Open Data Convergence

WHITE PAPER

International Institute of Information Technology, Bangalore

Contents

Abstract

Open Data initiative by data.gov.in has provide a great platform for sharing the datasets belonging to 55+ Departments consist of around 5000 datasets till date. These datasets are available in various format such as excel, xml, csv, etc.

But the consumer of these datasets are facing few major issue, first one is Data Integration, currently there is no mean to check what all dataset are linked with each other semantically. Second issue is Data Quality, there is no uniform data representation format, no quality check metric to show missing/invalid data.

In this white paper we are proposing a concept termed as ‘Data Convergence’ which provides unified view of data collected from different sources (or departments) in different formats. To demonstrate this we built a software application which will provide unified view by means of HTTP API in JSON/XML formats.

Main objective is to maintain loose coupling between underline storage structure and consumer client.

"Make each bit count – by creating network of datasets"

Introduction

Major shortcoming in most of open data portal (such as http://data.gov.in) is that they do not provide any API to their datasets. Some of portal like http://data.worldbank.org/, http://data.gov provide API but they are limited to produce output for a particular dataset only, yes they do have custom API builder but no way to merge/combine two or more different dataset which are having one or more common attribute/dimension.

Just having a lot of data is useless unless we can derive some useful information from it. For example from the current representation of open data, It is not obvious to identify effect of weather on number of tourist visited as both of this dataset belongs to two different department earth science and tourism respectively. And consumer may not be aware of existence of dependent dataset which could have worked as catalyst in his/her analysis. It would have been much better to have an API which will identify how datasets are connected / linked on which attribute/dimension and produce a unified view of multiple datasets.

Linked Dataset is a basement of complete Data Convergence system which consists of large number of graphs where a node represents dataset and edge between them denotes the existence of relationship. Linked Data helps to identify correlation among different parameters in different dataset.

Common flaws such as different file format, lack of consistent representation, missing/ invalid values can also be solved to certain extent in preprocessing phase of data convergence.

Problem Definition

In most of Open data portal consumer has to browse through a collection of datasets and select one dataset at a time. User has no mean to know relationship/linking among different datasets. Unless one goes through all dataset manually he/she won't get a clear idea about dependency, connectivity among datasets on certain parameter which could have served as more aggregated information.

There should be provision where consumer can provide what all data he/she wants, in which format as well as quantity and system should be able to identify how to process this query and display relevant information.

There should also be a mechanism for searching a dataset not only by its name / department, but also by its content / field / attribute / dimension.

System should be capable of displaying all linked datasets for a given dataset, where user can visualize a graph of linked datasets, also it must be able to converge this datasets and produce a unified view.

"Leveraging open data platform with the help of converged datasets"

Solution

Overview

Data convergence system at a high level takes three inputs from consumer viz., what all datasets user wants to converge? In which format? How many records per page? Later it processes the input and generates API which will give unified view of datasets.

Description

A typical use case in Data convergence system is stated below

1. Data convergence system has a Smart API builder which provides two kind of user interface to select a collection of datasets.

   1.1 Navigational

   a. Step by step navigation to select datasets provided with filter such as country, state, district, etc.,

   b. User has to first select department, then master dataset and finally choose as many to select all data.

   1.2 Search based

   a. User can search by data, dataset name, dataset description, and dataset field/attribute/dimension name in open data repository.

   b. Select any number of datasets which are displayed in query result.

2. From a given collection of datasets system identifies connected components. For example- Assume user selects dataset of Authorize Travel agency, tourism statistics, rainfall, storm and temperature. Then two connected component are identified viz.,{[Authorize Travel agency, tourism statistics], [rainfall, storm,

temperature]} depending upon what relationship is specified by dataset producer while uploading into open data repository

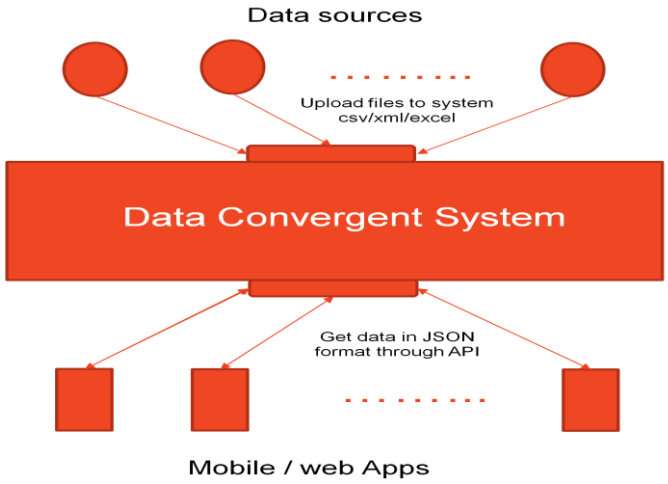
3. API is generated for each connected components which provide flexibility to modify number of records per page, offset and output format json/xml. Sample API:

   http://opendata.com/api.jsp?key=fbd7939d674997cdb4692d34de8633c4& offset=1&limit=10& format=json

4. Whenever user perform GET on API using key our data convergence system will retrieve metadata of API query create a temporary unified view and return result set in JSON/XML format

5. JSON/XML are standard data exchange format, they can be easily integrated into any application.

Data Convergence system also provides visualization of all connected dataset up to 3rd level (extensible). Few more functionalities like filtering attributes, records in unified view / result set can be added.



