify the gap of information for the local requirement and help local authors and publishers to publish and make them more visible on the Internet. Of course, when addressing the health workers and community, it requires looking into barriers like language, access. Hence it is required to develop to and pro (community to publishers and publishers to community) information flow which requires to develop plans, more strategic plan in its design and development of National Digital Repository System for Health information in the country.

1.4 Institutional Repositories: Growth and Development

Institutional repository (IR) has emerged as a means of storing digital contents produced by research organizations, universities etc, which is, in turn, reforming the
process of scholarly communication, and also becoming a key component topic within library and open access communities. Hernad, a key advocate of open access movement, views Institutional Repositories (IRs) also as the open access archives which are cost effective and immediate route to providing maximum access to the results of public funded results. As per McCord, these are capable of maintaining large amount and diversity of scholarly material in enhancing teaching, learning and research at the host institution and beyond, also enhance an institution’s prestige or branding by showcasing research output. These are an essential infrastructure for the reform of the enterprise of scholarly communication and publishing.

The increasing availability and use of digital form has led the institutions and other research organizations to reclaim responsibility for the management and distribution of scholarship through the creation of institutional repositories. It is revealed from the literature published in the last 10 years, that Open Access (OA) means electronic scholarly articles that are available freely at the point of use. There were two ways in the beginning to distinguish OA delivery:

- **Gold route** - author or author institution can pay a fee to the publisher at publication time, the publisher thereafter making the material available 'free' at the point of access.
- **Green route** - author can self-archive at the time of submission of the publication including grey literature (usually internal non-peer-reviewed), a peer-reviewed conference proceedings paper or a monograph.

In the green route model, the focus is on setting up institutional repository and to make it interoperable so as to become a part of aggregation model.

### 1.4.1 Development of Institutional Repositories

A large number of IRs have been established worldwide, more than 2000 repositories have been registered with Registry of Open Access Repository (ROAR) as maintained by Open Archive Initiative’s web site. (http://roar.eprints.org/) as well as by Open Directory of Open Access Registry Open Directory (OPenDOAR). This is in contrast to the discipline or subject-based repository such as Arxiv of Physics
archive (http://www.arxiv.org) or the Cogprint archive (http://www.cogprint.org) whose depositing policies are determined by the community.

1.4.1.1 Glance at IRs in India

Institutional repositories from India as registered with ROAR, OpenDOAR and also through Internet are detailed below:

- Existing IRs in India- there are 62 IRs registered with ROAR, OpenDOAR and also through Internet, out of which 11 are not functioning. Majority of IRs (77.4%) are having record ranging up to 2000. Indian Institute of Science is one institute having more than 20,000 documents in its repository. The major concern after establishing IRs is the filling up IRs. There is need for concerted efforts in this area.

- In Health Sciences, there are only four IRs, Out of four, two are institutional and two are disciplined based repositories. OpenMed of NIC is the leading subject domain repository in the area of health sciences as per survey conducted for latest ranking of World’s open access repositories (10).

- DSpace is the leading software adopted by the libraries as it is supported by continuous training programmes organized by various professional bodies, institutions, maintaining the List Server Digital Library Research Group (DLRG) by Documentation Research Training Centre (DRTC), Bangalore.

- Promotion of IRs in India is by conducting training programme on Open Source Software such as DSpace, EPrints, and GreenStone Digital Library (GSDL) by various professional bodies as well organizing conferences by many national as well international organizations.

- National Knowledge Commission has recommended that all public funded research should be made open access. It has been observed that there is a continuous growth in new IRs being registered and also there is a surge in the number of records over the period.

1.4.2 Components of Institutional repositories

Institutional repositories are having the following major important components:
1.4.2.1 Content

Content is intellectual product like research papers, theses, research data, learning resources, or even the administrative records of an institution. New IR software’s handle a wide range of resource types, where different types of output will have different requirements of metadata or access, whether peer reviewed or yet to be reviewed, internally developed by organization or published elsewhere with permission to host on their IRs. In fact volume and quality of content, where they allow non peer-reviewed material (e.g. preprints, working papers, newspaper opinion pieces), automatically support popular formats like Hyper Text Metadata Language (HTML), Portable Documents Format (PDF), PostScript, or Tagged Image File Format (TIFF), although it may be able to consider well documented other formats on request. Research Communities along with the developer of IRs has to decide the kind of collections to be deposited in the IRs. Some of the types of contents found in the IRs are:

