Managing Research Data in Academic Institutions: Role of Libraries

Mallikarjun Dora H Anil Kumar

Abstract

One of the global emerging trends in academic libraries is to facilitate the management of research data for the benefit of researchers and institutions. The purpose of this paper is to explore the role of a library in offering such research data management services. The paper discusses the importance of research data, its preservation, organization, dissemination and critical role in the scholarly research lifecycle. The authors attempt to provide a vivid description of Research Data Management (RDM) as a service and in the process review the existing literature on the topic in addition to the indicating the tools and technologies that could be adopted in successful RDM service implementation. The paper also is an attempt to share the experience of creating the Vikram Sarabhai Library’s research data repository that was developed by adopting the open source software - CKAN.

Keywords: Research Data Management, Research Data Service, CKAN, Research Data, Data Repository

1. Introduction

Scholarly research is an important indicator of national development and reflects the potential of a nation to harness its human resources to solving the problems of mankind. Global problems of health, education, poverty, etc. can be better understood and addressed effectively through research. Scholarly research broadens the horizon of policy thinking in addressing many critical issues that governments face.

In most of the cases, scholarly research has in its core, the need for data. EPSRC (Engineering and Physical Sciences Research Council) defines research data as “recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created”. Research data can be a raw data directly produced from the lab or survey, or it can be processed data which has been cleaned, refined, arranged, and combined in a manner that it is useful in research. Research data also include data which are already published in the journal or in other scientific communication. Research data includes analogue sources as well as discrete digital objects (text, files, image, audio, video), complex digital objects (discrete digital object made by combining a number of other digital objects) and databases (Whyte and Tedd, 2011).

Researchers all over the world are generating large volumes of data sets for their research purpose. The advancements in information technology, availability of large number of electronic data sources and
powerful data analysis software, have together facilitated the researcher to generate and work with large data sets. The challenge is in preserving and disseminating these large data sets for future needs that would be very valuable to the academic community.

At the individual researcher level, they are not only generating data but also attempting to preserve for reuse and sharing. The researchers either save and manage the data by preserving it with them or place the data directly in the open repositories like figshare (www.figshare.com) or dryad (www.datadryad.org). Journals publishers have started maintaining their own data repositories to facilitate its authors to host data for preservation, verification and management. Research institutions are also initiating research data management services to preserve and manage data for future reuse and transparency in research. Management of research data has recently emerged as a strategic priority for the university (Pryor, 2012). There are a number of universities, scientific institutions and governments that are thinking on similar lines and have initiated the building their research data infrastructure and management (Pinfield, Cox and Smith, 2014).

The growing number in the registry of data repositories validates the need for such repositories by the academic and research fraternity. There are about 1000 research data repositories registered in each re3data (http://www.re3data.org/) and databib, (http://databib.org/index.php), which are among the major research data registries. Simultaneously the popularity of services like Figshare and Dryad, where individual researchers can keep his/her research data also indicate the potential value and need for research data management.

It is in this context that it may be worthwhile to review the role of libraries in providing research support to its users. Traditionally libraries have been the playing the role of a partner in research, right from the time a researcher initiates the process of identifying a research problem to the final publication of a paper in a journal or presentation in a conference. Library has a key role in organising, managing, preserving, discovering and disseminating of research to a wide and relevant audience. Extending its tradition role in providing research support, the libraries today have opportunity to facilitate management of research data also.

This paper explores the role of libraries and library professionals in Research Data Management (RDM) and also presents a case study of CKAN software and its implementation at Indian Institute of Management, Ahmedabad which can be used for creating and preserving research data.

2. Research Data Management

"Research data management concerns the organisation of data, from its entry into the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results and permits new and innovative research built on existing information" (Whyte and Tedds, 2011). A service to manage research data, in other words, a Research Data Management Service (RDM S), would have to consider the entire life cycle of research data, starting from the creation of data to its reuse. The UK Data Service provides for a clear depiction of the research data life cycle as shown in Figure 1.
Cox and Pinfield (2014) describe research data management service as consisting of different activities and processes that include creation, storage, security, preservation, retrieval, sharing and re-use of data including technical capabilities, ethical consideration, legal issues and governance framework.

There are many benefits an institution can derive through an effective research data management service. Some of them include:

- Long term preservation of data provides for validation check, in the future, of the data and this enhances the credibility and transparency of the research data used.
- Research data can be reused by the same researchers or maybe even others who may like to extend the use of such data for purposes unseen by the initial researcher.
- Well managed research data can always be updated to enhance or extend the understanding of existing research on this data.

