Digital Curation of Research Data in Social Science Research Library: A Successful Experiment at ISEC, Bangalore (India)

Mr. B.B. Chand
Deputy Librarian, Institute for Social Economic Change (ISEC)
and Research Scholar, DLISc., Bangalore University
Nagarabhavi, Bangalore - 560 072, E-mail: bbchand@isec.ac.in

Dr. Ramesha
Professor, Dept. of Library and Information Science
Bangalore University, Bangalore – 560 056, E-mail: bbramesha@gmail.com

Abstract

Digital curation (DC) is the selection, preservation, maintenance, collection and archiving of digital assets. It establishes, maintains and adds value to repositories of digital data for present and future use. Digital curation of research data across disciplines has significant advantage to the research community. Gathering research data is not only time consuming but involves huge cost. Moreover by making the data available to the wider research community will not only enhance the knowledge on the subject with increased research output but also improve the rate of return on the research investment. Social scientists use micro data to study the social and economic behaviour of research units. Over the last decade policy change by the sponsoring agencies also make it mandatory to archive research data for the benefit of the research community. In this paper authors have studied the different standards such as Data Documentation Initiative, Qualitative Data Exchange, PREMIS, Statistical Data, Metadata Exchange, and tools available for sharing social science research data. Five major metadata standards are used for description and exchange of research data is being explained. Further the article highlight the tools that can be used for digital curation of research data. Major data archives in social science are also mentioned to study the best practices and current situation. Authors proposed a working and viable model for digital curation of social science research data available in Dr. VKRV Rao Library, Institute for Social and Economic Change, Bangalore, India.

Keywords: Digital Curation; Data Curation; Research Data; Social Science Research Data; Research Data Service; ISEC, Banalore.

1. Introduction

Research stands on data and often need access to quality data for analysis and evaluation. Across disciplines whether science, technology, medicine or social sciences, data is an intrinsic part of the research process. Use of micro data in social science research is quite common now considering the research value it generates by studying the base unit. Frequently these data are being used by the researchers to study and interrelationship between these micro units. Administrative agencies and planners often use these data for effective policy making.

Though the data are being generated by various institutions and agencies in the past much effort were not being made to make this available to the wider research community. Considering the sheer advantages particularly to the research community and to the society generally institutions strongly felt the need for wider access to the research data. Technological advances and access to computing infrastructure has played a major role in this regard. Off late most of the research data are born digital only. Here it needs a special mention that the born digital is not only the final data set that was supposed to be archived but also the data were collected through digital devices (Mobile phone, I-pads, Online etc.). Hence, becoming easier to manage and process the data and make them available to research community.

2. Digital Curation

The words archives and preservation is being used the library literature since the very beginning and with the adaptation of digital technology this become digital archives and digital preservation. On the other hand curation was largely being used by the museum professionals. With the adaption of cyberinfrastructre for research environment this terminology was used as a process of data management which focuses beyond the digital archives and digital preservation. The lifecycle approach was not developed much in the other cases where as the digital curation focuses more on lifecycle approach to manage research data. The e-Science Curation report (Lord and McDonald, 2003)[1] differentiates digital curation from archiving and preservation as follows:

Curation. The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose.

Archiving. A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.

Preservation. An archiving activity in which specific items of data are maintained over time so that they can still be accessed and understood through successive change and obsolescence of technologies.

Digital curation involves selection and appraisal by creators and archivists; evolving provision of intellectual access; redundant storage; data transformations; and, for some materials, a commitment to long-term preservation. Digital curation is stewardship that provides for the reproducibility and re-use of authentic digital data and other digital assets (Lee and Tibbo, 2007)[2].

Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle. The active management of research data reduces threats to their long-term research value and mitigates the risk of digital obsolescence (Digital Curation Center)[3].

Digital curation of research data emphasizes on two most important factor long term preservation and persistent access. But at the same time utmost care also needs to be taken for data security and integrity and fair use. According to Jacobs and Humphrey (2004)[4], 'Data archiving is a process, not an end state where data is simply turned over to a repository at the conclusion of a study. Rather, data archiving should begin early in a project and incorporate a schedule for depositing products over the course of a project's life cycle and the creation and

preservation of accurate metadata, ensuring the usability of the research data itself. Such practices would incorporate archiving as part of the research method.