- EPrints – Preprints/Postprints
- Working Papers and Reports
- Conference papers & Proceedings
- Electronic Thesis and Dissertations
- Data Sets
- Supplementary Materials
- Online and overlay journals
- Books
1.4.2.2 Technology

Another impetus for the recent growth of IRs is the easy availability of open-source software applications for setting up repositories and support from professional organizations. The highly distributed nature of resources scattered across the World Wide Web has made difficult to locate the relevant information. The interoperability standard developed by Open Access Initiative (OAI) has made these institutional repositories to interoperate to form a seamless global repository. The Open Archive Initiative, the Protocol for Metadata Harvesting (OAI-PMH) is designed to harvest metadata and associated resources that are distributed across different OAI-compliant servers. It has built-in support for basic Dublin Core metadata, standard used in digital-resource cataloguing. (http://dublincore.org) Some of the major digital repository software is given below; however, this list is far from a complete catalogue (11).

The selection of software model depends entirely on the needs and resources of the institute. For example, in case, one is comfortable using an open-source software platform that is open for community development, or someone would prefer a more conventional software vendor-customer relationship, where one can pay for technical support, upgrades, and consulting as needed. One must be cautious so as to aware of hidden costs in all product models. It might be beneficial to speak with other university librarians who have built institutional repositories with the systems one is evaluating. It is noticed that consulting companies may build on top of open source products, building services and offering hosting along with the free, open source software (12).

1.4.2.3 Intellectual Property Rights
Predominantly it covers publishers' copyright policies, and the opinions of data and service providers; this work all being used to explore the possibility of how information about rights might best be communicated as metadata. Also like publishers' copyright policies on self-archiving. The transfer of copyright to journal publishers has become a focus point for a discussion of the respective requirements of authors and publishers. Traditionally authors grant publishers 'exclusive license' to publish, where publishers not permit the inclusion of a paper in an institutional repository. It requires special copyright permission for inclusion in Intellectual Property Rights (IPR). Academics were not sure who owned the copyrights in research papers, as of now, changing attitudes of publishers mean that copyright assignment to publishers is probably less of a problem than it appeared to be a few years ago.

One response to the copyright problem has been for institutions or research funding bodies to attempt to reassert their own rights. Some universities are beginning to insert corporate ownership of intellectual property rights in university statutes and employment contracts, especially with regard to patents or learning resources. Bide (14) has described the question of ownership of intellectual property rights of academics as one of the more contentious issues facing higher education today. He says that the terms of the UK's Copyright, Designs and Patents Act, 1988 means that copyright in works made in the course of employment would normally pass to the employer. However, in practice, he notes, "most academic institutions do not exercise this right with respect to copyrights in journal articles or in textbooks." Perceptively, Lynch (9) has warned against universities using institutional repositories as a means of asserting control or ownership of that intellectual work that has traditionally be controlled by academics. Unlike universities, those organizations that fund research cannot claim copyright over the publications that result from its grants. However, a number of funding bodies have recently expressed their willingness to require grantees to provide open-access to such publications. For example, the Wellcome Trust (http://www.wellcome.ac.uk/assets/wtx022820.pdf) are proposing that its grantees will be required to deposit electronic versions of research papers in PubMed Central.
(http://www.pubmedcentral.org) (or its European counterpart) within six months of publication. It is perhaps worth noting that policy changes that 'mandate' the deposit of research papers in institutional repositories could have significant organizational implications for those who manage them.

1.4.2.4 Access

Crow's short definition of institutional repositories says that they should be "...accessible to end users both within and outside of the institution, with few if any barriers to access" (2). While this is fully in accord with open-access principles and may be desirable for most research papers, there are a number of reasons why institutional repositories may not make all content publicly available. Potentially restricted content might include research papers that have been retracted, reports or theses that contain commercially sensitive information, datasets that are in the process of being refined, learning resources, or administrative records. Some of this material may be legitimately distributed within the institution itself (or parts of it), but the content (and its descriptive metadata) would not routinely be made available outside. Repositories, therefore, may need to define access levels for different types of content, e.g. to place access control mechanisms on restricted content types while ensuring that adequate descriptive metadata is made available for that content that can be shared more widely.

