A Survey of Indian Open Data

Sweety Agrawal
IIIT-Bangalore,26/C,
Electronics City
Hosur Road, Bangalore
sweety.v.agrawal@iiitb.org

Jayati Deshmukh
IIIT-Bangalore,26/C,
Electronics City
Hosur Road, Bangalore
jayati.deshmukh@iiitb.org

Srinath Srinivasa
IIIT-Bangalore,26/C,
Electronics City
Hosur Road, Bangalore
sri@iiitb.ac.in

Chinmay Jog
IIIT-Bangalore,26/C,
Electronics City
Hosur Road, Bangalore
jog.chinmay@iiitb.org

KS Bhavaani
IIIT-Bangalore,26/C,
Electronics City
Hosur Road, Bangalore
Srisayi.Bhavani@iiitb.org

Rahul Dhek
IIIT-Bangalore,26/C,
Electronics City
Hosur Road, Bangalore
rahulsingh.dhek@iiitb.org

ABSTRACT

Publishing structured datasets[1] openly, called "open data" is gaining momentum worldwide. Government of India has started an open data initiative with the launch of its own portal[2]. In addition, open data in India is also published by private firms, research institutes and NGOs (Non - Governmental Organizations). Although structured, the datasets do not comply to any a priori defined schema and semantic elements are fragmented across the disparate datasets. We are working on a project in which we aim to semantically aggregate Indian open data. This will help users to browse open data related to an entity and get a one stop entry to all the aspects of that entity. In this paper we discuss a survey conducted as part of this project to get a fair idea of the open data scene in India. This survey helped us to categorize the open data which makes the modeling easier.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Data Sharing

General Terms

Survey

Keywords

Indian Open data, Datasets, Survey

1. INTRODUCTION

Lots of structured datasets collectively called, *open data* are available online for public use. The UN and several governments have dedicated websites to publish open data for public use. Government of India has also started to publish open data on its own data portal called data.gov.in. Apart from the government, significant amounts of open data is being published by private firms and NGOs. Other than this, unstructured data is also available in the form of survey reports, blogs, news articles, etc. which also counts as open data. Some example sites hosting open data in different domains include: agriculture[3], education[4], astronomy[5], election[6], finance[7], etc.

While large amounts of data are publicly available, these datasets do not conform to any well known, a priori agreed upon schema. Also, data pertinent to a given entity (for instance, data relevant to a city) is fragmented across disparate datasets. This makes it tedious to get a consolidated understanding of any entity of interest. In addition, the datasets are about no one domain – making it impractical to force fit them to any specific publicly available ontology.

While we found some related survey on Open Data, they did not specifically target open data of India, and were very domain specific: MIDAS [1] and GovWild [2]. The MIDAS project surveys and integrates financial data over several data sources. GovWild integrates several US and EU public sector data sources and weaves them with NYT data sources. The authors of [3] surveys Open Data worldwide but it does not talk much about Indian Open Data. All the Open Government Data repository considered by them is available at this website[8]. During that period, only 19 datasets were published on `data.gov.in`.

At the Open Systems Lab[9] in IIIT Bangalore, we are working on a project named Sandesh (acronym for *SemANtic Data mESH*), where we aim to provide an underlying infrastructure using which, different semantic structures can be woven over open data. In order to build this model, a better understanding of the open data available was required. This survey was conducted in this backdrop, and its results are presented here.

[1]Collection of data in tabular form
[2]http://www.data.gov.in

[3]http://agricoop.nic.in/Agristatistics.htm
[4]http://www.dise.in/index.htm
[5]http://www.packolkata.org/
[6]http://eci.nic.in/eci_main1/ElectionStatistics.aspx
[7]http://dbie.rbi.org.in/DBIE/dbie.rbi?site=statistics
[8]http://wwwdb.inf.tu-dresden.de/opendatasurvey/
[9]http://osl.iiitb.ac.in/dokuwiki/

I-CARE 2013 2013 IIT Delhi and IIIT-Delhi, Delhi, India

2. SURVEY STRATEGIES

We commenced the survey by populating the URLs of the websites that publish data online into a crawler. This gave us a fair idea of the data sources, the kind of datasets they host and also their licensing terms. We then manually looked into the datasets available on these websites to gather information regarding these datasets. Filenames of some of the datasets highlighted the topic to which the dataset was related. But, we realized that only looking into the filenames and headers of the datasets did not provide sufficient information regarding the content of the datasets. For such datasets, we gathered information by looking at the data inside the datasets.