3. Research Data:

Data are the basic elements for most of the research and is being used quite diversely in the literature. "Data are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors." (National Research Council. 1999)[5]. A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen." (Consultative Committee for Space Data Systems. 2002)[6]. "Any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment." (National Science Board. 2005)[7].

The term "data" as used in this document is meant to be broadly inclusive. In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, or simulations), it refers as well to forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socioeconomic data; and other forms of data either generated or compiled, by humans or machines. (Borgman, 2012)[8].

Research data can be generated for different purposes and through different processes i.e field survey, laboratory experimentation, may be captured through sophisticated technologies. For a deeper understanding the research data can be divided into different categories. Table 1 explains the six basic categories into which the data can be categorized on the basis of generation of the data itself. These data are generated across disciplines.

Table 1 – Types of Research Data

Data Type	Characteristics	Example
Observational	data captured in real-time, usually irreplaceable	sensor data, survey data, sample data, neurological images
Experimental	data from lab equipment, often reproducible, but can be expensive	gene sequences, chromatograms, toroid magnetic field data
Computational or Simulation	data generated from test models where model and metadata are more important than output data.	climate models, economic models
Derived or compiled	data is reproducible but expensive	text and data mining, compiled database, 3D models
Reference or canonical	a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated	gene sequence databanks, chemical structures, or spatial data portals.
Records	Available as reports largely printed or digital in government or business or even public	Administrative documents, Business records

Source: http://www.bu.edu/datamanagement/background/whatisdata/ [9]

Research data may include all anything and everything related to the project. From the initial project note to the final report is part of the research data. Data recorded from the field in questionnaire to organized data as SPSS tables are research data.

4. Software Tools for Research Data Management

Since the idea of sharing research data in the area of social sciences gathered momentum many software were developed to cater to the needs of researchers and research institutions especially in the area of social sciences. Some of them were made available as open source where as few were developed for commercial purpose. Few such popular one which has essential features and is used extensively are as follows:

4.1 Data verse Network [10]: The Dataverse Network is an open source application to publish, share, analyze and preserve research data and associated documents. It makes data openly available to the fellow researchers and aims at facilitating full data sets with clear descriptions. Hence, understanding research and replicating the procedure become easier and research can be verified. Dataverse network uses Data Documentation Initiative metadata for detailed description. Highly organised structure and flexible containers are key to the software.

4.2. National Data Archive [11]: National Data Archive (NADA) is an open source web-based data management application developed by International Household Survey Network, which allows users to browse, search, compare, apply for access, and download relevant census or survey information through portals. NADA also uses Data Documentation Initiative (DDI), XML-based international metadata standard. This software has also good access control features to each documents in the containers.

4.3. NESSTAR Software System[12]: The Nesstar Server & WebView are commercial software developed by Norwegian Social Science Data Services (NSD) to publish statistical information that can be searched, browsed, analysed and downloaded by researchers. Nesstar software suite provides a way to make statistical data easy to publish, locate, access and protect from unauthorised access. Nesstar software systems also confirm to DDI metadata standards. Nesstar Publihser which is only a cataloguing software developed by the same agency is available as a free software.

5. Metadata Standards for Digital Curation

Description of research data is very important and critical as it not only informs the research data users about the data but also the different conditions in which the data were collected. It helps the users in getting a clear idea about the content. Many standards were developed by many agencies to detailed description of the data. Mentioned below are some of the well crafted metadata standard that is being used widely to describe data in the field of social sciences.

5.1 DDI - Data Documentation Initiative[13]: The Data Documentation Initiative (DDI) is an international standard for describing data from the social, behavioral, and economic sciences. DDI metadata specification supports the entire research data life cycle. DDI metadata captures descriptions related to data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.

5.2 OAI-ORE - Open Archives Initiative Object Reuse and Exchange[14]: Open Archives Initiative Object Reuse and Exchange (OAI-ORE) are standards for the description and exchange of aggregations of Web resources. Irrespective of the media types whether text, images, data or video can be described in detail for aggregation.The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation.

