Article

Data Repositories in India with Specific Reference to "ICSSR Data Service: Indian Social Science Data Repository"

Pallab Pradhan, Scientist B (LS)

The raw research data / datasets used in the process of research analysis has immense inheritance value as like the research results itself. So, it is very evident and becoming very important to collect, store and manage such data / datasets properly for further reuse. In this regard, data repositories are becoming the trend setters in research data management and playing a great role for facilitating open access to data & promoting data sharing. Currently, hundreds of data repositories are available on the web covering a wide range of disciplines from around the world; and India is not an exception to it. The ICSSR Data Service has been initiated by ICSSR, New Delhi with aim to promote data sharing at a national level and to provide open access to social science research data / datasets hugely generated by various social science research institutes in the country. This article elaborates the current status of data repositories in the country, specifically the "ICSSR Data Service: Indian Social Science Data Repository". Further, it enumerates in details about ICSSR Data Service i.e. its objectives, development, current status, its features & functionalities, datasets, etc. Also, it explores the in-built online "ICSSR Data Analytics" and visualization tool developed using "R" language.

1. Introduction

Thanks to the “Open Access Movement”, the scientific and academic world is witnessing a methodical move from subscription-based access to open access to scientific literatures. From last two decades, a growing and persistent demand is being made across the globe for free and open access to science and scientific research literatures. In perfect alignment to the open access movements, open data movement is getting momentum worldwide. The raw data / datasets associated with a research article possess a great value for reuse and further research. Data is the key to any scientific analysis of research as the research and its results solely depends on the quality and accuracy of data. Open data facilitates free availability, sharing and long-term use of data. According to National Data Sharing and Accessibility Policy (NDSAP, 2014), “A dataset is said to be open if anyone is free to use, reuse, and redistribute it – Open data shall be machine readable and it should also be easily accessible.” Whereas, data sharing is nothing but the practice of making data available to be used for scholarly research to other investigators / researchers. Several funding agencies, institutions, and publishers are coming-up with their well-defined policies for open data and sharing to promote open data and data sharing amongst researchers. .

Generally speaking, data repositories, institutional repositories, and self preservations are being used for storing, preserving, managing and sharing research data and research articles. Institutional repositories are mostly used for storing and sharing the institutional research output i.e. research papers, articles, and theses produced by its own academic community. Whereas, data repositories are used for storing the raw data / datasets that are generated during an experiment or are collected while carrying out a survey. Data repository generically refers to a central place where data is stored, managed and maintained, often for safety or preservation. It can be a place where multiple databases, datasets or files are located for preservation and distribution over a network that is directly accessible to the user. It may include requisite infrastructure often referred as data archives or data centres required for obtaining and depositing data to facilitate further sharing, analysis and reuse. Now-a-days, various data repositories are evolving at international and national level and Institutions to disciplines specific repositories to support the end users / researchers for reuse and advance research. The article elaborates on current scenario of data repositories in India with specific reference to ICSSR Data Service including its genesis, objectives, development, features & functionalities, available datasets, and its current status, etc. At the end, it explores the in-built online ICSSR Data Analytics and visualization tool developed using "R" language.

2. Current Indian Scenario in Data Repositories

As per re3data.org, the registry of data repositories, there are 30 data repositories run by different institutions in India. ("http://service.re3data.org/browse/by-country/") Some of the major Indian data repositories are:

2.1 Open Government Data (OGD) Platform India: The Government of India has committed itself for data sharing through a policy document entitled "The National Data Sharing and Accessibility Policy (NDSAP)" published in the Gazette of India in March 2012 with an aim to share non sensitive data available either in digital or analogue format that is generated using public funds by various ministries / departments / subordinate offices / organizations / agencies of Government of India as well as states. The NDSAP policy is designed to promote data sharing and enable access to Government of India owned data for national planning, development and awareness. Based on the NDSAP Policy, the flagship initiative "data.gov.in" of the Government of India was launched on the Open Data Government (OGD) platform in the year 2012 to act as single point access to all resources (datasets / apps) in open format published by various ministries / departments / organizations of Government of India. (https://data.gov.in/)

2.2 National Data Repository (NDR): In accordance to Petroleum and Natural Gas Amendment Rules 2006, the Directorate General of Hydrocarbons (DGH), a technical arm and nodal agency under Ministry of Petroleum and Natural Gas, Government of India has developed a National Data Repository (NDR) to preserve different kinds of data i.e. oil field data, cultural data, geological data, petro physical data, seismic data, well data, production data, reservoir data, and various unstructured data such as reports, documents, etc. It is hosted at Directorate General of Hydrocarbons (DGH), Noida, Sector - 73 , UP 201301, India. NDR is a fully Government of India owned integrated data repository of Exploration and Production (E&P) data of Indian sedimentary basins. NDR offers an unique platform to all E&P Operators, E&P Service Companies, E&P Investors, Academia to delve inside diverse E&P datasets of Indian sedimentary basins. (https://www.ndrdgh.gov.in/NDR/)

2.3 KRISHI - Knowledge based Resources Information Systems Hub for Innovations in agriculture:

KRISHI is an initiative of Indian Council of Agricultural Research (ICAR) to bring its knowledge resources to all stakeholders at one place. The portal is being developed as a centralized data repository system of ICAR consisting of technology, data generated through experiments/ surveys/ observational studies, geo-spatial data, publications, learning resources etc. (http://krishi.icar.gov.in/)

2.4 ICRISAT Dataverse Network: ICRISAT's data repository collects, preserves and facilitates access to the datasets produced by ICRISAT researchers to all users who are interested in. Currently, the data repository has 11 dataverses (datasets) comprising 75 studies and 549 files. (http://dataverse.icrisat.org/)

2.5 ICSSR Data Service: The "ICSSR Data Service" is culmination of signing of Memorandum of Understanding (MoU) between Indian Council of Social Science Research (ICSSR) and Ministry of Statistics and Programme Implementation (MoSPI). The MoU provides for setting up of "ICSSR Data Service: Social Science Data Repository" and host NSS and ASI datasets generated by MoSPI. The ICSSR Data Service includes social science and statistical datasets of various national level surveys on industries, employment and un employment, household consumer expenditure, enterprise, land holdings survey, census data, etc. into its repository. Currently, the data repository is hosted at Information and Library Network (INFLIBNET) Centre, Gandhinagar, Gujarat which was assigned the task of setting up the data repository. (http://icssrdataservice.in/)

The details about ICSSR Data Service is broadly discussed in detail in this article.