Jones, Prior and White (2013) have developed a useful working level guide for higher education institutions planning to create Research Data Management (RDM) services for their institutions. The guide pro-
provides working description of why the RDM services are required, what are the roles and responsibilities of the stakeholders, the process of developing the service and a detailed description of the various components of the RDM service. The guide is also very useful for institutions planning to implement RDM services as it provides relevant links to useful links to a large body of training materials.

It may be worthwhile to look at the experiences of various institutions around the world that have been offering RDM services. Akers, Sferdan, Nicholls and Green (2014) review eight US universities, classified as research universities engaged in ‘very high’ research activity by Carnegie Foundation. The universities include Cornell, Emory, John Hopkins, Pennsylvania State, Purdue, Illinois at Urbana-champaign, Michigan and Virginia. The authors found that despite their differences in approaching the RDM implementation, most of the institutions face a common challenge in developing RDM support programs. The other major issues highlighted in the study include, challenges in reaching out to the researchers to improve research data management practices and seeking funding for new staff positions and infrastructure. It is also interesting that in the case of all these universities, respective libraries played a prominent role in the design and development of RDM services in their institutions.

3. Libraries and Research Data Management

To further explore the role of libraries in RDM services being offered at research institutions, it may be interesting to take note a few interesting papers existing on this topic. Gold (2007) described the potential role of libraries in managing data, with a focus on social science data, geo referenced data and bioinformatics data. Henty (2008) surveyed Australian universities to identify the existing data management practices and trends. He also explored the possible roles of libraries and librarians in this context. Lewis (2010) examined in detail the roles and skills of university librarians in UK in the context RDM and suggested upskilling of the existing library workforce through education and training on research data management.

One of the early surveys to study the preparation and attitude of librarians towards research data service was undertaken by Tenopir, Sandusky, Allard and Birch (2012). The survey was conducted among 223 librarians of the Association of Research Libraries (ARL) and the findings indicated that although there was very low percentage of libraries involved in RDM offerings, the librarians believed that this was an important service for academic research libraries to render. Similar findings were reported from the survey conducted by Corrall, Kennan and Afzal (2013) among 140 libraries in Australia, New Zealand, Ireland and United Kingdom. They also found that RDM service represents a relatively new development in library service offering though there was an interest among the libraries to offer RDS, with a high proportion of libraries in the process of planning to offer RDM services support.

Meanwhile, Tenopir, Sandusky, Allard and Birch (2014) extended their previous study to survey and understand the perceptions of 223 library directors in US and Canadian libraries towards RDM service. They found that RDM was not frequently employed in libraries but there were academic and research libraries that were already offering RDM services with more planning to initiate RDM in the next couple of years. There was a small but growing number of libraries that were becoming more in-
involved in RDM by helping with data management plans and preparing and preserving research data.

To examine the contribution of academic libraries to research data management (Pinfield, Cox and Smith, 2014) interviewed (semi structured) 26 library staff from different UK institutions. The study found that though libraries were playing an important role in RDM, there was a lack of consistent support from various stakeholders at the institution. The study also identified various factors and issues that were important for successful RDM service implementation. The study, based on its findings, proposed a new model for a RDM programme that could help overcome barriers to successful implementation of such RDM services.

Another report worth mentioning in this context is the final report of the LIBER working group on E-Science/Research Data Management (Christensen-Dalsgaard, 2012) that concluded with ten recommendations to libraries for providing RDM support.

Existing literature and studies on Research Data Management services indicate quite clearly that there are new opportunities for libraries and library professionals in this area. Library professionals have already established themselves as experts in metadata; data curation and preservation techniques and hence can now extend their role to research data management also. The library professionals can not only create infrastructure for the research data management but also extend help in designing institutional policies and frameworks, build a bridge between administrative staff and researchers in developing research data management services.

4. Research Data Registry, Repository and Software

Prior to embarking on development of a RDM service, it would be important to study various existing initiatives in the form of registries and repositories. A registry would typically list out various research data repositories, while repositories in themselves would be hosting the research data. Table 1 indicates that two main registries that list out research data repositories on various topics.

<table>
<thead>
<tr>
<th>Service</th>
<th>About</th>
<th>Founded by</th>
<th>Number</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Databib</td>
<td>Tool to locate online repositories of research data, originally sponsored by a Sparks! Innovation National Leadership Grant</td>
<td>Institute of Museum and Library Services, hosting by University of Purdue</td>
<td>993</td>
<td><a href="http://databib.org/index.php">http://databib.org/index.php</a></td>
</tr>
<tr>
<td>Re3data</td>
<td>Registry for research data repository that covers research data from different academic discipline</td>
<td>German Research Foundation Partner with • Berlin School of Library and Information Science • GFZ German Research Centre for Geosciences • KIT Library • Purdue University</td>
<td>1093</td>
<td><a href="http://www.re3data.org/">http://www.re3data.org/</a></td>
</tr>
</tbody>
</table>

The popular repositories of research data include Dryad, Figshare and Harvard Dataverse. Table 2 depicts the main features of these three research data repositories.

