Discovering the Hidden Cross-Dataset Links in Data.gov

Johanna Flores
Tetherless World Constellation
Rensselaer Polytechnic Institute, 110 8th St.
Troy, NY 12180, USA
florej6@rpi.edu

Li Ding
Tetherless World Constellation
Rensselaer Polytechnic Institute, 110 8th St.
Troy, NY 12180, USA
dingl@cs.rpi.edu

ABSTRACT

Data mash-ups unleash users from the limit of accessing one dataset at a time, and enable a broader view based on the integrated data. A key challenge to mashing up Data.gov datasets lies in the fact that cross-dataset links are rarely published explicitly by dataset owners, making it hard for users to find related datasets for building mash-ups. In this paper, we show several types of hidden cross-dataset links found in Data.gov and explain how they can be obtained using semantic technologies in Rensselaer's Linking Open Government Data project.

Keywords

Data.gov, Linked Data

1. INTRODUCTION

Data.gov is a web portal serving over 300K open government data (OGD) datasets from US federal government agencies. This effort not only opens up lots of raw government data on the Web for reuse, but also promotes the development of mash-ups, wherein multiple OGD datasets are integrated to support global analysis (Berners-Lee 2009). Data mash-ups unleash users from the limit of accessing one dataset at a time, and enable a broader view based on the integrated data. They are often facilitated by explicitly declared links between datasets, either provided within the datasets themselves or in the supplementary metadata. Creating mash-ups, however, can often be challenging since that relies on users' knowledge of these cross-links, which help to determine how datasets can be integrated meaningfully. Most federal government datasets published on Data.gov unfortunately lack any cross-links between their own data (or any outgoing links to other widely used data corpuses, for that matter). While explicit links may not be presented in Data.gov datasets, there are automated techniques which can be used to infer possible links between data with reasonably high accuracy. In this paper, we examine these methods and demonstrate two examples of how they can be used to provide a means for users to raise and view the hidden cross-links in Data.gov datasets.

2. THE PROBLEM

In linked data, references between datasets often appear in the form of named identifiers, such as keywords and URIs. A URI, (similar to a URL), is an exclusive type of identifier which typically allows a user uniquely identify something and then follow certain name-resolution protocols (e.g. HTTP) to locate the original definition/metadata of the referred thing. In comparison with string identifiers (e.g. "New York") which carries multiple meaning (e.g. the city and the New York state), a URI (e.g. dbpedia:New_York) offers a precise reference to New York State without any ambiguity. In Data.gov dataset, unfortunately, things are mentioned in datasets using string identifiers, and New York state can be mentioned using different labels, e.g. "New York", "NY" , and "36" (in FIPS code[1]). Therefore, even building one cross-link could be a non-trivial task for human users, and Natural Language Processing (NLP) technologies are needed for automation.

Given the enormous size of many OGD datasets (some numbering in the millions of data rows), it is unlikely that the casual data consumer will expend such considerable effort to manually review the raw data (of multiple datasets) and extract any cross-links between datasets. Therefore an automatic approach is preferable not only for the sake of efficiency but also to make OGD equally accessible.

A key challenge to mashing up Data.gov datasets lies in the fact that cross-dataset links are rarely published explicitly by dataset owners, making it hard for users to find related datasets for building mash-ups. These datasets contain plentiful amounts of metadata and string literals, however, which can be treated as guidance terms and used to discover cross-dataset links. In this paper, we show several types of hidden cross-dataset links found in Data.gov and explain how they can be obtained using semantic technologies in Rensselaer's Linking Open Government Data project (Ding, et al. 2010).

3. TECHNIQUES FOR DISCOVERING HIDDEN CROSS-DATASET LINKS

In principle, Data.gov datasets can be linked at several levels:

i. metadata-level: linked by common attributes in metadata
ii. content-level: linked by common entities in content
iii. social-level: linked by social context

While the third level has already been explored in the TWC LOGD portal via web page navigation, the first two levels are the focus of this paper - showing how automatic semantic technologies are used to reveal hidden cross-dataset links.

