# "Not Elementary, My Dear Watson.."- Extending Watson for Question Answering on Linked Open Government Data

**James Hendler[1] (hendler@cs.rpi.edu), Amar Viswanathan[1] (kannaa@rpi.edu), Siddharth Patwardhan[2] (siddharth@.us.ibm.com),**

**([1]Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States,**

**[2] IBM T.J. Watson Research Center)**

## TWC Linking Open Government Data (LOGD)

- The Linking Open Government Data (LOGD) project investigates opening and linking government data using Semantic web technologies. We translate government-published datasets to **RDF** and link them to the Web of Data. In addition we also discover, document and analyze data catalogs that are published by different governments all over the world.

- As part of the IOGDS project a practical dataset catalog metadata model has been created for integrating these diverse dataset catalogs and then linking them on the Linked Data Cloud and 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.


Figure 3. Languages in IOGDS

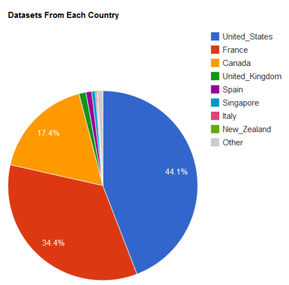


Figure 2. Dataset distribution

## 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.

**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.

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.

## What's out there : Analyzing the Linked Open Government Datasets Corpus

- 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.

- A casual glance at the data shows us that underneath the linked structure, the datasets describe a wealth of information about the datasets in the textual metadata.

- To utilize these datasets for Watson, we decided to summarize these RDF datasets in the ITA Controlled English(CE) format.

| 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 4. Table showing the different categories published by the top five countries


Figure 5. Top keywords found in the dataset catalog Data.gov.uk from the United Kingdom


Figure 6. Top keywords found in the dataset catalog Data.gov from the United States

## Watson for Linked Open Government Data

- We use the **Watson-Mini system** - a minimal implementation of the Watson system deployed at RPI for the Question Answering Phase. Watson-mini relies on text based knowledge sources to answer questions, we converted the entire RDF dataset dump into a textual knowledge base in the Controlled Natural Language(CNL) format. While this preserves the entire RDF knowledge base as is, it also serves as an excellent knowledge base and a corpus to Watson-mini.

- The whole RDF dataset dump is then fed to the INDRI index in the Watson-Mini system as title-oriented documents. Each of the dataset is translated to a title oriented document by using the term dataset title i.e. the DCMI Metadata term for the title of the dataset as the title of a single document. The metadata of the document is then added to the textual knowledge base for that document.

- Watson's TIC Passage search component then uses these characteristics of title-oriented documents to reduce the search to relevant documents (or) datasets in this case. The titles of the datasets present a vague clue for the system to find an answer. Watson at this stage relies on INDRI's passage search capability, which uses a combination of language modeling and inference networks for ranking documents.

## Question types, Answer Formats and initial results

- We created a set of 500 training question-answer pair that were then fed to the Watson-mini as an initial experiment. These training questions were based on the following categories – **Health, Agriculture, Economics and Environment.** These were the focus topics that are present across datasets. The categories were also chosen because of the nature of the diverse datasets that are present under them i.e. these categories have datasets from different countries.

- The questions were also pruned manually and were given an initial classification of the following types :**TITLE ORIENTED** i.e. questions that have answers on their title or if the dataset title provides the answer e.g. ***What datasets talk about Child care in the corpus?*, LOCATION ORIENTED** i.e. questions ,whose answers are focused on the location of the dataset type as well i.e. ***What are the datasets that describe food production in Europe* , DESCRIPTION ORIENTED** i.e. questions that rely on the description of the metadata to provide matched to the answer i.e. **What datasets provide details about locations of movies shot in San Francisco from the year 1924 to 2010?** Yes the above question actually returns an answer that points to dataset number 22 published by the "San Francisco Film Commission" and even has the title "Film Locations in San Francisco"

## Conclusions and Future Work

- In this poster we present some initial work on the Watson Mini system that can effectively use Government Datasets to answer questions related to Government Data. While the initial demos suggest that such a system is possible, it is still in the nascent stages and further work needs to be done for Answer Ranking and also using RDF graphs as a corpus instead of text data.