3. ANALYSIS OF OPEN DATA

3.1 Initial Survey

For the initial survey, we populated more than 100 website URLs that hosted data pertaining to India. We looked at all the datasets provided by these websites manually. We targeted the survey to gather statistics pertaining to licensing terms, geographical distribution, time granularity and update frequencies of the provided datasets.

During the survey of licensing terms, we observed that a few of the websites had not mentioned licensing and privacy terms for the datasets that they published. We have assumed these datasets to be not open and discarded them from further study. For geographical distribution survey, we gathered statistics based on the geographical divisions of India. We looked for columns having mention of geotags, cities, districts, states and so on. For time granularity, we searched for columns that described temporal metrics like hour, day, month and year. We also tried to categorize websites by tagging them on the basis of the information that they provided.

This survey could not be automated because not all of the datasets published were in a machine-readable format. Some datasets were published as PDF files, some were published as Word documents. Some were Excel sheets while others were CSV files.

3.2 Detailed Survey

In order to get a deeper insight into the contents of open data, we downloaded around 2500 datasets from Indian Data portal[10] on July 10, 2013. The files which we collected from the Indian data portal were present in different file formats including XLS, HTML, CSV, etc. Aggregating data across different formats is not feasible. It indicates a need to standardize data formats for open data. We converted all the datasets into CSV format to simplify the analysis process.

Generating a heat map

We extracted the file names of the dataset and stored them in a CSV file. We looked for states of India in each CSV file to get the frequency of occurrence of states in the datasets. With this frequency distribution, we generated a heat-map (shown in the Figure 7) using a freely available colour-coding scheme[11]. The transition from green to red shows the decrease in the number of times a state was featured in the datasets.

[10] data.gov.in

[11] indzara.blogspot.in/2013/04/geoheatmapindia.html

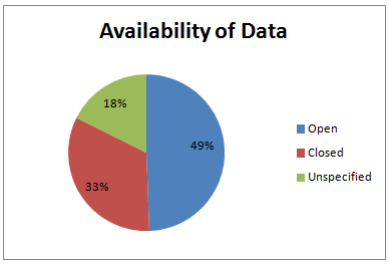

Figure 1: Availability of Open Datasets

Extracting file names with headers and their count across all the files

We generated a table containing filenames, terms in the header and the count of that term across the files with similar file names (for example, Wheat_2007.csv, Wheat_2008.csv, etc).

For each term in the dataset, we constructed a graph where the term is the central node connected to other terms which occur with the respective term across the datasets. We assigned a weight to connected edge between the nodes which is equal to the number of times both the term occur together. We constructed graphs for first 700 most occurring terms. The maximum weight of a node in a graph implies the frequency of term across all the datasets.

4. SURVEY RESULTS

4.1 Results of Initial survey

Availability of Open Datasets

Figure 1 shows the open, closed and unspecified distribution of the datasets. We found that around 49% of the websites (the URLs we collected that contained structured datasets) provided datasets with clearly specified open licensing. It means that the data could be further modified and used by people for any commercial/non-commercial use. 33% of the websites provided data along with strict terms and conditions and privacy policies. These datasets were termed "closed" and were not considered in this survey. The remaining 18% of the websites did not specify the terms of use and we have termed then "unspecified".

Geospatial Granularity of Datasets

Figure 2 shows the geographical granularity of datasets on the basis of geospatial tags associated with the datasets. We found that most of the datasets have a city-wise geotag (around 41%). Fewer datasets were available with a state-wise and country-wise granularity. Even fewer datasets (mostly related to weather) were available at the district level.

Format of Datasets

Figure 3 depicts the distribution of the datasets over the formats of the datasets provided by various sources. Firstly, we observed that the datasets are available in a different formats. At the time of the survey, most of the datasets were found to be in PDF, XLS, HTML and CSV formats.

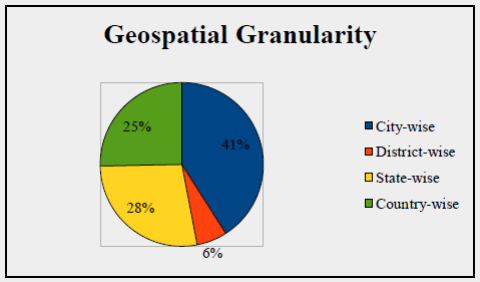

Figure 2: Geospatial Granularity of Datasets

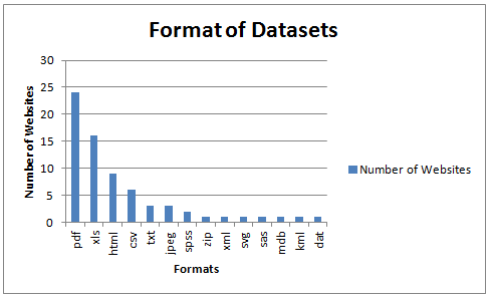

Figure 3: Format of Datasets

Classification of Open Data

The datasets can be classified into 7 different categories. Figure 4 shows the distribution of datasets in seven categories. Any dataset can be associated with at least one of the identified categories.