WebSci '11, June 14-17, 2011, Koblenz, Germany.

[1] http://www.itl.nist.gov/fipspubs/fip5-2.htm

3.1 Metadata-level Linking

Data.gov datasets are published with a moderate amount of text – based metadata, including dataset title, publishing agency, keywords, description, category, links to additional metadata and etc. It is very easy to process the metadata of two datasets by looking for any overlap, such as shared keywords. Any mutual content discovered can be utilized as a link between two or more datasets. The casual use of this method involves almost no semantic or natural language processing beyond searching for string matches and therefore it is fairly easy to implement. This type of linking can offer a meaningful (although somewhat topical) view of the cross-dataset links contained in multiple datasets.

3.2 Content-Level Linking

Sometimes keyword search is insufficient for finding links between datasets and so a deeper semantic analysis of the content data they contain is required. In content-level linking, the literal string values comprising the data are viewed as potential named entities. A named entity can be anything which is conceivable, such as a specific city, a person, and a chemical. The aim of content-level linking is to perform entity resolution, in which a variety of NLP and semantic verification techniques are used to make the determination of whether or not a string representation can be resolved to a named entity. Resolved entities can be uniquely identified by URIs (which represent named entities) extending the benefit of not only providing a standardized (and machine-readable) representation for discovered entities but also opening up the dataset to possible linking with other foreign sources which also contain those URIs. Mapping entities to URIs can be tricky since a single named entity can have a variety of legitimate textual representations. Data.gov datasets are published by different agencies, each of which has its own way of referring to named entities, such as states. For example, one agency might use a state's name to refer to it (i.e., Alabama), whereas a different agency may instead use a state's FIPS or postal code (i.e., 01, AL). Entity resolution attempts to (ideally) capture all semantically-correct possible representations of a target entity within the data. Although content-level linking is much more involved and expensive to perform than metadata-linking, it has the potential to provide a comprehensive and accurate portrayal of the cross-dataset links which may exist in data.

4. CASE STUDIES

In the following sections, we will present two implementations which utilize the previously described linking techniques to reveal hidden links in OGD datasets.

4.1 Multiword Tag Cloud

This work partitions one dataset's title into a list of multi-word phrases, so that users may quickly find relevant datasets by clicking an entry in a tag-cloud. In this case, the hidden links are exposed from the shared multi-word phrases in the datasets' metadata. The work has two highlights. (i) Meaningful *Links*. The "Multi-Word TagCloud" approach better captures the meaning conveyed in the title of a dataset. As shown in Figures 1 and 2, the single-word tags "release" and "inventory" are less meaningful cross-dataset links than the corresponding multi-word-tag "toxic-release-inventory-data". (ii) *Automated Process*. The Multi-Word TagCloud is automatically generated by computer program, powered by the Microsoft Web N-gram service (Wang et al. 2010). The segmentation algorithm leverages the service (using the dataset titles' corpus) to glue together highly dependent words into phrases.

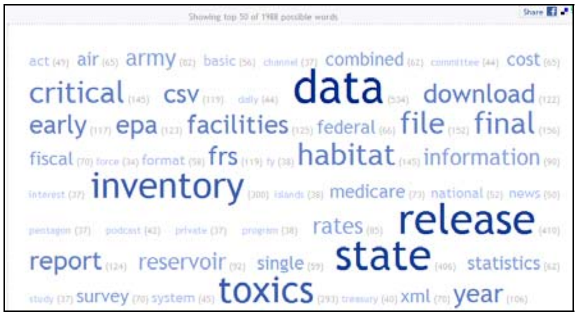

Showing top 50 of 1988 possible words
Share
act (49) air (65) army (82) basic (56) channel (37) combined (62) committee (44) cost (65)
critical (145) csv (119) daily (44) data (534) download (122)
early (117) epa (123) facilities (125) federal (66) file (152) final (156)
fiscal (70) force (34) format (58) frs (119) fy (38) habitat (145) information (90)
interest (37) inventory (300) islands (38) medicare (73) national (52) news (50)
pentagon (37) podcast (42) private (37) program (38) rates (85) release (410)
report (124) reservoir (92) single (59) state (406) statistics (62)
study (37) survey (70) system (45) toxics (293) treasury (40) xml (70) year (106)

