Text Analysis of International Open Government Data

John S. Erickson (erickj@rpi.edu), Amar Viswanathan (kannaa@rpi.edu), Josh Shinavier(shinaj@rpi.edu), Yongmei Shi(shiy@rpi.edu), James A. Hendler (hendler@cs.rpi.edu)

Tetherless World Constellation
(Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States)

International Open Government Data Search

•**I**nternational **O**pen **G**overnment **D**ata **S**earch aims to discover, document and analyze data catalogs that are published by different governments all over the world.

•IOGDS demonstrates a practical dataset catalog metadata model for integrating these diverse dataset catalogs and then linking them on the Linked Data Cloud.

•By early 2013 the IODGS project had harvested metadata for over 1,022,787 datasets from 192 catalogs in 24 languages, representing 43 countries and international organizations.

•RPI's aggregate catalog, implemented in RDF and published using both a public SPARQL endpoint and a faceted user interface, has proven to be a valuable tool for gaining insight into the nature of open government data publication.

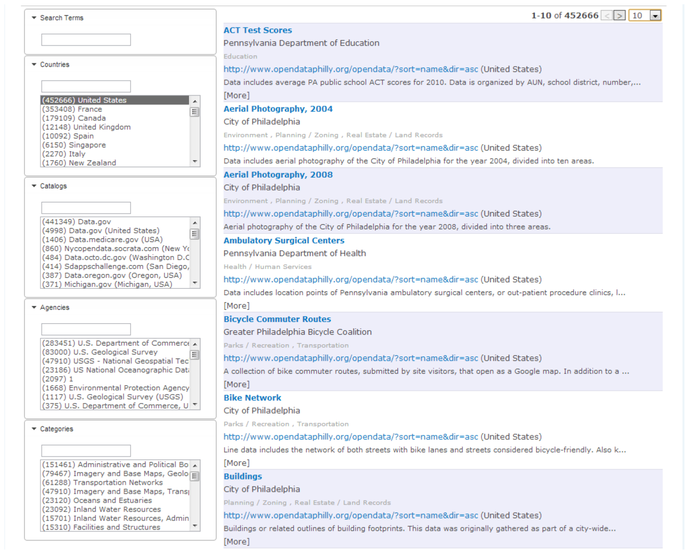

Figure 1. Faceted Browsing and Search developed by TWC

Open Government Data Collection and Linked Data Production

•**O**pen **G**overnment **D**ata catalogs are collected for IOGDS through a semi-automated process that includes manual catalog identification and customized semi-automated catalog harvesting.

•Government Data Catalogs are located by both Open Knowledge Foundation's catalog referred to as the "catalog of catalogs" and also through manual Google search.

•Once the catalog is identified, we look for the ways in which it is published. We pick catalogs in which content is static and the structure is fairly evident. A custom scraper then scrapes the meta data and harvests them to csv files

•Although most catalogs share common dataset properties like the title, description, the metadata from different catalogs adopt different metadata structures and vocabularies. To integrate all the catalogs and make them available as Linked Data we developed a metadata model to describe *catalog* and *dataset* concepts. Based on a earlier proposal for government dataset description and informed by DCAT, our vocabulary was developed to fit existing dataset catalogs found anywhere.

Inside Linked Government Data : IOGDS demystified

The IOGDS catalog has served as an observatory for exploring the diversity of datasets published by different governments. We provide some insights from the analysis of the data here.

Catalogs published by individual countries: During our collection period the United States was the global leader in publishing (453,859 datasets), followed by France (353,394), Canada (179,131), the United Kingdom (12,131) and Spain (10,076). These statistics include geographic datasets of various kinds, which account for the largest subset of entries.

Languages across all catalogs: Our analytics found a total of 24 languages represented across all catalogs in the IOGDS collection. English is by far the most prominent language (98 catalogs / 652,176 datasets), since most of the open government data published through late 2012 came from English-speaking countries. Other notable languages included French (19 catalogs / 528,153 datasets), Spanish (18 catalogs / 8,444 datasets), Italian (14 catalogs / 2,256 datasets), and German (12 catalogs / 1,584 datasets). 19 languages were represented in the remaining 31 catalogs.