1.4.2.5 Sustainability

A more general collection management issue is the long-term sustainability of the repository itself. This has two main aspects; firstly the need for ongoing strategic and financial support from the host institution, secondly the need to ensure continued long-term access to the content of repositories.
Lynch (9) has said that it is "vital that institutions recognize institutional repositories as a serious and long-lasting commitment to the campus community (and to the scholarly world, and the public at large) that should not be made lightly." He warns that repositories can fail for a number of reasons, e.g. lack of strategic or financial support from institutions, management failure, and technical problems. In addition, institutions are not the stable entities that they sometimes appear to be. For example, new departments or research centers can be opened, old ones can be closed, merged with others, or move to a different institution. Also, while higher education institutions rarely close down completely although this could theoretically happen, they do increasingly merge with others. (http://education.guardian.co.uk/universitymergers/)

For this reason it is important that institutional repositories secure high-level political support within institutions. Also that they should develop contingency plans that can be implemented if and when circumstances change. This may mean making arrangements with other institutions or with 'repositories of last resort' like national libraries. In this context, it is perhaps interesting that a growing number of journal publishers have begun to make similar arrangements, e.g. publishers like Elsevier Science, Kluwer Academic, Blackwell, BioMed Central, Oxford University Press and Taylor & Francis have all recently signed deposit agreements with the National Library of the Netherlands. (http://www.kb.nl/nieuws/)

Ensuring long-term access to the content of repositories is yet another challenge that will need to be faced by institutional repositories. A JISC-funded feasibility study on the preservation of eprints highlighted the importance of file formats, metadata and organisational strategies (15). Existing e-print repositories often only accept a limited number of file formats, usually based on the perceived download preferences of users. These typically include a mixture of proprietary and 'open' formats, e.g. HTML, PDF, PostScript, MS Word, MS PowerPoint and TIFF. It is recommended in the Joint Information Systems Committee (JISC) study that repositories should assess the preservation risks of file formats in their collections and consider format conversion, e.g. to those based on open standards or Extended
Markup Language (XML). There is also a need to maintain information about stored file formats, possibly in co-operation with third party format registries like the proposed by Global Digital Format Registry. (http://hul.harvard.edu/gdfr/) The need for appropriate metadata to support digital preservation processes has been recognized for some time. Further, repositories should collaborate on the production of a common set of preservation metadata. Progress on both of these issues will be important but ultimately the long-term preservation of repository content will be dependent on the development of appropriate organisational strategies. This suggests that institutional repositories may eventually need to become trusted digital repositories.

A working group sponsored by the Research Libraries Group and Online Computer Library Center (OCLC) has defined some of the main attributes of trusted digital repositories. To summarize, these include the need to accept responsibility for the preservation of content, to obtain sufficient control over content in order to be able to preserve it, to demonstrate financial sustainability and organizational viability, to ensure that there are documented policies and procedures that can be monitored and evaluated, and to adhere to standards and best practice (16). Most existing repositories would have difficulty fulfilling all of these criteria. Instead, it is observed that many existing e-print repositories are focused primarily on access rather than preservation and that project-based funding may not be the best way of building long-term sustainability. The best way forward may be for repositories, where necessary, to co-operate with specialist sources of preservation expertise or third party preservation services.

1.4.2.6 Digital Preservation

The growth of institutional repositories and their valuable digital content compel digital preservation, which is a complex process such organizational, managerial and technical issues. In addition it is required to solve problem of creating repositories, depositing content and discovery thereby encouraging the necessary cultural change. Digital preservation has not been embedded as an integral part of the
repositories workflow and also there is neither much experience nor commonly agreed practice, as to how content would be preserved in a long term basis.

Digital preservation is a process is to ensure the usability, durability and intellectual integrity of the information contained in IRs, which in precise is the storage, maintenance, and accessibility of a digital object over the long term, it is viewed as ‘series of managed activities necessary to ensure continued access to digital materials’ and in managing the data and the accessibility of these data, not just storing the bits and performing regular backups, but it involves understanding the data over the years (17).

### 1.4.2.7 Requirement of expertise and collaboration

The development of IR is a collective effort, involving scientists/researchers, IT personnel, computer services and the library. Institutional repository requires a large digital storage, powerful servers and technology expertise which are likely to be found within a computer services division. The role of library in establishing IR is in metadata submission, metadata application, discovery mechanism, and preservation etc that are identical with the core functions of the libraries. Further it includes existing relationship with the researchers and level of trust that researchers have for the libraries. The expertise required by the librarians are metadata and preservation, good graphic and visual design, network abilities including authentication, firewalls, knowledge of database and data storage, backup facilities, marketing and personal relationship including IRs usefulness. Librarians also have to have good knowledge of copyright law and support from users group.