3. ICSSR Data Service

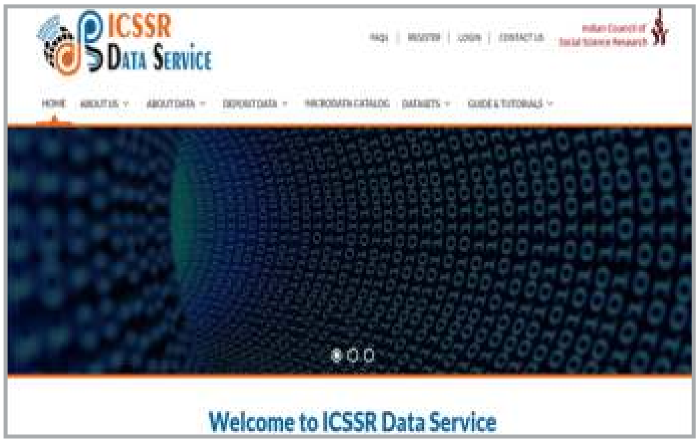
ICSSR Data Service is long term vision of Prof. Sukhdeo Thorat, Chairman, ICSSR, New Delhi. The ICSSR was established in 1969 with the specific objective of promoting socio-economic research in India. The Council is responsible for much of the funding to support and assist research activities with over 27 social science research institutes and 6 regional centres being directly under its purview.

A research commissioned by the ICSSR found that 417 institutions across India are involved in doing social science research including over 230 universities, 51 Institutes of National Importance and numerous autonomous research institutes. All these institutes are generating huge amount of social science research data. Also, ministries like Ministry of Statistics and Programme Implementation (MoSPI), and Ministry of Home Affairs, Government of India are generating abundance of data of interest to the social scientists from national surveys like National Sample Survey (NSS), Annual Survey of Industries (ASI), and Census data respectively. But, making data available for research is an ongoing challenge in India, with open access an exception; Government departments are highly cautious to grant access to data; and private sector data, wherever accessible, is very expensive for researchers.

To overcome these issues and with a vision to promote and facilitate open access to social science research data in the country, Prof. Thorat met Dr. Matthew Woollard, Director, UK Data Service in 2013, and discussed a blueprint for development of Social Science data service in India. A road map and implementation plan for ICSSR Data Service was produced in 2014, outlining the proposed information architecture for the data repository, setting out key components for data acquisition, data pre-processing and analytics, metadata, software requirements and technology platforms; providing an excellent starting point for developing a data service for India. Senior representatives of the ICSSR and other organisations interested in implementation of the ICSSR Data Service visited the UK Data Archive at the University of Essex, a partner in the UK Data Service, in January 2015, for a two-day training workshop (Moody, 2016).

Later, a Joint Advisory Committee was constituted by the ICSSR with representatives from ICSSR, MoSPI, other key agencies and academic social science research centres / institutions across India for setting-up ICSSR Data Service and to guide the process development. The task for developing, hosting and maintaining the data repository along with all related activities was assigned to the Information and Library Network (INFLIBNET) Centre, Gandhinagar as an ICSSR-sponsored project with initial funding from MoSPI and ICSSR. NADA, an open source software been used to built the data repository and customized extensively as per the requirements of social science researchers. The ICSSR Data Service was launched formally on 20th June, 2016 by Dr. T. C. A. Anant, Secretary and Chief Statistician of India, MoSPI at ICSSR, New Delhi.

Figure 1 is the screenshot of the "Home" page of ICSSR Data Service.



Welcome to ICSSR Data Service

Figure 1: "Home" Page of ICSSR Data Service

3.1 Objectives

The first and foremost objective of this portal is to provide seamless and integrated access to a wide range of datasets generated by the MoSPI, New Delhi, social science institutions under direct purview of ICSSR and other Government organizations, to researchers who are looking for high quality social and economic research datasets with following aims and objectives.

- To serve as a national data service for promoting powerful research environment through sharing and reuse of data among social science community in India;
- To acquire, process, organize, preserve and host research data along with its metadata with ETL (extract, transform and load) facilities of raw data in social sciences and related domains collected from diverse sources for easy sharing and access;
- To facilitate online submission, access, search, browse, discovery, conversion, analysis and

visualization of data through intuitive interfaces;

- To impart training and spread awareness about benefits of data sharing and reuse amongst social science research community in India; and
- Interact, cooperate and collaborate with other national and international data services and repositories for data and resource sharing and improved management of data services.

3.2 Stakeholders

Built on the basis of participatory approach, the ICSSR Data Service is a national-level social science data repository service set-up to facilitate data sharing and open access to social sciences data collected from various social science communities in the country. Any persons or institutions are welcome to contribute their secondary social science research data voluntarily or to use the data available data in the repository.

Major stakeholders of the ICSSR Data Service are:

- Ministry of Statistics and Programme Implementation (MoSPI) and its two organs, namely National Sample Survey Organization and Central Statistics Office (CSO);
- ICSSR: Indian Council of Social Science Research (ICSSR) and its 27 constituent research centres located across the country;
- Other ministries, Govt. Departments, and policymakers as users as well as contributors;
- Students, researchers and faculty as users as well contributors;
- Working professionals and NGOs as users as well as contributors;
- Universities and colleges as organizations that use and contribute to ICSSR Data Service and define policies on data generation and its delivery;
- Foreign Users: Students, research scholars, scientists, and faculty members from institutions abroad with which ICSSR has bilateral understandings / agreements on sharing of resources subject to the condition that similar facilities will be reciprocated by such institutions with respect to resources held by them; and
- Any others individual/independent researchers, government organizations, private firms, NGOs and any other institutions working on social sciences and related domains who wishes to deposit their data into the repository by complying to the data repository policies.