5.3 QuDEx - Qualitative Data Exchange Format[15]: QuDex is the Qualitative Data Exchange model for archiving and exchange of data and metadata between CAQDAS packages. The standard is developed by The UK Data Archive together with the social science data archiving community and international XML schema experts with respect to exchange both survey and qualitative research data based on XML schema.

5.4 PREMIS[16]: The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems.

5.5 SDMX - Statistical Data and Metadata Exchange[17]: For the efficient exchange and sharing of statistical data and metadata Statistical Data and Metadata Exchange (SDMX) was developed as technical and statistical standards and guidelines, together with an IT architecture and IT tools by the BIS (Bank for International Settlements), ECB (European Central Bank), EUROSTAT (Statistical Office of the European Union), IMF (International Monetary Fund), OECD (Organization for Economic Co-operation and Development), UN (United Nations), and the World Bank. This standard was developed taking into account the sharing of macro and aggregated data at a global level.

6. Social Science Research Data at ISEC

Social scientists often study the same subject for years through panel study to research on their economic progress and social behaviour. Information collected from these social units is used for various researches over years. These data are collected by different agencies with excellent planning and careful investigation to cater to the needs of larger audience. Often it is time consuming to do in-depth research which the primary data collection agency doesn't do or it may not be their goal. But many other researchers are keen to do further studies with critical analysis for which the availability of the data is must. Massive increases in the availability of informative social science data are making dramatic progress possible in analyzing, understanding, and addressing many major societal problems. (King, 2011). ISEC as a premier social science research institution acquire most of such micro data for in-depth analysis and to bring out the new evidences from the observed data. Following table provides the detailed microdata available at Dr. VKRV Rao Library ISEC.

Table - 2 : Microdata available at ISEC

Name of the Dataset	Primary Subject	Indicators	Ownership	Available at ISEC
Government Agencies				
Consumer Expenditure	Household Economy	Consumption, Food Security, Poverty	National Sample Survey Office	7 (1983, 1987-88, 1993-94, 1999-2000, 2004-05, 2009-10, 2001-12)

Employment and Unemployment	Labour	Employment, Unemployment, Education, Literacy, Work participation	National Sample Survey Office	8 (1983, 1987-88, 1993-94, 1999-2000, 2004-05,2007-08, 2009-10, 2001-12)
Education		Education expenditure, Tution	National Sample Survey Office	3 (1986-87, 1995-96, 2007-08)
Healthcare and Morbidity	Healthcare	Health facilities; Healt expenditure; Out patient care	National Sample Survey Office	3 (1986-87, 1995-96, 2004)
Others (Agriculture, Housing, Migration, Debt and investment)	Agriculture		National Sample Survey Office	5 (1991, 1992, 1993, 1998, 2003)
Annual Survey of Industries	Industry	Prodcution, Value Added, Labour productivity, Energy Use	National Sample Survey Office	5 (1994-95, 2000-01, 2004-05, 2007-08, 2009-10)
Independent Institutions				
Natioanl Family Health Survey	Healthcare	Literacy, Women Empowerment, Public Services	Indian Institue of Population Studies	3 (1992-93, 1998-1999, 2005-06)
District Level Health Survey	Healthcare	Literacy, Women Empowerment, Public Services	Indian Institue of Population Studies	
Rural Economic and Demographic Survey	Rural Development	Land Holdings, Household Expenditure, Gender Studies	National Council of Applied Economic Research	4 (1971, 1982, 1999, 2006)
India Human Development Survey	Human Development	Poverty, Education, Healthcare, Public services	National Council of Applied Economic Research	3 (1994, 2004-05, 2010-11)
Ageing	Demography	Social Security, Ageing	Institute for Social and Economic Change	1 (2011)

7. Research Data Curation at ISEC:

The objective of this digital curation is to facilitate single point access to microdata holdings at Dr VKRV Rao library. But at the same time we would like to know which users are using which data. Currently most of the data are available in CDs, and as and when the request comes the CDs are issued to them. Two main reasons why we consider using digital curation for managing research data are firstly it will provide a 24/7 access to research data to all the eligible users where as now the CDs were issued in library working hours only. Secondly we will be able to preserve the data and associated documentation for longer duration where as in the case of CDs it has to be replicated and preserved. The other most significant advantage that the model will achieve is that, the customized data tables that were generated from the datasets can again be curated along with data which may be useful for other researchers engaged with the similar research.