IR managers should aware of the potential users of IR, mainly researchers for their current practices of archiving their research papers and their interactions with their peers through informal survey of the researchers. For successful implementation of IRs, collaboration among various stake holders is very much required. However, top-down support must be required for implementing the IRs project. Many of the
studies suggest that a mandate from the organization would have less resistance in implementing IRs. Further, libraries can play a very important role in establishing the IRs as a successful venture as it has trust and relationship with the research community as well computer division (11).

1.4.2.8 Evaluation

Institutional repositories as data providers will also need to provide a full range of services for its user communities. These may include, e.g. managed data storage, metadata creation and enhancement, search and retrieval, the export of metadata about research outputs to research assessment procedures. As such, repositories will need periodic evaluation to ensure that they are fulfilling its basic institutional requirements and user needs.

1.5 Aggregator perspectives

In the model promulgated by the OAI-PMH, service providers are those that issue protocol requests to data providers and use the metadata as a basis for building value-added services. (http://www.openarchives.org/documents/FAQ.html) Some service providers concentrate on aggregating metadata content from multiple data providers. For example, the experimental ARC service from Old Dominion University harvests metadata from over 80 data providers and stores them in a searchable relational database. (http://arc.cs.odu.edu/) The University of Michigan's OAIster project harvests metadata describing a wide range of digital content from many different institutions; in 2011 the service gave access to over 23 million records from almost 1100 data providers. (http://www.oaister.org/) Other service providers focus on providing specific types of value-added functionality. For example, the experimental Citebase service developed by the University of Southampton extracts reference data from the full-text of papers and combines this with harvested metadata this to build a citation database. (http://citebase.eprints.org/) The ePrints UK project also developed an experimental service provider that was intended to aggregate and
federate access to UK-based repositories and test some added value services related to metadata enhancement and citation linking. (http://eprints-uk.rdn.ac.uk/search/)

The collection development policies needed for service providers will have some superficial similarities with those needed for data providers, however their implementation will be to a large extent dictated by the amount and quality of metadata made available by repositories. Before proceeding, it is perhaps worth noting that institutional repositories can themselves be service providers, e.g. aggregating metadata about content from a number of data providers under their control. In addition, that service providers can also be data providers, to the extent that they can make the content they harvest available through the OAI protocol.

1.5.1 Content

Services that aggregate content from multiple data providers have to resolve some of the same collection development questions as repositories. In addition to selecting which data providers to harvest metadata from, aggregators also need to make decisions, e.g. about the type, quality level, subject or geographical origin of the content that they make available through their service.

However, there needs to be sufficient metadata available to support these choices. For example, an aggregator wanting to focus on harvesting only metadata about peer-reviewed research outputs would be dependent on these being held in a separate repository from non-reviewed outputs or this information being clearly included in each record’s metadata. Similar principles would apply if the aggregator were only interested in harvesting metadata about research datasets, theses or learning resources.

Ascertaining subject coverage is potentially more problematic. Obviously, it may be possible to select content based on its provenance in particular academic
departments or schools, or on an analysis of the terms provided in subject fields. An alternative approach would be to harvest the full-text of content (where this is available) and undertake some kind of automatic classification process. For example, the ePrints UK project has considered using automatic classification tools developed by OCLC Research for helping to distribute repository content amongst the eight faculty-level services of the Resource Discovery Network (RDN).

Another potential issue related to content is the potential duplication of content. In practice, research outputs are sometimes going to be submitted to multiple repositories, e.g. where they originate in more than one institution. It will be the task of the aggregator to ascertain how much of a problem this might be for users of the service and, where necessary, take steps to duplicate.

There may be similar problems with different versions of the same content. Ultimately, the service provider has no direct control over content. Repositories may permit submitters to remove outputs, or may restructure or close down. The collection development strategies of aggregators will have to take this into account.

1.5.2 Technology

The recent growth of networking of institutional repositories is the emergence of enabling technology and also the availability of readily IR software applications for setting up repositories from open source software as well from propriety vendors. There are large numbers of harvester available as open source software and some of them are widely used in building national level system through harvesting number of repositories using OAI-PMH. The most prominent among open source software is Public Knowledge Publishing (PKP) Harvester used widely in India as well throughout the globe. Digital Repository Infrastructure Vision for European Research (DRIVER) is one of the largest network of more than 2000 repositories mainly European repositories has developed D-NET, open software based on service oriented architecture and make it available to end users through open software (18). Digital
Commons is a proprietary solution developed by BePress and marketed by ProQuest to build a network of repositories comprising repositories registered with Proquest and also institutions running EPrints or DSpace software. Therefore, it is appropriate for a national level institution to consolidate national resources into a single virtual archive to provide seamless information from a single window to end users locally as well to globally.