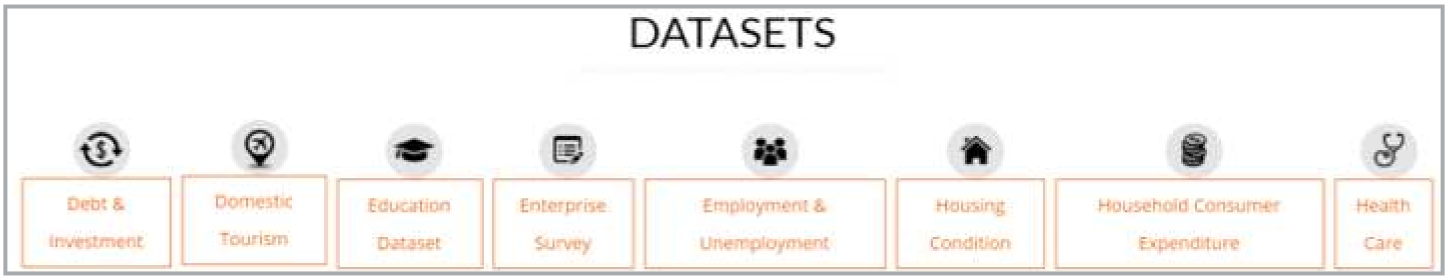
3.3 Current Status, Datasets and Access

Currently, ICSSR Data Service hosts 131 datasets (NSS and ASI datasets) provided by the MoSPI under the agreement in between the MOSPI and ICSSR. These datasets have been extracted, transformed and uploaded with details metadata into the repository. Also, the supporting documents, i.e. questionnaires, data collection methods, codebooks, and project summaries / descriptions, etc. are available on the repository along with their associated datasets. Further, it is proposed to expand the scope of ICSSR Data Service to include datasets from all social science institutions which are under direct purview of ICSSR, other social science research centres, NGOs and individuals researchers, others academic institutions as well as government agencies.

The ICSSR Data Service supports almost all preferred machine readable data formats for hosting into the repository. It mainly considers and accepts following kinds of datasets for inclusion in the repository:

- raw or preliminary data;
- data that are ready to use and ready for full release;
- unit level summary data; and
- tabulated, analyzed and derived data, etc.

The ICSSR Data Service employs DDI XML based descriptive metadata schema for assigning descriptive metadata to datasets deposited into the repository. Also, the ICSSR Data Service uses Humanities and Social Science Electronic Thesaurus (HASSET) developed by UK Data Service to assign keywords to the data / datasets which was granted to ICSSR Data Service under a licence in July 2015 by UK Data Service. All the datasets available in the ICSSR Data Service are categorized in eight categories / collections as defined by the MoSPI which are depicted in the Figure 2 given below.

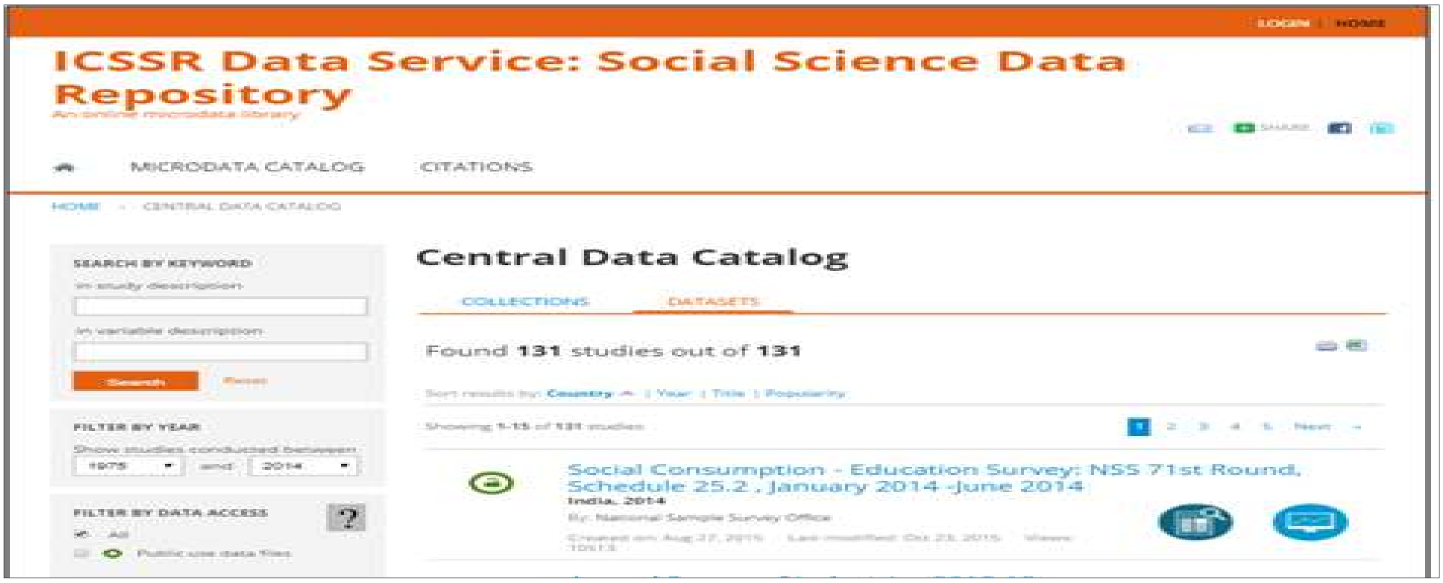


DATASETS

Debt & Investment | Domestic Tourism | Education Dataset | Enterprise Survey | Employment & Unemployment | Housing Condition | Household Consumer Expenditure | Health Care

Figure 2: Categorization / Collections of Datasets in ICSSR Data Service

All the datasets are available under “Microdata Catalog” in Central Data Catalog which is shown in the screenshot Figure 3 given below.