Figure 1. Conventional Single-word TagCloud generated from Data.gov dataset titles

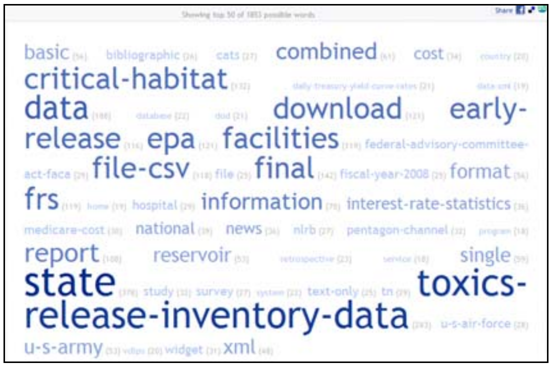

Showing top 50 of 1853 possible words
Share
basic (54) bibliographic (26) cats (27) combined (61) cost (34) country (20)
critical-habitat (132) daily-treasury-yield-curve-rates (21) data-xml (19)
data (188) database (22) dod (21) download (121) early-release (116)
epa (121) facilities (118) federal-advisory-committee-act-faca (29)
file-csv (118) file (25) final (142) fiscal-year-2008 (25) format (54)
frs (119) home (19) hospital (29) information (78) interest-rate-statistics (36)
medicare-cost (30) national (39) news (36) nlrb (27) pentagon-channel (32) program (18)
report (108) reservoir (53) retrospective (23) service (18) single (59)
state (376) study (33) survey (37) system (22) text-only (25) tn (29) toxics-release-inventory-data (293) u-s-air-force (28)
u-s-army (53) video (20) widget (31) xml (48)