1.5.3 Metadata harvesting

The sufficiency and quality of harvested metadata is the *sine qua non* for aggregator services. In this, however, they are again almost totally dependent on the data provider repositories that they harvest from. Unfortunately, studies of metadata usage in OAI contexts suggest that quality varies greatly, leading to 'collisions' between metadata formats and problems with authority control and de-duplication (19). Safeguarding the quality of metadata is fundamentally the role of the data provider repositories.

As we have noted already, this depends on the provision of well-designed submission workflows, interfaces and quality control, also (perhaps) support from library and information professionals. Unfortunately the last of these, when scaled to the task of growing repositories, may appear expensive. However, in a genuinely distributed repository system, descriptive metadata for each output would only need to be created once, which has some cost advantages over current library cataloguing practice. Lagoj (20) suggest some practical things that may help improve the creation of good-quality metadata. These include the provision of content guidelines for metadata, the improvement of metadata creation tools, and the implementation of appropriate quality control processes.

As with repositories themselves, service providers may also have an interest in the automatic generation, capture or enhancement of metadata. Those interested in digital preservation have already experimented with tools that can capture technical metadata about formats (e.g., [http://hul.harvard.edu/jhove/](http://hul.harvard.edu/jhove/)), but the capture of
descriptive metadata directly from the full-text of harvested content is something that needs to be explored further in an OAI context. The use of structured text formats like those based on XML may help with this, as might software based on the Document Object Model (DOM) developed by the World Wide Web Consortium. (http://www.w3.org/DOM/)

The explored the use of various proposed third party services (using Web services technology) that could be used to enhance the metadata harvested by aggregators. The first of these was a name authority service developed by OCLC Research that would validate the forms of personal and organisational names in use, thus improving consistency in the use of names. The second, also based on technology developed by OCLC Research, was for a service that would automatically assign subject classification terms, thus helping to ensure consistency in subject metadata across the service provider and enabling harvested content to be accurately distributed to more focused subject-based services. A third service, provided by the University of Southampton, would parse the bibliographic references in the full text of harvested content into structured forms, thus facilitating citation linking and informetric analysis. All of these services have some potential to support the added-value functionality that can be offered by both repositories and aggregators. Their needs, however, to be far more work in developing such third party tools and testing their use in realistic harvesting-based contexts (21).

1.5.4 Intellectual Property Rights

While institutional repositories as data providers are concerned with the intellectual property rights vested in both repository content and metadata, service providers are primarily concerned with metadata. This includes both the metadata that they harvest from data providers and any enhancements that they make to it in the process of offering their service (22). Gadd, Oppenheim & Probets (23) note that the copyright status of metadata under UK law is uncertain, although individual records could obtain protection as a compilation and collections of metadata as a database. They conclude that, in order to avoid infringing database right, service providers
wishing to make use of a data provider's metadata should seek permission to do so. The Rights MEtadata for Open archiving (RoMEO) project survey showed that while most data providers agreed that collections of metadata enjoyed database right, they also felt that this might be implicitly waived in OAI contexts. Service providers were equally divided between those who did check the rights status of metadata before harvesting, those who considered it to be implicitly free under OAI, and those who had never thought about it. A majority were happy for others to harvest their enhanced metadata, although some conditions of use were specified.

There is, perhaps, a need for aggregators to define their own working assumptions about intellectual property rights with regard to harvesting metadata (or full-text) from repositories. Service providers should also consider publishing conditions of use where they make their own enhanced metadata available for harvesting by others.

1.5.5 Sustainability

While aggregator services do not have content that requires long-term management and curation, they do have a need for organisational and financial stability. Service providers mediate between content providers and their users and, as such, their sustainability will depend on whether they provide a service that is required and how well they perform this in a potentially competitive market. As services, however, they can evolve over time, both with regard to the content they harvest and the added-value services they provide.

Many service provider initiatives are funded on a project basis, e.g. services demonstrating the added value that can be provided by metadata harvesting using the OAI protocol. In the future, there is likely to be a 'mixed economy' of service providers, with some supported by public funding while others will be provided on a commercial or quasi-commercial basis.
1.5.6 Evaluation

As user-focused services, aggregators will need ongoing evaluation to ensure that they continue to fulfill user needs and to help identify new requirements. The results of such evaluation may lead to changes in collection development policies.