LOGIN | HOME

ICSSR Data Service: Social Science Data Repository

An online microdata library

MICRODATA CATALOG CITATIONS

HOME › CENTRAL DATA CATALOG

SEARCH BY KEYWORD

in study description

in variable description

Search Reset

FILTER BY YEAR

Show studies conducted between 1975 and 2014

FILTER BY DATA ACCESS

All

Public use data files

Central Data Catalog

COLLECTIONS DATASETS

Found 131 studies out of 131

Sort results by: Country | Year | Title | Popularity

Showing 1-15 of 131 studies

1 2 3 4 5 Next →

Social Consumption - Education Survey: NSS 71st Round, Schedule 25.2 , January 2014 -June 2014

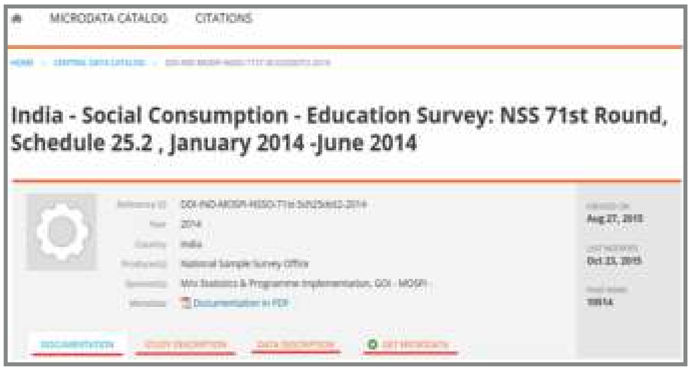
India, 2014

By: National Sample Survey Office

Created on: Aug 27, 2015 Last modified: Oct 23, 2015 Views: 10613

Figure 3: Display of Datasets in "Microdata Catalog"

Further, each dataset available in the repository is organized in following four tabs, i.e. i) Documentation, ii) Study Description, iii) Data Description, and iv) Get Microdata as depicted in below Figure 4 given below. Here, user can view / get details about the supporting documents, study, and data available in a specific dataset. User can get access or download the dataset in “Get Microdata” tab.



MICRODATA CATALOG CITATIONS

India - Social Consumption - Education Survey: NSS 71st Round, Schedule 25.2 , January 2014 -June 2014

Reference ID DDI-IND-MOSPI-NSSO-71st-Sch25Sch2-2014

Year 2014

Country India

Producer(s) National Sample Survey Office

Sponsor(s) M/o Statistics & Programme Implementation, GOI - MOSPI

Metadata Documentation in PDF

Created on Aug 27, 2015

Last modified Oct 23, 2015

Page views 10614

DOCUMENTATION STUDY DESCRIPTION DATA DESCRIPTION GET MICRODATA

Figure 4: Organization of a Dataset in ICCSR Data Service

ICSSR Data Service follows a stringent process to accept or reject data / dataset before its submission. The data is internally reviewed and evaluated by the data experts from ICSSR Data Service on the basis of the eligibility of the depositor, relevance to the scope of the collections, valid data formats, etc. Further, the data goes for an internal data quality check to ensure that the proper quality standards are maintained, i.e. accuracy, consistency, documentation, metadata, free from any sort of legal issues, privacy of individuals are maintained and does not compromise with the national security.

All the datasets, metadata information and their supporting materials available in the ICSSR Data Service can be searched, viewed, accessed and used freely for further study, teaching and research. The available datasets in the ICSSR Data Service are not for profit making use. In few circumstances, access to some specific data may be restricted. Basically, three types of access restrictions or access control mechanism are used on the datasets:

- Open Data: Access to open data which generally means data generated from public funding and meant for public should be freely available in open access without any access restrictions.
- Safeguarded Data: Data can be access only to the registered users with proper registration and authorization by the ICSSR Data Service. The user agree to the terms and conditions of data usage policy displayed to them at the time of request.
- Controlled Data: Access to these kinds of data would be controlled through Secured Lab, stored in a secured server. Sensitive and highly confidential data, as declared by Government of India policies will be accessible only through this mode.

3.4 Features and Functionalities

As mentioned before, the data repository of ICSSR Data Service is built on NADA, an open source software platform. The ICSSR Data Service offers following features and functionalities:

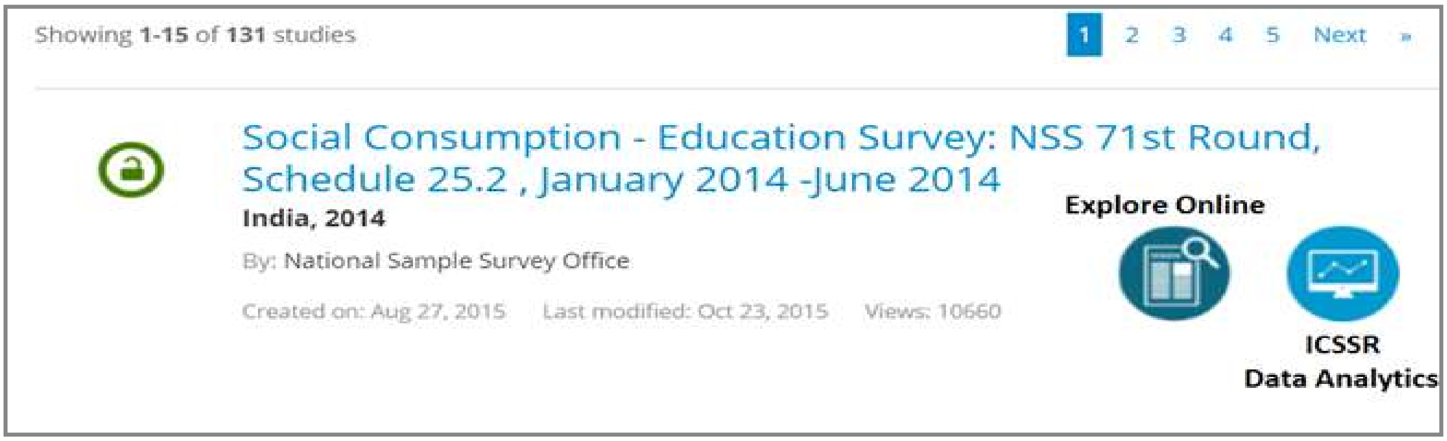
- Supports search and discovery through elaborate metadata description of each datasets using Data Documentation Initiative (DDI) Metadata Standard (DDI-MS);
- Provides raw as well as transformed data in multiple formats;
- Provides options for online analysis through inbuilt data analytical tool by choosing multiple variables;
- Generates cross tabulation for various datasets;
- Provides for data visualization through bar charts, line and scatter diagrams, pie charts, stacked charts, histogram, etc;
- Supports multiple output formats such as CSV, TSV, PDF, EXCEL, DTA, SAV, etc;
- Multiple search and browse options; and
- Visualization of geo coded data on maps available online.