Analytics of Government Data Publication

Figure 5. Languages in IOGDS

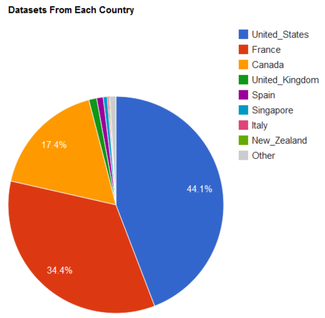

Figure 6. Dataset distribution

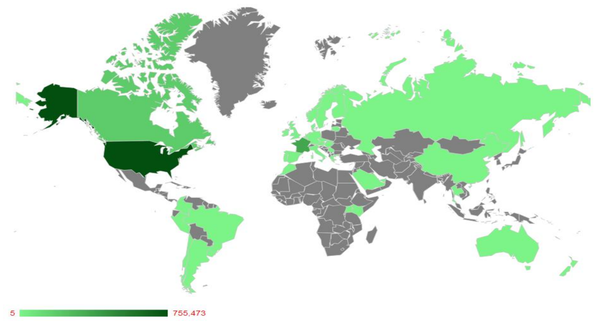

Figure 7. Countries that publish open government data

Analyzing the Textual Content of Metadata

•To identify trends present in the datasets we decided to perform text analysis on the keywords associated with the datasets and categories. This analysis gave us an idea of what is present in the various catalogs at a glance. Simple word clouds were used to illustrate this insight.

•The IOGDS category metadata characterizes subsets of datasets across catalogs. Analysis of the distribution of categories across countries gives us some insight into the priorities of governments as they publish their data.

•The analysis of this data was done on Microsoft's new F# platform along with the extensions provided by both D3.js and jQuery.

•The analysis results give us an insight into how well published the meta data is and also gives us a means of comparing datasets across countries in terms of what they are publishing.

United States	United Kingdom	Canada	Australia	Germany
Administrative and Political Boundaries (18.01%)	Health and Social Care(10.53%)	Economics and Industry(47.36%)	Community(11.09%)	Air (11.27%)
Imagery and Base Maps(12.87%)	Economy(4.07%)	Health and Safety(12.34%)	Geography(9.31%)	Water(11.06%)
Transportation Networks(12.49%)	Children, Education and Skills(3.67%)	Society and Culture(6.34%)	Government(8.79%)	Other(8.08%)
imageryBaseMapsEarthCover(11.083%)	People and Places(3.43%)	Nature and Environment(4.76%)	Business(8.79%)	Politics and government(5.14%)
boundaries(9.38%)	Population(2.93%)	Labour(3.71%)	Environment(7.68%)	Economy(4.16%)
Geological and Geophysical(7.91%)	Agriculture and Environment(2.54%)	Persons(3.24%)	Finance(6.89%)	Education and Science(3.92%)
Locations and Geodetic Networks(7.85%)	Crime and Justice(2.21%)	Agriculture(2.84%)	Sciences(6.38%)	Radiation
Inland Water Resources(5.54%)	Travel and Transport(2.04%)	Education and Training(2.66%)	Society(6.18%)	Public Administration,Budget & Taxes
transportation(3.66%)	Business and Energy(2.01%)	Transport(2.1%)	Industry(5.50%)	Traffic(3.18%)
Oceans and Estuaries(2.33%)	Parish(1.84%)	Government and Politics(1.84%)	Recreation(3.68%)	Labor market(2.94%)
inlandWaters(1.91%)	Government(1.69%)	Impacts & Environmental Change(1.41%)	Culture(2.85%)	People Family and Social Affairs(2.94%)
oceans(1.58%)	Health(1.53%)	Science and Technology(1.19%)	Health(2.33%)	Climate Environment and Nature(2.69%)
Facilities and Structures(1.51%)	Demographics(1.32%)	Information and Communications(1.14%)	Law(2.06%)	Basic data and Geosciences(2.20%)
structure(0.87%)	Labour Market(1.32%)	Energy & GHG Emissions(1.03%)	Transport(1.86%)	Building and living(1.96%)
climatologyMeteorologyAtmosphere(0.44%)	Education(1.02%)	Law(0.86%)	Planning(1.66%)	Trade commerce(1.71%)
Geography and Environment(0.20%)	health(0.95%)	Business & Economic Development(0.48%)	General(1.58%)	Leisure Culture and Tourism(1.71%)
elevation(0.1%)	Employment and Skills(0.85%)	Census(0.48%)	Safety(1.42%)	Population growth(1.71%)
biota(0.09%)	Environment(0.79%)	Processes(0.33%)	Property(1.22%)	Demographics(1.47%)
Environment and Conservation(0.06%)	transport(0.75%)	Development(0.25%)	Communication(1.14%)	Environment and Climate(1.47%)

Figure 2. Table showing the different categories published by the top five countries

Figure 3. Top keywords found in the dataset catalog Data.gov.uk

Figure 4. Top keywords found in the dataset catalog Data.gov from the United States

Conclusion : Where do we go from here

Conclusion : This study was done to find out what the datasets have in terms of their content and also their diversity. Moving forward this data can be fed into Deep NLP systems like Watson for intelligent question answering.

Acknowledgments:
This work has been made possible by a generous gift to the Tetherless World Constellation at Rensselaer Polytechnic Institute from Microsoft Research.