1.6 Proposed Model of NDRS

NDRS is largely the products of inter-institutional efforts to accommodate digital-based materials and cost-effective solutions to collections (pool resources and storage) which offer a shared space, subjected to common standards and managed by a single organization. In the present Internet era, it is must to design and build the infrastructure necessary to connect all of country’s digital research repositories as National Digital repository System (NDRS) that meet standards for interoperability and access to showcase and preserve the research outputs of the country. As such, there does not exist any national level consolidation in health sciences, however, the existing service providers such as Cross Archive Search Services of Indian Repositories (CASSIR) of Indian Institute of Science (IISc) is general covering 22 IRs of the country and is still project based. Therefore, it does not serve the national requirement as it does not have any defined policy for harvesting data from the repositories. There are other discipline based service providers such as SEED of IIT covering Technology based literature, which is not updated for a long time. CSIR’s Open Archive Repositories Harvester (OARH) is harvesting metadata from six institutes of CSIR consolidating metadata and yet to reach the critical mass in terms of articles harvested.

The present study evolved a model and its institutionalization called National Digital Repository system (NDRS) in the field of Health Science through cooperative collection, storage and management initiatives that can shape and support new strategies for institutional repositories in the country. From the various models available, the present study recommends Aggregation Model as it offers more advantages over other models. It will have its own participatory policy and
programmatic direction as it provides opportunity for representation from participatory IRs, experts from scholarly contributor in the subject, technologists, finance (a representative from the funding apex body) and NDRS administrator as Member-secretary. Further, It would enable cooperative-management, collection-sharing initiatives and ensure the collective interest of the IRs and maintain governance and budgetary control over them.

1.7 Methods and Materials

1.7.1 Need of the study

Libraries and scholarly communication are increasingly collaborating with each other and sharing as system-wide services to deliver comprehensive and high-quality information services to faculty and students. It also needs to set out strategic directions that will further leverage available resources of the libraries for expanded development and management of shared collections; elaboration of shared services; greater utilization of shared facilities; ensuring persistent access to digital information that result from and supports research and teaching and changing the economics of scholarly communication (24).

These strategies extend the growing cooperation and collaboration among the libraries to share collections, leverage technology and pool financial resources. Many of the benefits that have been achieved through the strategy are progressively expanding collaboration. The strategic directions set out are intended to address a number of challenges faced by the libraries. There are many enduring challenges as (24):

- Diminishing budgets and continued accumulation/preservation of information resources.
- Proliferation in the amount of scholarly information.
- Proliferation in the number of formats in which scholarly information is produced and disseminated.
• Lack of any unifying technology that would facilitate the easy capture, organization, presentation, use and long-term management of that growing component of scholarly information that is produced in digital form.

• Much higher expectations among users about the speed and ease with which such information (and associated user support services) should be discovered, located, accessed, and used.

• Rapid change in information technology and the accompanying need to adopt and adapt new technology to serve the information needs of faculty and students.

• The increasing financial and operational interdependence of the libraries in providing services to each other and supporting shared system-wide collections and services, and the resulting effects of independent campus budgetary and operational decisions on the quality of system-wide library services.

To meet these challenges and to sustain the excellent collections and services that characterize the libraries in the face of budget constraints, increasing cost of information, changing information technologies, and the shifting service expectations of the libraries’ clientele. Libraries are optimizing through the management and access to information resources by reducing unnecessary duplication, leveraging shared assets, while meeting the unique information needs of library users at each campus and enhancing the distinctiveness of each library’s collections and services.

Libraries, particularly as institutional repository need to collaborate in developing and implementing new services, and re-engineering existing services, so as to enhance existing resources – staff, budget, system so as to achieve cost avoidance and increased cost-effectiveness. The shared infrastructure can be economically supported, as national facility for improvements in existing services. Also it provides an enhanced experience for library users by more effective integration of information resources and services generated by their own faculty, students and research scholars including policy makers.
In the national consolidation it is to ensure that the digital information resources that are increasingly essential to support world-class teaching, research and scholarship are cost-effectively captured, preserved and managed by developing digital preservation infrastructure. In turn, it can be used centrally to preserve the digital information of common interest and that can be used by libraries to preserve digital assets. It also expands efforts to work with faculty and administration, scholarly societies and publishers to address problems. Thus, the economically sustainable development and maintenance of high-quality and readily accessible research collections is made possible.