3.5 Data Analysis and Exploration: “Explore Online” and “ICSSR Data Analytic Tool”

Two options namely “Explore Online” and “ICSSR Data Analytic Tool” are in-built in ICSSR Data Service to facilitate exploration, analysis and visualization of dataset.

3.5.1 Explore Online: User can explore a dataset online through “Explore Online” option or can analyse and visualize the data using “ICSSR Data Analytic Tool”. Figure 5 given below shows the above mentioned two options alongside a dataset for example.

- Controlled Data: Access to these kinds of data would be controlled through Secured Lab, stored in a secured server. Sensitive and highly confidential data, as declared by Government of India policies will be accessible only through this mode.



Showing 1-15 of 131 studies

1 2 3 4 5 Next »

Social Consumption - Education Survey: NSS 71st Round, Schedule 25.2 , January 2014 -June 2014

India, 2014

By: National Sample Survey Office

Created on: Aug 27, 2015 Last modified: Oct 23, 2015 Views: 10660

Figure 5: Display of "Explore Online" and "ICCSR Data Analytic Tool" Options

In “Explore Online”, user can visualize the results as charts and tables as shown below in Figure 6 to 8 as examples. Further, user have the option to select pre-derived and pre-selected data from the drop-down menu / list available to generate charts and tables

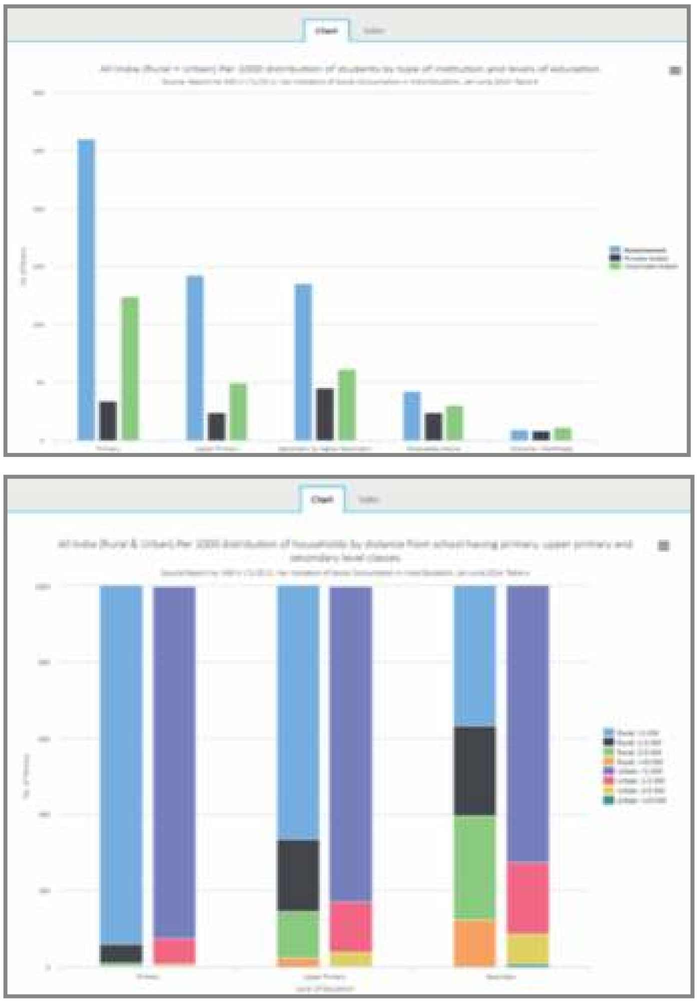


Figure 6 and 7: Display of Charts

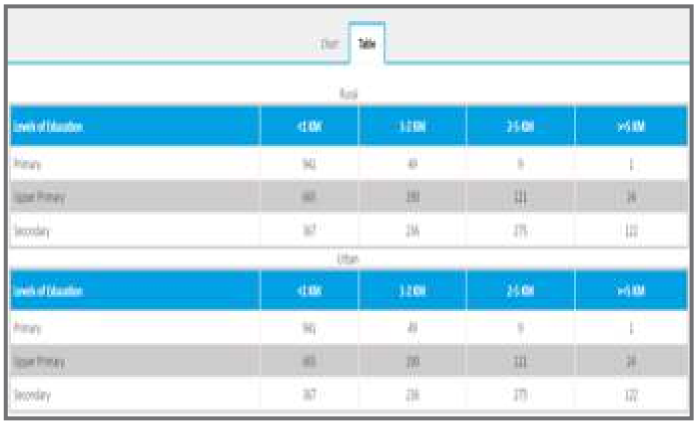


Chart
Table

Rural

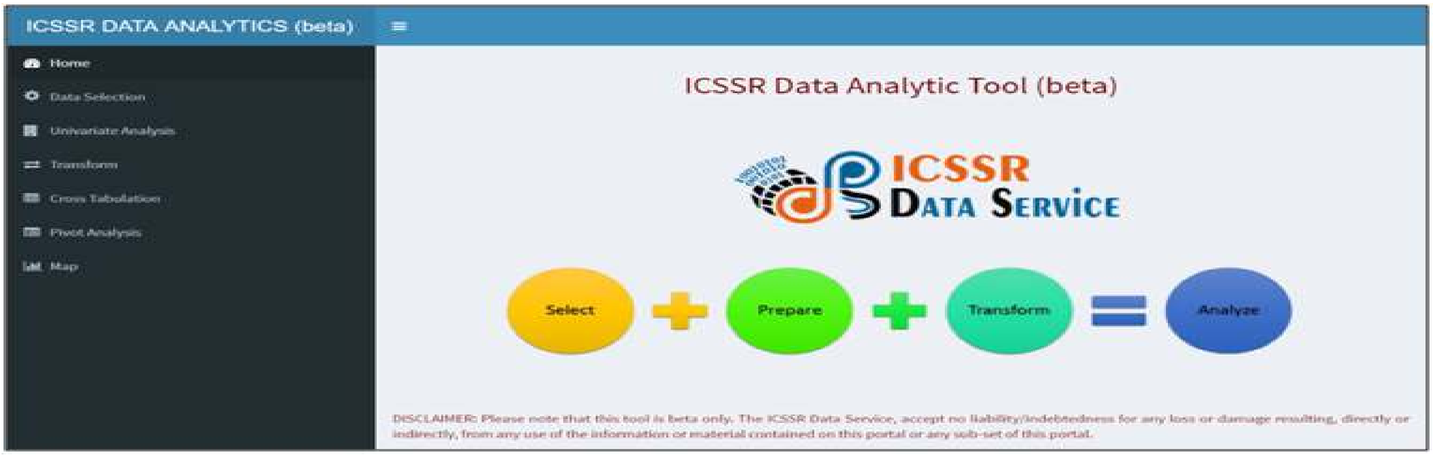
Levels of Education	<1 KM	1-2 KM	2-5 KM	>5 KM
Primary	941	49	9	1
Upper Primary	665	193	121	24
Secondary	367	236	275	122