(fig. converged ouput in json and xml format)

Following image show converged output of two different dataset Road Accident and Person Injured for a state Arunachal Pradesh in two different format json and xml.


```
{
  "Total_Count": 1,
  "Dataset": [
    {
      "Road_Accident": {
        "d8_id": "1",
        "_sl__no_": "1",
        "_states_uts": "Andhra Pradesh",
        "_2003": " 34,826 ",
        "_2004": " 38,940 ",
        "_2005": " 37,131 ",
        "_2006": " 43,559 ",
        "_2007": " 44,325 ",
        "_2008": " 42,657 ",
        "_2009": " 43,600 ",
        "_2010": " 44,599 ",
        "_2011": " 44,165 "
      },
      "Person_Injured": {
        "d9_id": "1",
        "_sl__no_": "1",
        "_states_uts": "Andhra Pradesh",
        "_2003": "47,477 ",
        "_2004": "50,895 ",
        "_2005": "46,613 ",
        "_2006": "58,520 ",
        "_2007": "59,213 ",
        "_2008": "58,741 ",
        "_2009": "52,157 ",
        "_2010": "53,928 "
      }
    }
  ]
}
```

```
<gov.in>
 -<Dataset>
  -<e>
   -<Person_Injured>
      <_2003>47,477 </_2003>
      <_2004>50,895 </_2004>
      <_2005>46,613 </_2005>
      <_2006>58,520 </_2006>
      <_2007>59,213 </_2007>
      <_2008>58,741 </_2008>
      <_2009>52,157 </_2009>
      <_2010>53,928 </_2010>
      <_sl__no_>1</_sl__no_>
      <_states_uts>Andhra Pradesh</_states_uts>
      <d9_id>1</d9_id>
    </Person_Injured>
   -<Road_Accident>
      <_2003> 34,826 </_2003>
      <_2004> 38,940 </_2004>
      <_2005> 37,131 </_2005>
      <_2006> 43,559 </_2006>
      <_2007> 44,325 </_2007>
      <_2008> 42,657 </_2008>
      <_2009> 43,600 </_2009>
      <_2010> 44,599 </_2010>
      <_2011> 44,165 </_2011>
      <_sl__no_>1</_sl__no_>
      <_states_uts>Andhra Pradesh</_states_uts>
      <d8_id>1</d8_id>
    </Road_Accident>
   </e>
  </Dataset>
  <Total_Count>1</Total_Count>
</gov.in>
```

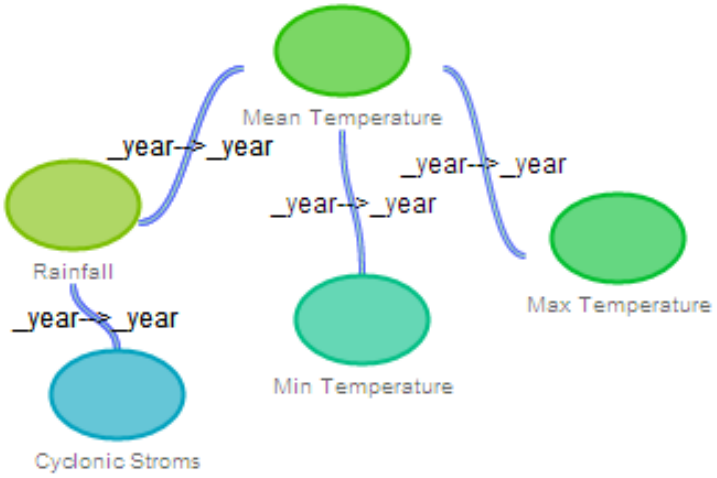

(fig. API response in JSON and XML format)

Benefits

Data Convergence System will provide a single uniform HTTP API access to a unified view of dataset. It also provides user a facility to visualize the network of connected datasets.

Key Highlights

- Easy Access to converged data through API,
- Standard data exchange format (JSON and XML) are supported
- Flexible (select dimension as required)
- Unified view support real time data
- Loose coupling
- Notifications generated with change data capture (CDC)



(fig. Visualization of Datasets relationship)

Conclusion

An attempt is been made to introduce a concept which will provide an unified view for Datasets present in open data portal based on the concept of Data Convergence which will leverage the way of accessing open data in this document.

Consumer of Open Data can now have a much more idea and knowledge of Datasets and their relationships, eventually it will stimulate their data analysis process.

Contact Information

Prof. Chandrashekhar Ramanathan
Tel: +91 80 4140 7777
rc@iiitb.ac.in

Kodamasimham Pridhvi
Tel: +91 8123160887
Pridhvi.kodamasimham @iiitb.org

Bisen Vikrantsingh M.
Tel: +91 8792708719
Bisenvikrantsingh.mohansingh@iiitb.org

Institute

IIIT Bangalore
IIIT-Bangalore,26/C, Electronics City, Hosur Road, Bangalore, 560100
Tel+91 80 4140 7777 / 2852 7627
http://www.iiitb.ac.in