Managing Research Data in Academic Institutions...
One of the most important factors in successful implementation of a RDM service is selection of the software to effectively manage the research data. There are options available from the basket of Open Source software and include Databank, CKAN and Dataverse. Table 3 includes these open source options along with relevant website links.

### Table 3: Open Source Software Solutions for RDM

<table>
<thead>
<tr>
<th>Software</th>
<th>Created by</th>
<th>Software and Platform</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN</td>
<td>Open Knowledge Foundation</td>
<td>Open Source Software Linux Platform</td>
<td><a href="http://ckan.org/">http://ckan.org/</a></td>
</tr>
<tr>
<td>Dataverse</td>
<td>Institute of Quantitative Social Science at Harvard University</td>
<td>Open Source Software Linux Platform</td>
<td><a href="http://dataverse.org/">http://dataverse.org/</a></td>
</tr>
</tbody>
</table>

DataBank: DataBank is a scalable data repository designed for institutional deployment. It is an open source project that promotes free to use cloud hosted systems for management, preservation and publication of research data sets. DataBank was created by Oxford Bodleian Libraries.

CKAN: The Comprehensive Knowledge Archive Network (CKAN) is an open source data management platform adopted by numerous governments, organizations and communities around the world. CKAN is supported by Open Knowledge Foundation (OKFN) and is one of the most popular research data management software available. There are 116 CKAN instances around the world which...
cover all types of organizations including Government/Local Government/Academic/Community and Other organizations. Most of the Government open data sites are run on CKAN software. There are very few academic institutions who have adopted CKAN for their research data management needs.

5. Research Data Management Repository at IIMA

Indian Institute of Management, Ahmedabad is one of the leading business schools in India and the world. Its library, Vikram Sarabhai Library (VSL) services about 100 core faculty members, 100 research students, 1000 students and a good number of research associates working at IIM Ahmedabad. VSL also caters to researchers from around the country, mainly in the subject of business and management.

Business and management researchers need different kind of data sets for their research which they source from subscribed resources, open data or primary data collected by themselves through surveys and other instruments. The library and its staff play a key facilitating role in research data collection, in the context of subscribed and open sources. In addition to identifying the sources, library staff is actively involved in downloading, collecting, organising and disseminating the datasets to the library users, as per individual needs. These datasets can include raw data, processed data or analysed data. Over a period of time, with increased focus on research at the Institute, large volumes of datasets are expected to be generated. The challenge would be in appropriately preserving these datasets for future access and reuse.

It is pertinent that the library and academic research fraternity at Indian Institute of Management, Ahmedabad requires implementing a service that will be helpful to preserve the data in a standardised format for long-term access and reuse. An effective research data management service maybe the ultimate solution to this challenge and this paper proposes a RDM service for the Institute.

5.1 CKAN - Software solution adopted for the RDM service at IIMA

The existing open source software options were reviewed and the options considered for this review included Harvard Dataverse network, Databank and CKAN. One of the major factors in favour of CKAN was that there were many data repositories adopting it and it seemed to be quite popular with various national governments for hosting their data repositories. In addition to excellent features, CKAN also have a very active user community and this was one of the main reasons why it was selected to develop the IIMA RDM service.

CKAN had all the important function like data storage, licensing, metadata, persistent URL, authentication etc. which are very much required in research data management. The major features are:

- Integration with the institutional research environment (e.g. hooks into CRIS system, Institutional Repository, DMP Online, networked storage)
- Capturing the research process/context/activity; notation, not just data
- Controlled access to research partners
- Good, comprehensive search tools
- Version control for data and metadata
- Customisable, extensible metadata
- Adherence to data standards e.g. RDF
- Multi-level access policies
Managing Research Data in Academic Institutions...

- Secure, backed up, scalable file storage for anywhere access to files and file sharing (e.g. Dropbox)
- Command-line tools and good web UI for deposit/update of data
- Permanent URIs for citation e.g. DOIs
- Import/export of common data formats
- Linking datasets (by project, type, research output, person, etc.)
- Rights/license management
- Commercial support/widely used, popular platform
- Documentation for installation and customization is not comprehensive and could be a barrier for non-IT users.