Urban

Levels of Education	<1 KM	1-2 KM	2-5 KM	>5 KM
Primary	941	49	9	1
Upper Primary	665	200	111	24
Secondary	367	236	275	122

Figure 8: Display of Table

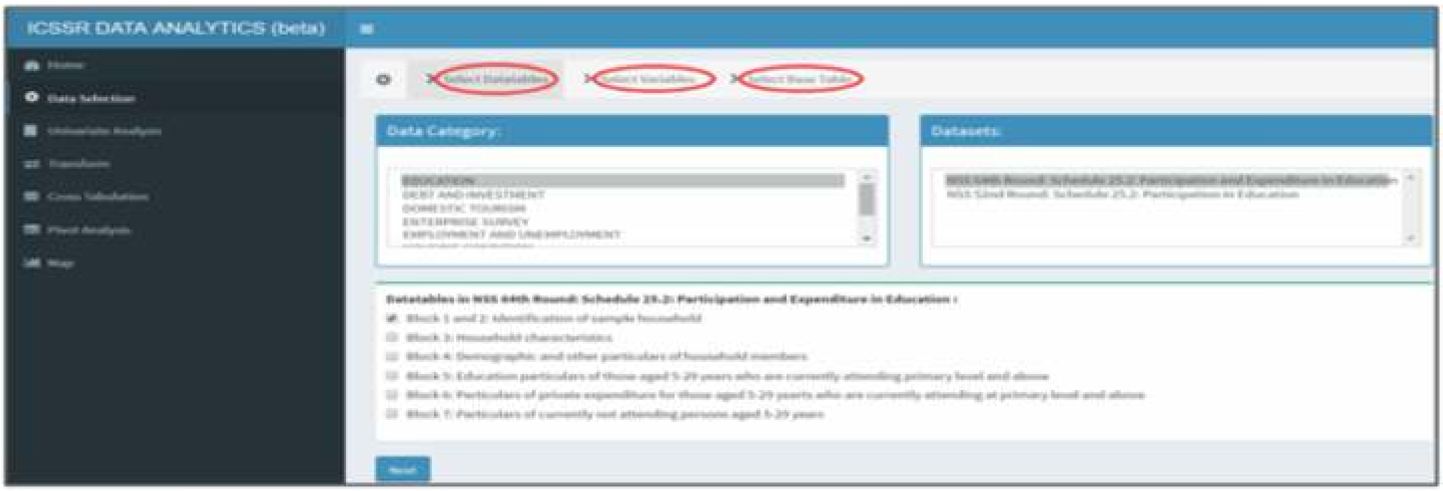
3.5.2 ICSSR Data Analytic Tool: The ICSSR Data Analytic Tool was developed in "R" language for advance analysis and visualization of datasets available in the ICSSR Data Service. Figure 9 given below is the screenshot of the dashboard of ICSSR Data Analytic Tool (beta).



DISCLAIMER: Please note that this tool is beta only. The ICSSR Data Service, accept no liability/indebtedness for any loss or damage resulting, directly or indirectly, from any use of the information or material contained on this portal or any sub-set of this portal.

Figure 9: Dashboard of ICSSR Data Analytic Tool (beta)

User can perform a number of analyses by using ICSSR Data Analytics such as: univariate analysis, data transformation, cross tabulation, pivot analysis and for generation of various maps and charts, etc. To perform any analysis in the data analytic tool, user has to first select the required data tables, variables, and base table by clicking the "Data Selection" tab available on the dashboard as shown in Figure 10.



ICSSR DATA ANALYTICS (beta)

Home
Data Selection
Univariate Analysis
Transform
Cross Tabulation
Pivot Analysis
Map

Select Datatables | Select Variables | Select Base Table

Data Category:

EDUCATION
DEBT AND INVESTMENT
DOMESTIC TOURISM
ENTERPRISE SURVEY
EMPLOYMENT AND UNEMPLOYMENT

Datasets:

NSS 64th Round: Schedule 25.2: Participation and Expenditure in Education
NSS 52nd Round: Schedule 25.2: Participation in Education

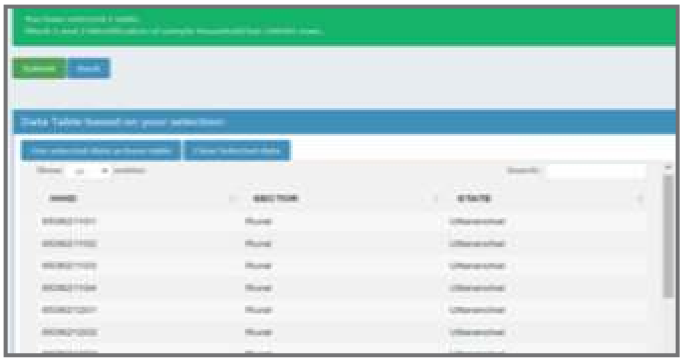
Datatables in NSS 64th Round: Schedule 25.2: Participation and Expenditure in Education :

- ☑ Block 1 and 2: Identification of sample household
- ☐ Block 3: Household characteristics
- ☐ Block 4: Demographic and other particulars of household members
- ☐ Block 5: Education particulars of those aged 5-29 years who are currently attending primary level and above
- ☐ Block 6: Particulars of private expenditure for those aged 5-29 years who are currently attending at primary level and above
- ☐ Block 7: Particulars of currently not attending persons aged 5-29 years

Next

Figure 10: "Data Selection" Tab on the Dashboard of ICSSR Data Analytics

Selection of data tables and variables is shown in below in Figure 11 and 12 as examples. Subsequently, the result of univariate analysis from selected data tables and variables is depicted in given bar chart below and pie chart as shown in Figure 13 and 14 respectively.