Distribution of Datasets based on Time granularity

Figure 5 shows the granularity based on time. It shows the updation frequency of different datasets i.e., year-wise, month-wise, etc. We observed that most of the datasets are available on a five-yearly and yearly basis. A few datasets are also present on monthly and quarterly basis. It shows that the information of some datasets is long lasting (example, datasets related to economy) while some datasets are updated more frequently (example, datasets related to weather).

4.2 Results of Detailed Survey

The detailed survey went into the contents of the data to look for commonly occurring entity names and their distribution.

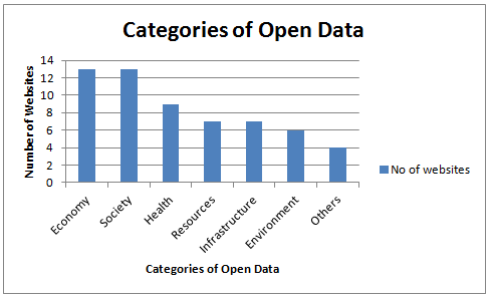

Figure 4: Classification of Open Data

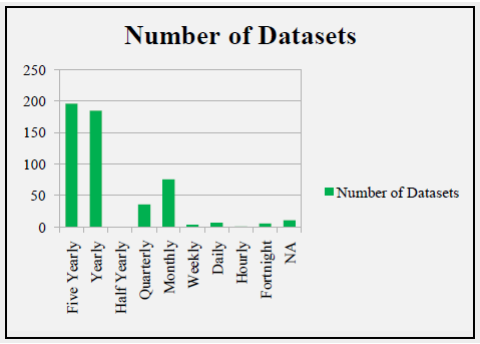

Figure 5: Distribution of Datasets based on Time granularity

State-wise distribution of open data

Here, the aim was to identify states and union territories publishing data in varying amounts. The results showed that some states feature more than the rest in the 2500 datasets that were analyzed. We found that Gujarat and Karnataka are the top-most states mentioned in 1461 and 1371 datasets respectively. Chattisgarh and Uttar Pradesh were the lowest with 80 and 107 datasets respectively. Among the union territories, Chandigarh was the highest with 208 datasets and Dadara & Nagar Haveli and Lakshawadeep were mentioned in zero datasets. Figure 7 shows the heatmap of the open datasets distribution across all the states and union territories of India.

Figure 6 shows the bar graph of the distribution of datasets across all the states and union territories of India. All the datasets which did not have a reference of any state or union territory of India have been tagged India. It is our assumption that a dataset without a reference to any state or UT generically refers to India as a whole.

5. OBSERVATIONS

We observed some points after doing the analysis of the datasets. Most nouns representing entity names are present in the name of the dataset and attributes in column names, specially in agricultural datasets. However, there is a significant problem of de-duplication. There was ambiguity in column names, for example the header for state was named "State" in some datasets and "StateName" in others, months were named in multiple ways like "January" and "Jan."

Some columns contained multiple types of values and the header was also named accordingly, for example there was a column "State/Year" which contained state names or year values in the same column. Term "Variety" was the most commonly occurring header name across all the datasets. "State" and "Commodity" co-occurred maximum number of times.

Almost one-third of the datasets were related to agriculture and animal husbandry. The column distribution across the datasets is sparse and follows a Zipf's law. Two column tuples are also sparsely distributed.

6. CONCLUSIONS AND FUTURE WORK

From this survey, we concluded that the semantic mesh we aim to create will not be dense. The mesh will have sparse central nodes (column names of the datasets) using which

