Congregating Socio- Economic Datasets for Scholastic Research: A Case Study in IIMB Library

K Rama Patnaik
Librarian,
Indian Institute of Management Bangalore
Bannerghatta Road
Benguluru-560076
Tel: 91-80-26993016
Mob: 91-9740240038
Email: rama.patnaik@iimb.ernet.in

ABSTRACT

Digital curation initiatives with an intention to preserve the intellectual content of an institute gained momentum in the early 1990s with open source technologies facilitating such efforts. Indian Institute of Management Library (hitherto referred as IIMB library in this paper) heralded onto a new path of making primary datasets about socio-economic datasets, including the massive census reports in the digital domain. In the year 2014 all National Sample Surveys from the Government of India and Census reports from 1881 to 1941 were digitized for internal circulation purposes only. Though these efforts were a small step towards digital curation, it raised expectations from the user community on the computational potential and data mining abilities of these datasets. But to accomplish it, the challenges of digital perpetuity, technological obsolescence, dissemination expanded to the public, copyright issues have to be overcome.

General Terms

Institutional opportunities

Keywords

Social science data sets social surveys. India

1. INTRODUCTION

Indian Institute of Management Bangalore (IIMB http://www.iimb.ernet.in) is a leading business school located in the southern state of Karnataka in India. IIMB offers a myriad of courses spanning student population in the age group of 20 to 50 years from postgraduate courses & doctoral courses in Management to executive management courses. Its user base primarily comprises the faculty, research scholars, and students, also to substantial number of walk-in users. IIMB Library is predominantly a management resource library with more than 80% of the annual budget spent on digital resources (http://www.iimb.ernet.in/library).

1.2 Objectives

i. **Improve accessibility: T**o make the information available on the Internet. To ensure longevity of data, by digitizing content that are stored and organized in high-density servers, with searchable indexing terms for easy retrieval.
ii. **Preservation:** Preservation of original data for a longer period by the deployment of meticulous preservation techniques to protect data from deteriorating**.**
iii. **Enhance search capability:** Implement a web enabled integrated digital library through which the content can be managed, catalogued and searched.
iv. **Centre for Social Science Data Online:** Create a centre responsible for online dissemination of census data and other social science datasets, for easy and wider access.

2. DIGITAL CURATION

2.1 Digitization and Data Acquisition

IIMB Library had acquired census reports in microfiche format in 1988 and access to these reports were enabled by associated accessories such as microfiche reader and printer. The library committee amidst growing demand from the faculty of Economics and Finance area recommended for the digitization of Census reports with data mining and computational abilities. As a first step towards digital curation, these reports were converted into digital format (PDF Images) in collaboration with International Institute of Population Sciences, Mumbai

The other important datasets that were digitized were National Sample Survey Organization reports published by Ministry of Statistics, Government of India. The **National Sample Survey Organisation (NSSO)**, now known as **National Sample Survey Office** is an organization under the Ministry of Statistics of the Government of India. It is the largest organization in India that conducts regular socio-economic surveys. Digital reports are available from 38th round onwards. However, IIMB has a print collection of earlier rounds, which were also digitized in PDF image format.

2.2 Preservation

The first stage was to preserve and make them accessible in PDF image format; we started exploring for a dedicated technology that would facilitate not only preservation but also disseminate the content. The reports were delivered in CD format; a mirroring technology was chosen to copy the files to the hard disk. Adhering to Dublin Core Schema, Metadata was granulated to include a

chapter level description and make it discoverable. It also allows loading of pre-designed thesaurus unique to each type of collection, thereby allowing assignment of authoritative descriptors only. As of now, all the content is searchable, apart from browsing facility. These files will be converted into a durable format identified by the technical team for long-term preservation.

2.3 Strategies

Currently, we are evaluating from the sample PDF image files that can be converted into a data mining capable format for the tables and textual information in the reports, along with the costs incurred in the computational ability and preservation strategy.

A proposal will be submitted to the sub-committee constituted for this purpose to enable the technical team to evaluate the strategies available in the order of preference along with the cost of preservation.

a. Microfiche: Strategy one: Preserve in Microfiche either at local site or off site with a third-party vendor and digitize again as and when the current formats turn obsolete.

b. Format Migration: Reduce the risk of obsolescence by storing in multiple storage locations and then data is migrated to a new media when it is appropriate. Explore Technical registry services and digital archives projects initiated by Universities Archives to keep abreast of the formats and, software and hardware requirements 'rendering platform' by extracting the technical metadata and their durability. There are number of projects such as PRONOM of UK data archives, Jhove of Harvard University, NLNZ by National Library of New Zealand, COPTR in open planet foundations, PANIC of University of Queensland Center

c. Convert all PDF images in PDF/A-2, Use of ISO 32000-1 (PDF 1.7)

d. Tables from both census reports and NSSO data sets will be in XML from which content can be extracted in multiple formats which includes spreadsheet format. This format is more durable and adaptable to changes especially XML-based mark-up formats, with included or accessible DTD/schema, XSD/XSL presentation stylesheet(s), and explicitly stated character encoding

e. Participate and collaborate in Global archives alliances such as SafeArchive, Data–PASS and Private LOCKSS.

3. COMPUTATIONAL AND DATA MINING NEEDS:

Digitization of census reports exacerbated the demand for making these reports in a format that are downloadable and amenable for maneuvering of the numerical data. In the second phase, only the statistical data in tabular format will be made available using technology that survives obsolescence. Sample files are already being tested and it is possible to convert 90 percent of the content into XML format. We gave a few samples to evaluating the conversion process and deliverables as per our requirement.

Stage one: PDF images are converted into by PDF searchable format a by an OCR

Stage two: An application is used to convert the PDF searchable into XML.

Stage three. From this XML format, a proprietary application was run to get tables in exportable spreadsheet format. Other formats that can be generated from this are epub and HTML

4. CONCLUSIONS:

IIMB's digital curation efforts will move forward to enable the tables of census data and other social science datasets from government sources and publicly funded research projects.

An outline of the above proposal was presented to the Library Committee in the first week of September, 2015 The Committee suggested for a holistic plan to cover all digital assets of the Institute for long term preservation. This includes the Digital Institutional repositories, Electronic Journal subscriptions, Primary research datasets, MOOCs and other video lessons contributed by the faculty. A two-year time line and budgetary resources are sanctioned for the current year to initiate the digitization process for census and NSSO datasets and explore viability of strategies proposed.

ISEC (www.isec.ac.in) has digitized NSSO datasets and propose to use Dataverse Network. IIPS (www.iipsindia.org) digitized in PDF image format and offer the census content as a browsable datasets without computational and data mining capabilities. However, complete set of digitized data is currently available with IIPS and Registrar General and Census Office, Government of India from 1881 to 1991 in PDF image format. But reports of 1991, 2001 and 2011 are completely available in digital format with extractable features.

Census workstation is being set up at IIMB by the Office of Registrar General and Census office that provides access to complete published tables from 1991, 2001 and 2011.

There is a lot of potential to build theme-based collection leading to subject repositories, but our priority right now is to strengthen the socio-economic datasets. Once the data is enabled with computational and data mining abilities, we would like to make this accessible to all research scholars who are interested in this area, and extend this facility to other socio-economic data primary datasets. Preservation efforts will be simultaneously revised and implemented after the committee evaluates the strategies on efficacy and cost of maintenance. A sub-committee constituted for the purpose will evaluate after submission of the proposal for creating infrastructure for digital asset management