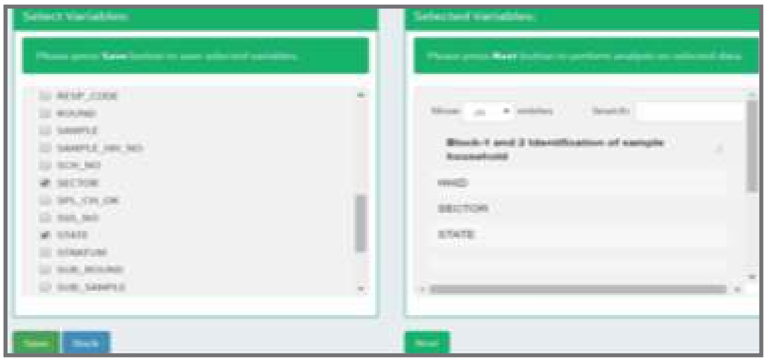


You have selected 1 table.
Block 1 and 2 Identification of sample household has [illegible] rows.

Submit | Back

Data Table based on your selection:

HHID	SECTOR	STATE
[illegible]	Rural	Uttaranchal
[illegible]	Rural	Uttaranchal
[illegible]	Rural	Uttaranchal
[illegible]	Rural	Uttaranchal
[illegible]	Rural	Uttaranchal
[illegible]	Rural	Uttaranchal



Select Variables:

Please press **Save** button to save selected variables.

- ☐ RESP_CODE
- ☐ ROUND
- ☐ SAMPLE
- ☐ SAMPLE_HH_NO
- ☐ SCH_NO
- ☑ SECTOR
- ☐ SPL_CH_OK
- ☐ SSS_NO
- ☑ STATE
- ☐ STRATUM
- ☐ SUB_ROUND
- ☐ SUB_SAMPLE

Save | Back

Selected Variables:

Please press **Next** button to perform analysis on selected data.

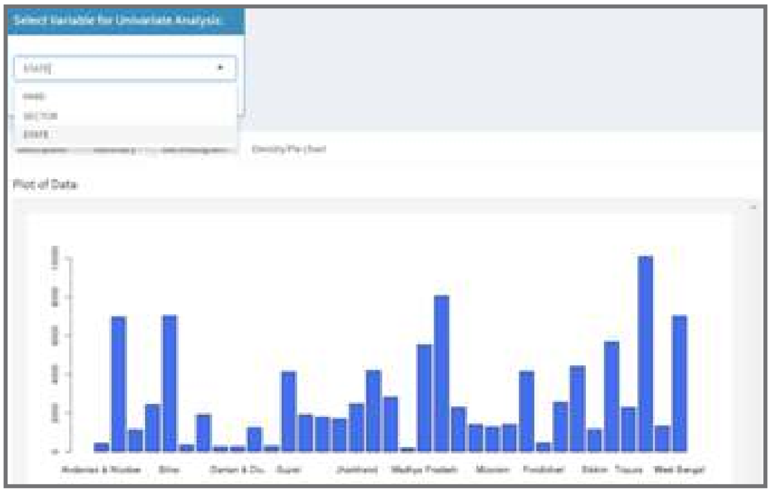
Block-1 and 2 Identification of sample household

HHID
SECTOR
STATE

Next

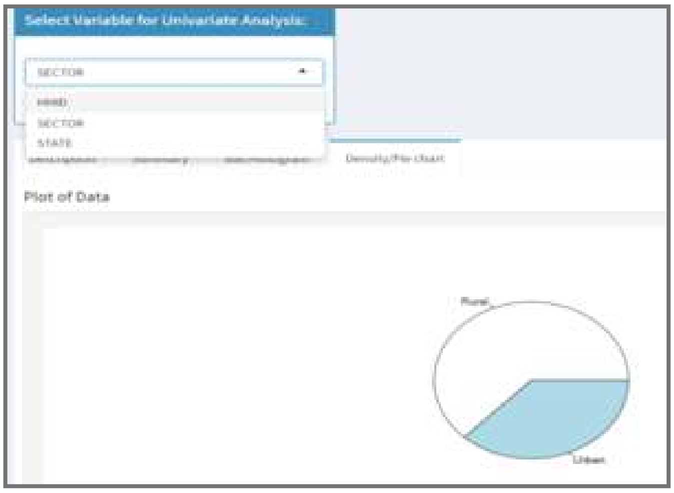
Figure 11: Selection of "Datatables"

Figure 12: Selection of "Variables"



Select Variable for Univariate Analysis:

STATE



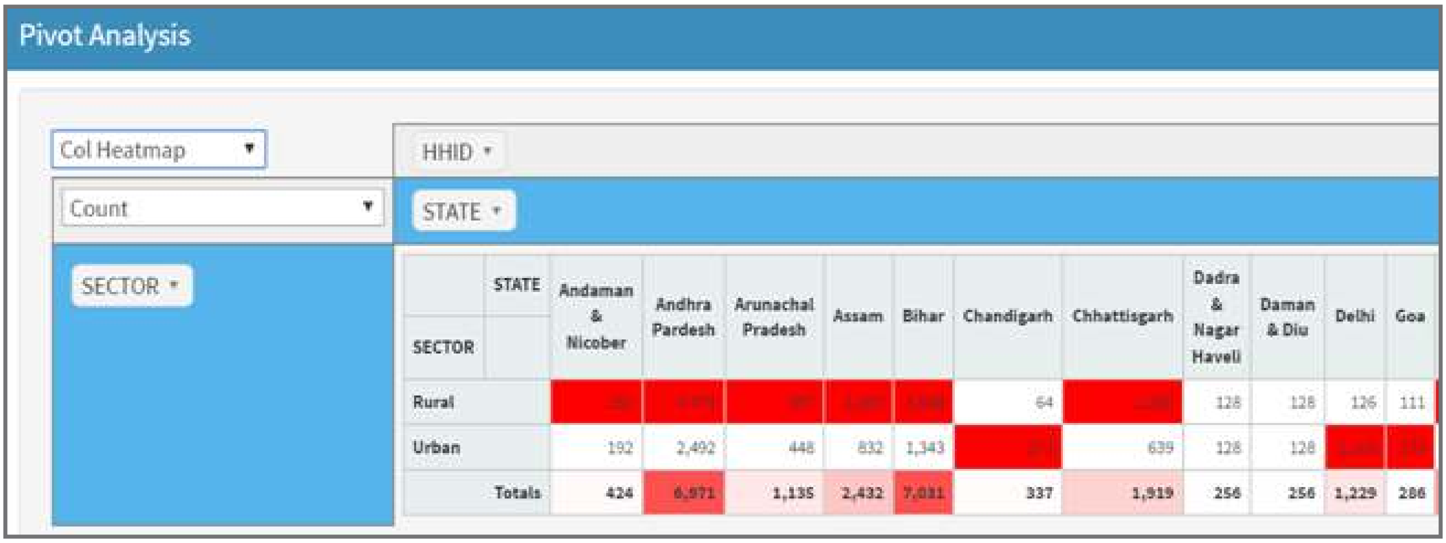
Select Variable for Univariate Analysis:

SECTOR

Figure 13: Univariate Analysis Result as Bar Chart

Figure14: Univariate Analysis Result as Pie Chart

Further, from the pre-selected data tables and variables, pivot analysis has been performed in ICSSR Data Analytics. Results are depicted as fire table, area chart and stacked bar chart which are shown below in Figure 15, 16 and 17 respectively.



Pivot Analysis

Col Heatmap ▾

Count ▾

HHID ▾

STATE ▾

SECTOR ▾

	STATE	Andaman & Nicober	Andhra Pardesh	Arunachal Pradesh	Assam	Bihar	Chandigarh	Chhattisgarh	Dadra & Nagar Haveli	Daman & Diu	Delhi	Goa
SECTOR												
Rural		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	64	[illegible]	128	128	126	111
Urban		192	2,492	448	832	1,343	[illegible]	639	128	128	[illegible]	[illegible]
	Totals	424	6,971	1,135	2,432	7,031	337	1,919	256	256	1,229	286

Figure 15: Pivot Analysis Result as "Fire Table"

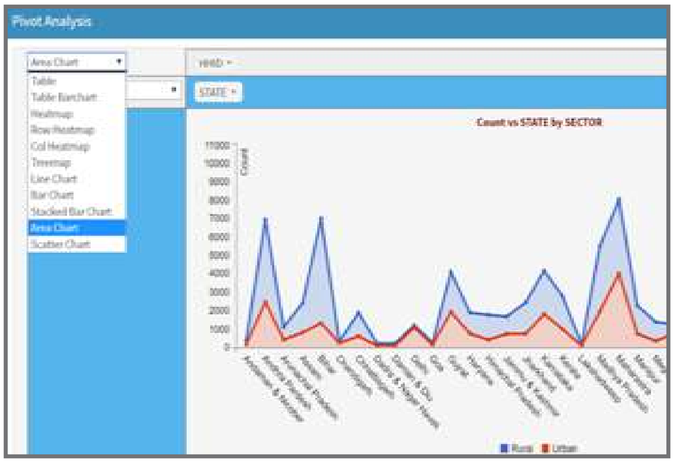


Figure 16: Pivot Analysis Result as Area Chart

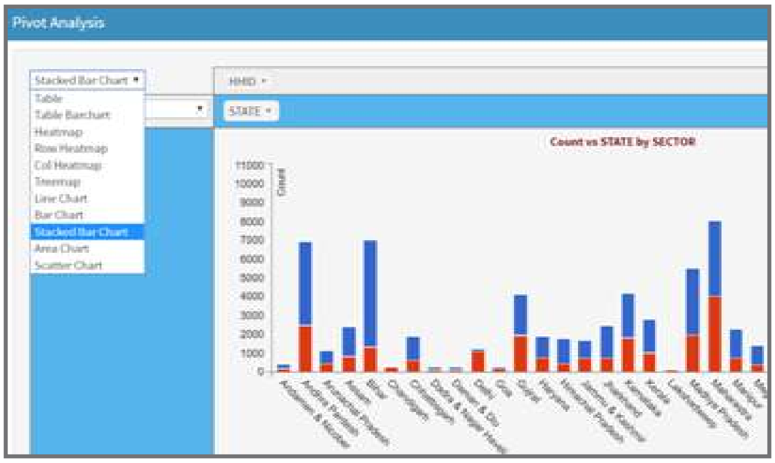


Figure 17: Pivot Analysis Result as Stacked Bar Chart

Similar to the examples given above, user can perform a number of analyses and visualize its results in different formats and types of tables, charts, etc. using ICSSR Data Analytic Tool.

4. Conclusion

The ICSSR Data Service is one of the foremost and an emerging social science data repository in India. It is a great endeavour by ICSSR to share social science research data collected or developed through public funding from various government agencies, institutions and social science research centres. As a policy, the ICSSR Data Service promotes data sharing to encourage reuse of data and provide information on developing and generating social science research data and its management.

References

I. Moody, V. (2016, June 21). Welcoming progress on the new Indian Council for Social Science Research (ICCSR) Data Service [Web log post]. Retrieved from http://blog.ukdataservice.ac.uk/welcoming-progress-on-the-new-indian-councl-for-social-science-research-iccsr-data-service/

ii. National Informatics Centre (NIC), Government of India. (2014). Implementation Guidelines for National Data Sharing and Accessibility Policy (NDSAP) (Ver. 2.2). https://data.gov.in/sites/default/files/NDSAP_Implementation_Guidelines_2.2.pdf

iii. http://dataverse.icrisat.org/

iv. http//www.icssrdataservice.in

v. http://krishi.icar.gov.in/

vi. https://www.ndrdgh.gov.in/NDR/