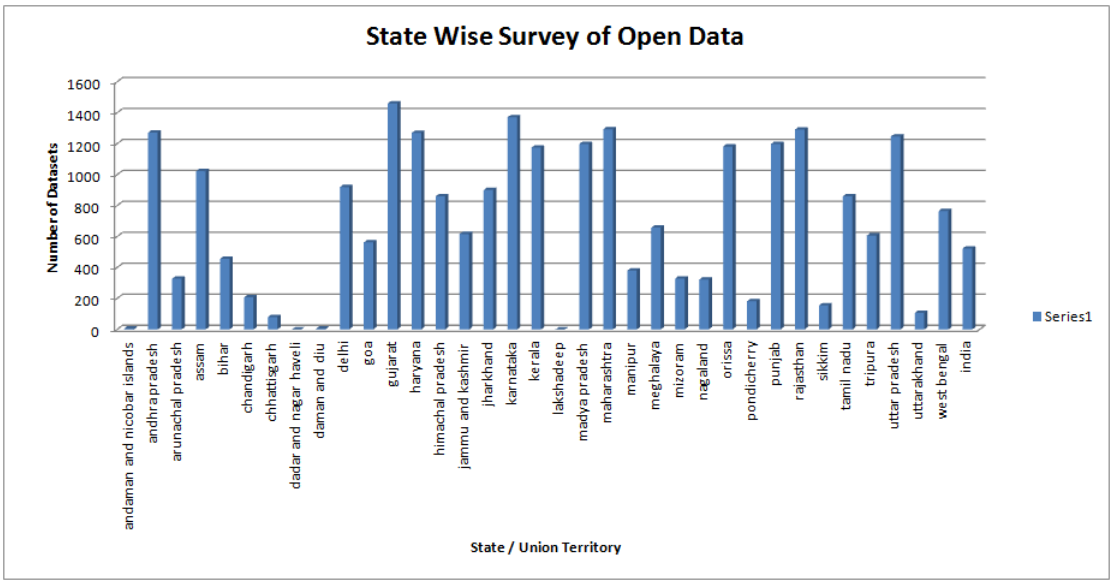

Figure 6: Bar graph of state wise distribution of open data

the mesh needs to be constructed. We will use the findings of the survey to model the open data. The categories we identified can form the basic framework for building one or more ontologies over the Indian Open Data.

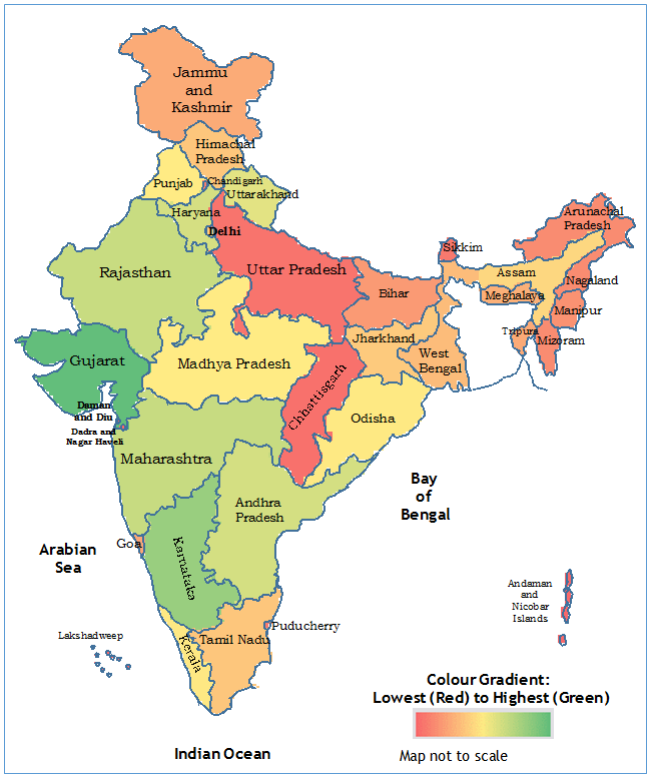

Figure 7: Heatmap of state-wise distribution of open data

7. ADDITIONAL AUTHORS

Additional authors: Sneha Deshpande (Open Systems Lab, IIITB, email: `sneha.deshpande@iiitb.org`) and Sana Javed (Open Systems Lab, IIITB, email: `Sana.Javed@iiitb.org`) and Vikas Mohandoss (NIT Surathkal, email: `vikasmohandoss@gmail.com`).

8. REFERENCES

[1] S. Balakrishnan, V. Chu, M. A. Hernández, H. Ho, R. Krishnamurthy, S. X. Liu, J. H. Pieper, J. S. Pierce, L. Popa, C. M. Robson, L. Shi, I. R. Stanoi, E. L. Ting, S. Vaithyanathan, and H. Yang. Midas: integrating public financial data. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1187–1190. ACM, 2010.

[2] C. Böhm, F. Naumann, M. Freitag, S. George, N. Höfler, M. Köppelmann, C. Lehmann, A. Mascher, and T. Schmidt. Linking open government data: what journalists wish they had known. In *Proceedings of the 6th International Conference on Semantic Systems*, page 34. ACM, 2010.

[3] K. Braunschweig, J. Eberius, M. Thiele, and W. Lehner. The state of open data. *WWW2012, Lyon, France: ACM.*, 2012.