5.2 Installation and Customization

Installation of CKAN can be done through three procedures (1) Installation from an operating system package (2) install from source (3) install using a Docker image. The recommended operating system for CKAN is Ubuntu 12.04 64bit. CKAN is written in Python, use Solr for search and relational database is PostgreSQL. A detailed manual has been given in the CKAN website for each procedure; one can use these procedures for installing CKAN for the respective data repository.

5.3 Creating a IIMA Dataset Repository

One may need to understand that dataset is unit of data or group of data and it can demographic data of a country, financial data, economic data, etc. Each dataset comprises of two parts, first one is the metadata of dataset which contains the following fields:

- Title, URL, Description, Tags, License, Organization, Visibility, State, Source, Author, maintainer

The second part of the dataset is the data which needs to be managed. CKAN supports most of the data formats that include Excel, CSV, PDF, XML, RDF, etc. The data could physically be stored internally or on an external link could be provided to the data host. It may be noted that a dataset may contain multiple types of data and each can be added separately.

5.4 Policy and Planning

CKAN provides for one master administrator who has full rights in designing the system. There can unlimited number of users of this system. Before adding datasets, it is required that the administrator creates ‘organizations’ and ‘groups’. Each organisation has its own administrator, who can decide the rights and responsibilities of registered users in that organisation. Each dataset is owned by an ‘organization’. Each ‘organization’ could have its own workflow and authorization procedure. Datasets can be made private or public, facilitating organisation administrator to decide whether a dataset can be limited to specific organizations or not. Administrator of the organization can also assign the role of editor or viewer to each registered user of that organisation.

The content of the RDM service for IIMA that has been initially planned includes:

- Datasets produced by IIMA researchers for their scholarly research and project and it may include raw data, processed data or analysed data.
- Datasets generated from open access or subscribed sources, by the library staff from dif-
The major objectives of a RDM service of IIMA are articulated as:

- Long term preservation of datasets generated by its researchers.
- Sharing data sets for collaborative research.
- Develop an archive of datasets already compiled in response to earlier reference queries so as to facilitate reuse of the datasets, not only by the researchers but also by the reference staff in addressing future reference requests.
- Standardize various datasets and integrate them into a single platform to facilitate search and discovery in addition to preservation.
- Assist the researchers in avoiding duplication efforts in compiling datasets.

Status of IIMA RDM Service

Currently the IIMA research data repository has 62 data sets in 2 organisations.

![Home Page of Data Repository](image-url)
Figure 3: Data Sets

Figure 4: Details of data set

Figure 5: Data Set Preview
6. Conclusion

Research Data Management service is an emerging as an important offering by academic. Infrastructure requirements, policy making, planning may differ from institution to institution in the context of RDM services. However one common critical success factor in the success of a RDM service is the active participation of the major stakeholders like director, administrator, library, IT staff and researchers. It is also important that a policy is developed and stated at the institutional level to give the right impetus and direction to such initiatives. Within the available options, institutions can create policies that indicate their desire to share their data openly or restrict the use for private communities.

The sharing of data is now increasingly becoming a globally accepted objective with governments mandating Open Access of data and research funded by it. The opening up of the World Bank data is a testimony to the fact that international organisations are also working on opening up access to their data.

At IIMA research is being emphasised for faculty recruitment and evaluation, leading to a situation wherein large volumes of research data would be generated. In addition to this type of data, an increase in research data based reference services by the library has forced the implementation of RDM service in the library. Instead of waiting to develop institutional policies for sharing or securing, the library has embarked on a path to create a research dataset repository first and then look at policy issues relating to access. The library is sure of the use of this RDM service for its users and also to the reference staff in the library.

Selecting the software, CKAN did involve considerable time and was finally identified to be appropriate as it was the best available in the open source domain. As the importance and implementation of such services increase among the libraries, in future, we will see that a number of technologies and tools would be available for adoption. It is envisaged that the research community of IIMA will find the existing RDM service useful and lead to more research at the Institute.

References


15. University of Lincoln, Orbital Project. Choosing CKAN for research data management. Available at: http://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management/


Website visited
1. http://ckan.org/
7. https://thedata.harvard.edu/dvn/

About Authors

Mr. Mallikarjun Dora, Professional Assistant
Indian Institute of Management Ahmedabad
Vastrapur, Ahmedabad.
Email: mallikarjun@iimahd.ernet.in

Dr. H Anil Kumar, Librarian, Indian Institute of Management Ahmedabad, Vastrapur, Ahmedabad
Email: anilkumar@iimahd.ernet.in