Therefore, it is sincerely felt that study of designing a national digital repository system to project scholarly output of the country to all those want to have an access at a minimum possible time. As it is seen in most of developed and developing countries have made considerable progress in developing a dynamic national system to provide the visibility of scholarly content to users’ community. Realizing the paucity of consolidated and well structured central repository system, an attempt has been made in this study to design and develop a system which can enable user’s community to have a virtual seamless single source of health information in all kind of data types.

1.7.2 Statement of the Problem

Researcher being in the national level organization concerned with consolidation of information, building databases and applications for the government sectors and disseminating over the net, it was felt to make positive contribution by developing a system to facilitate users community of the country the research topic entitled “Design and Development of National Digital Repository System (NDRS) for Health Information in India: Descriptive Study.” is undertaken for study. The aim of the study is to facilitate a process by which National Digital Repository System (NDRS) can
determine coordination in fulfilling the needs for institutional repository capabilities in health sciences.

1.7.3 Objectives of the Study

The study undertaken is to identify the goals of institutional repositories with respect to coordination in context of National Digital Repository System, following objectives are derived:

- To identify the needs of the institutional repositories and deployment strategies to satisfy the institutional needs
- To identify the problems and prospects in preservation of internally and externally produced digital objects including technological and technical problems in maintaining the institutional repositories.
- To study the standards deployed to deposit various kinds of materials in digital repositories
- To determine the possible areas of synergy between the organization in the implementation of IRs.
- To study and derive functional and sustainable NDRS model to digital collection, management, preservation and communication with interoperation among institution in the country.

1.7.4 Hypotheses

- NDRS establishes better impact, effectiveness and visibility for the institutional repositories.
- Facilitates IR in storing, migrating and creating appropriate metadata and guarantee to depositor intellectual content preserved for a long time.
- Enables management effectiveness by providing expertise and technological solution to lessen the burden on individual institution repository.
- NDRS is more cost effective as cooperative programme and act as testing tools for preservation within the defined framework.
Success of National Digital Repository system depends on updating periodical updating of standards, best practices, protocols and processes.

1.7.5 Methodology

Basically, it is a descriptive study based on published literature in the form of research articles, conference proceedings, reports, thesis, products and services developed by institution. Interactions with system administrators to understand challenges involved in establishing national level systems, problems encountered, trouble shootings and maintenance of the system. Researcher being developer of one of the largest repository developed containing the official debates of Rajya Sabha (http://rajyasabha.nic.in) understands all aspects of institutional repositories. Researcher has conducted various workshops on institutional repository software and demonstrated various products developed using DSpace, Greenstone Digital Library (GSDL) to several organizations. The experience gained through participation in policy formulation and implementation programme in the government sector is appropriately integrated in the context of study. Evaluation of projects and programmes relating to research study has also added a greater insight to the study. The study has also undertaken the various National level systems developed in the country by visiting their websites and understands the kind of services offered by them.

Thus, for the pilot test of the study, it was found that the awareness about the national consolidation of the Institutional repository or their compliance to develop national repository is poor. Hence it was felt appropriate to apply “meta-analysis approach for systematic conduct of research work. A brief outline of this approach is given in the following paragraph.

Meta-analysis as defined (25) is “The statistical analysis of a large collection of analysis results for the purpose of integrating the findings. The primary purpose of meta-analysis is to provide the same methodological rigor to a literature review that we require from experimental research.
There are two types of meta-analyses

- Common use of meta-analysis has been in *quantitative literature reviews*. These are review articles where the authors select a research finding or effect that has been investigated in primary research under a large number of different circumstances. They then use meta-analysis to help them describe the overall strength of the effect, and under what circumstances it is stronger and weaker.

- Recently, as knowledge of meta-analytic techniques has become more widespread, researchers have begun to use *meta-analytic summaries* within primary research papers. In this case, meta-analysis is used to provide information supporting a specific theoretical statement, usually about the overall strength or consistency of a relationship within the studies being conducted. As might be expected, calculating a meta-analytic summary is typically a much simpler procedure than performing a full quantitative literature review.

The study being on compliance and comparison of tools, technology, experts and expertise, it needs to depend on primary research undertaken at an informed environment, hence the meta-analysis is considered more suitable to develop on the basis of analysis of the practiced successful models and best practices including primary research.