7.1 Software selection: After due evaluation of all the software available with us we decided to use Dataverse Network which use DDI metadata for detailed description. Also it has features which are specifically designed keeping digital

curation in mind. Features such as integration with R and linking between the microdata sets and subsequent research studies were quite noteworthy. The three main reasons for choosing Dataverse Network as the software are firstly it is open source software with a strong backing of Harvard University. Secondly the unique features make it way above the other software and thirdly, continuous development and improvement to the software along with a clear roadmap shows the commitment of the developers as well as the funding agencies.

7.2 Dataverse Network Software: The Dataverse Netwrok was released for the first time in 2008. Since then many versions were released for public and the latest version is 3.6.2. The pre-requisite software includes JDK1.6 or latest, Postgresql8.3 or higher and glassfish 3.1.2 or latest. It runs on both Linux as well as Windows environment. The key features that makes the software different from others includes;

- Both Basic and Advanced search are available across all dataverses and within a dataverse.
- Generate a formal data citation with a persistent identifier and URL.
- Files uploaded in SPSS, R and STATA offer additional subsetting and analysis services, and can be downloaded in multiple formats.
- User roles for contributing and reviewing.
- Study restriction levels: Public (complete access to all), restrict individual file(s) but leave cataloging information public and can restrict entire study.
- Study Templates allows creating templates based on a subset of cataloging fields, with pre-filled values.

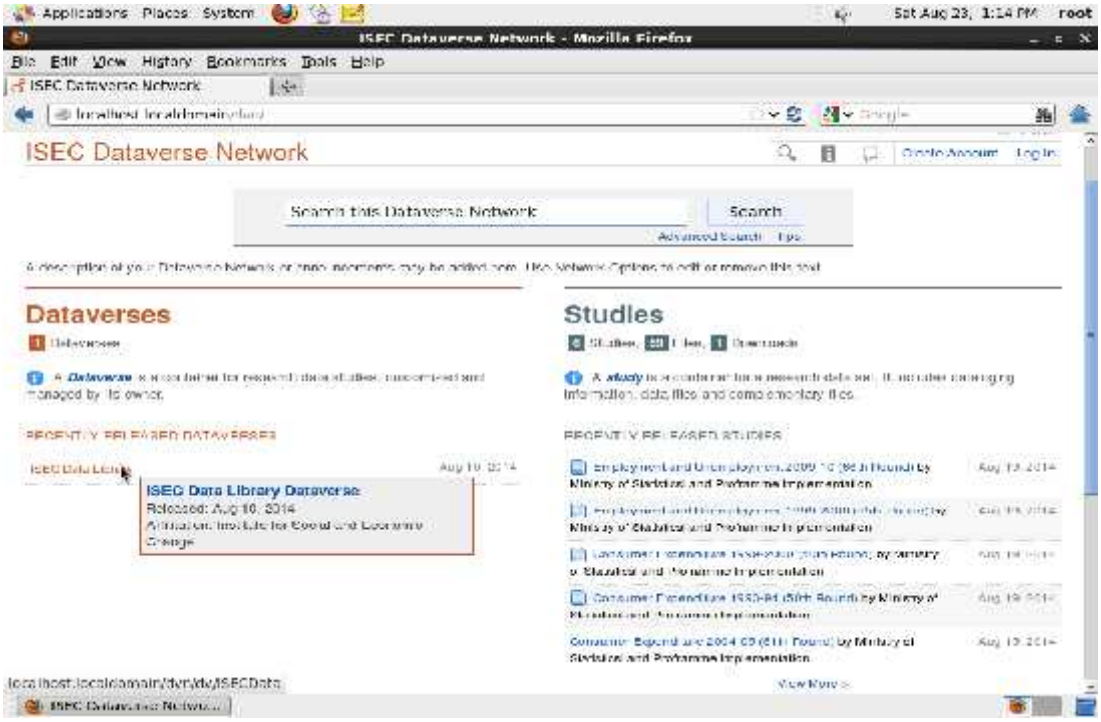

Fig-1: Dataverse Network home page

For our system environment we choose a CentOS 6.5 server as the operating system environment with default java and postgresql. Glassfish 3.1.2 was installed and configured for running the Dataverse Network Server.