### 1.7.6 Study parameters

Literature review has revealed the experiences of the earlier researchers as well as system administrators as to analysis and systematic evaluation of the IRs in general and different subject fields and developed a model for future. Personal experience gained in the workstation also guided the researcher to identify some of the parameters which are very effective in developing the system. Thus, the selected parameters are listed below:

- Visibility
- Research impact
- Accessibility
As per the objectives of the study there are two main components

- Institutional Repositories (as data providers)
- National consolidation (Service Providers) – In context of this study it is called as National Digital Repository System (NDRS). Hence the first part of the study is on the existing status and it compliance to develop NDRS.

1.7.7 Scope and Limitations

Broadly speaking, the development of institutional repositories in the country is in general in nature, it is not grown up in any particular discipline. Though attempt has been made to cover all possible IRs in the country to understand the nature of work in these IRs, however, the focus of study is in the design and development of National level system in the field of health sciences. The duration of the work is limited for last decade.

1.7.8 Chapterisation

- Chapter 1 - Introduction
  This chapter introduces the topic of research and highlights the studies which have initiated to design the study. Further, the objective hypothesis and
methodology for the collection and analysis of data is mentioned briefly. In addition, limitation of the study and chapter scheme is provided.

• **Chapter 2 - Review of Related Literature**
  This chapter covers the review of the published and unpublished literature pertaining to the research topic. The entire review is divided into five facets.

• **Chapter 3 - Growth and Development of Institutional Repositories**
  This chapter deals with the growth and development of IRs and also highlighting the structure, functions, required software, including benefits to stakeholders.

• **Chapter 4 - Models of National Digital Repositories System (NDRS)**
  This chapter provides very selective narration of the recognized models which are in operational including salient features of the models. It also covers IPR issues, technical and long term preservation issues etc.

  **Chapter 5 - Analysis & Interpretation of Existing IRS & NDRS in India**
  This chapter covers the analysis and interpretation of the collected and tabulated data.

• **Chapter 6 - Proposed Model of National Digital Repositories System and Institutionalization in Health Sciences**
  This chapter contains the detailed description of the proposed NDRS with essential components. The accepted pre-requisites of the design and development of the model is discussed.

• **Chapter 7 - Findings, Recommendations & Conclusion**
  This chapter presents the findings of the study and feasible suggestions in the form of recommendations along with the conclusion of the entire research study. Finally, references are accumulated and listed at the end in an alphabetical order so as to enable the future researchers to get an idea of volume and variety of the literature published in the field.

### 1.8 Conclusion

This report has attempted to identify some of the main types of collection development issues that need to be considered when setting up both institutional
repositories and aggregator services. Collection development is largely concerned with content. As with any other digital library, decisions need to be made about the types of digital object should be included and which should be excluded from repositories. While the primary focus of a repository is defined by its institutional scope, additional decisions need to be made about the exact type of content required, for example, whether it should just be for peer-reviewed research papers and theses, or for all research outputs and learning resources.

It is recommended that institutional repositories (as data providers) should consider the production of written collection development policies. These should at the very least define the intended scope of the collection with regard to subject matter, object type and quality (this could be part of a collection-level description). The policy could also contain information on submission workflows (including intellectual property rights clearance), access, and include some consideration of sustainability. Some of this information, for example, is included in the guide for depositors produced by University College London, which includes information on the eligibility of papers, copyright clearance, eligible formats, metadata creation and upload procedure (http://eprints.ucl.ac.uk/deposit.html).

Aggregators need to make similar decisions about content to repositories, e.g. about subject coverage, the type of resources to be included, whether it needs to be peer reviewed. However, their choices may be limited by the extent and consistency of metadata made available by repositories. For example, the consistent use of a standard type-list would facilitate the selection of repository content by output type. In the absence of consistent metadata, it is possible that aggregators could use software to support the selection of content, e.g. by using automatic subject classification or output type identification tools. In some contexts, it may be possible for repositories and aggregators to co-operate on the development of common metadata schemas that can support a wider range of functionality than is possible with the simple Dublin Core metadata mandated by the OAI protocol for basic cross-domain interoperability.
It is recommended that repositories should develop (or adopt) metadata generation tools that facilitate the production of consistent metadata. This may be helped by the use of suitable metadata content guidelines, e.g. those for the use of simple Dublin Core to describe e-prints. Consideration should also be given to the use of software tools that can automatically capture some types of metadata, also to the integration of third party metadata enhancement services like OCLC’s name authority service.

In addition, co-operating repositories and aggregators may need to consider if there are additional types of metadata (e.g. about rights) that are necessary to support value-added services, and make arrangements to share this using the OAI protocol.

1.9 References