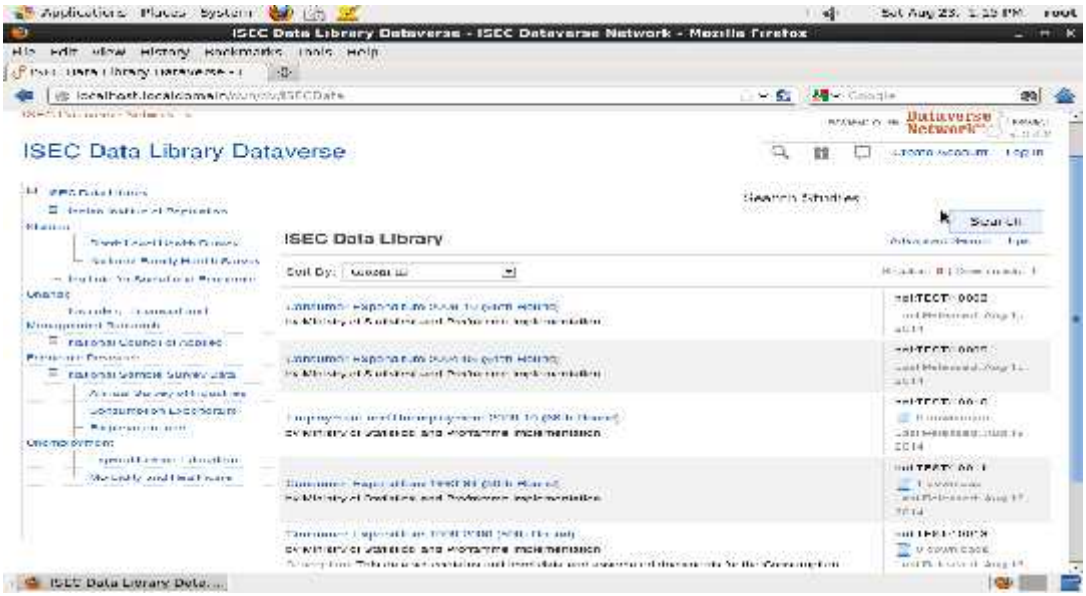

Fig-2: ISEC Data Library the container

7.3 Data Preparation and Uploading: Data acquired from different agencies mostly supported by good documentation. Survey tools, coding manual, tabulation manual are generally provided with the raw data. Where as in the case of small study sometimes it requires some special efforts to get the desired level of documentation which is essential for digital curation. After due discussion with the researchers it was decided to curate the data in all the formats whether it is simple text or coma separated values or even excel file. But preference was to preserve organized data either in SPSS, STATA or SAS files. Based upon the datasets the classifications were created. The records for the studies were created with the initial data available. For each classification a template was saved for all future cataloguing. Once the studies were created, it was mapped to respective classification for easy search and browsing.

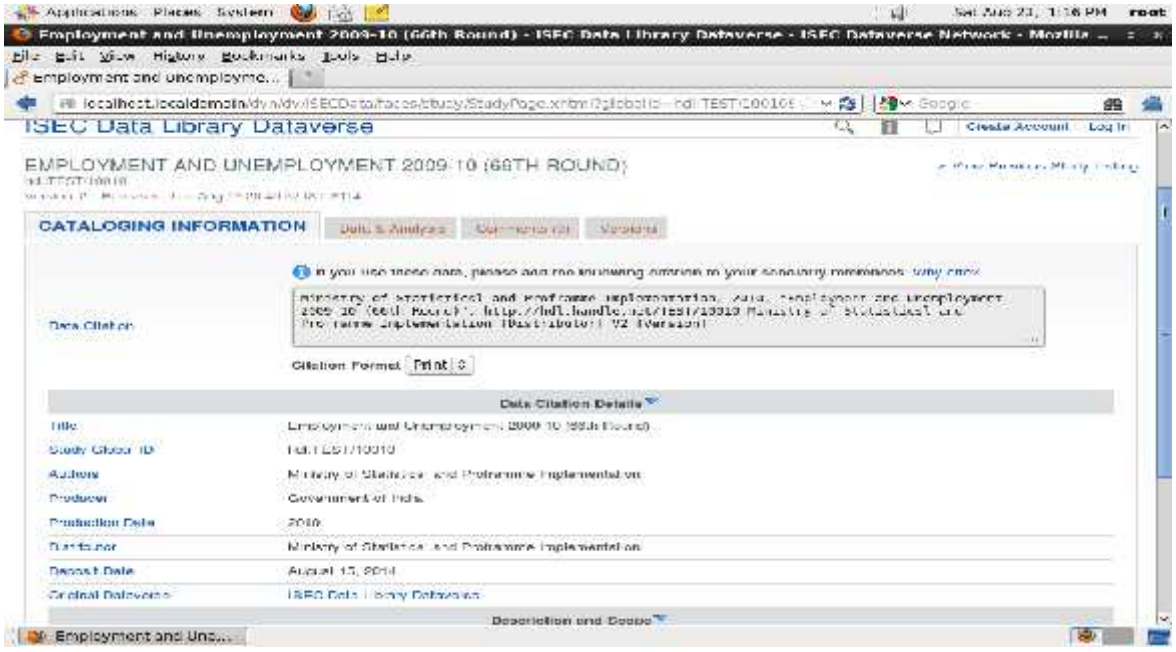

Fig-3: Metadata records for data sets

8. Conclusion:

Quick and efficient research data services to the users in social science research libraries are challenge to the librarians. Digital curation is a service by which the librarians can provide well-organized and easy access to the research data by using technology to the research community. Researcher's desire for new data drives data acquisition and that need to be curated for wider access and preservation. Open source software such as Dataverse Network is well architected to cater to the needs of both the research community as well as the data producers. Excellent features such as integration with statistical software for data visualization and strong security features with rights management, both data producers as well as the librarians shows keen interests in populating their data archives for the research community.

References:

1. Lord, P., & Macdonald, A. (2003). e-Science curation report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. Retrieved Oct, 25, 2008. http://www.jisc.ac.uk/media/documents/programmes/preservation/e-sciencereportfinal.pdf [Accessed on 16th July 2014]
2. Lee, C. A., & Tibbo, H. R. (2007). Digital curation and trusted repositories: steps toward success. Journal of Digital Information, 8(2). https://journals.tdl.org/jodi/index.php/jodi/article/view/229/183 [Accessed on 13th July 2014]
3. Digital Curation Center. What is digital curation. http://www.dcc.ac.uk/digital-curation/what-digital-curation [Accessed on 8th July 2014].
4. Jacobs, James A., and Charles Humphrey. (2004). 'Preserving Research Data.'. Communications of the ACM 47(9): 27-29.
5. National Research Council. 1999. A Question of balance: private rights and the public interest in scientific and technical databases. Washington, DC: National Academy Press, p. 15 http://www.nap.edu/catalog.php?record_id=9692 [Accessed on 16/01/2014]
6. Consultative Committee for Space Data Systems. 2002. Reference model for an open archival information system (OAIS). Washington, DC: National Aeronautics and Space Administration, p. 1-9. http://public.ccsds.org/publications/archive/650x0b1.pdf [Accessed on 16th January 2014[
7. National Science Board. 2005. Long-lived data collections: Enabling research and education in the 21st Century. Arlington, VA: National Science Foundation, p. 13. www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf [Accessed on 16th January 2014].
8. Borgman, Christine L. (2012). The Conundrum of Sharing Research Data. Journal of the American Society for Information Science and Technology 63(6): 1059-1078.
9. Boston University. What is research data. http://www.bu.edu/datamanagement/background/whatisdata/ [Accessed on 08 July 2014].
10. thedata.org [Accessed on 15th July 2014].
11. http://www.ihsn.org/HOME/node/98 [Accessed on 15th July 2014].
12. www.nesstar.com/help/4.0/server/index.html [Accessed on 15th July 2014].
13. http://www.ddialliance.org/ [Accessed on 13th July 2014].
14. http://www.openarchives.org/ore/ [Accessed on 13th July 2014].

15. http://www.data-archive.ac.uk/create-manage/projects/qudex [Accessed on 13th July 2014].
16. http://www.loc.gov/standards/premis/ [Accessed on 13th July 2014].
17. http://sdmx.org/ [Accessed on 13th July 2014].