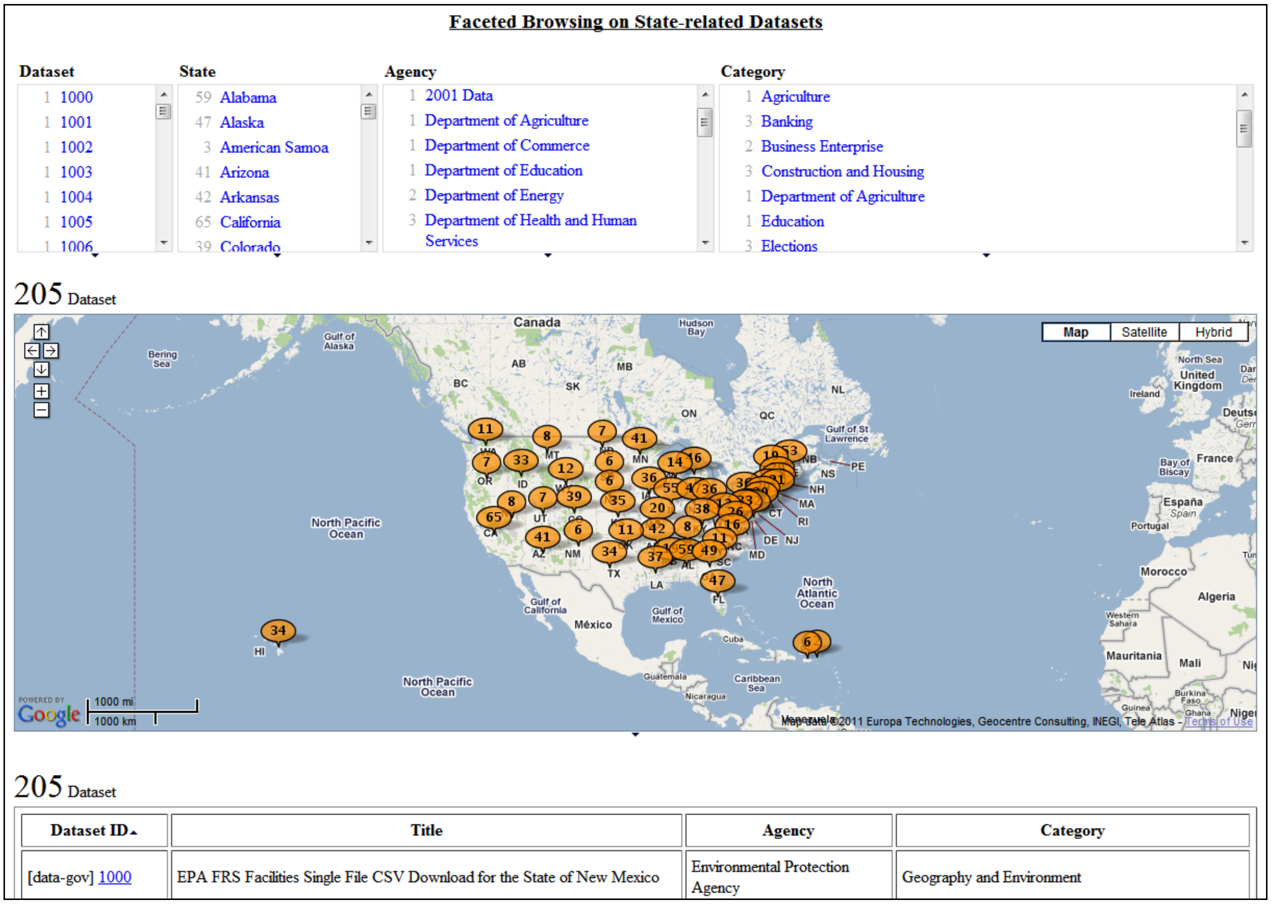
Figure 2. Novel Multi-word TagCloud generated from Data.gov dataset titles

4.2 Faceted Browsing on State-related Datasets

This work identified US states mentioned in the content of Data.gov datasets, and provides a faceted browsing interface for users to list which Data.gov datasets are linked by a given US state (see Figure 3).

4.2.1 Implementation

A content-level linking approach is taken to create an automated process which allows the discovery of the state-related cross-dataset links in Data.gov datasets. This work is highlighted by its semantic analysis, which leverages Yahoo! Boss search to guess DBpedia URIs for complete/abridged state names. Adapted natural language processing heuristics are employed on relational structures (Bhattacharya and Getoor 2007) to deal with other state representations, i.e., abbreviations, FIPS code, etc. The heuristics are primarily used to perform verification on the candidates for entity resolution and generate a confidence measure for how certain it is that the literal and corresponding entity should be mapped. Once states have been identified within a dataset, they are recorded and output in a new file linking the states to the dataset. The output datasets are converted to JSON and are then loaded into a faceted browser created using the Simile Exhibit visualization tool. Detailed algorithm of this implementation is described in (Flores 2011).



Faceted Browsing on State-related Datasets

	Dataset
1	1000
1	1001
1	1002
1	1003
1	1004
1	1005
1	1006

	State
59	Alabama
47	Alaska
3	American Samoa
41	Arizona
42	Arkansas
65	California
39	Colorado

	Agency
1	2001 Data
1	Department of Agriculture
1	Department of Commerce
1	Department of Education
2	Department of Energy
3	Department of Health and Human Services

	Category
1	Agriculture
3	Banking
2	Business Enterprise
3	Construction and Housing
1	Department of Agriculture
1	Education
3	Elections

205 Dataset

205 Dataset

Dataset ID	Title	Agency	Category
[data-gov] 1000	EPA FRS Facilities Single File CSV Download for the State of New Mexico	Environmental Protection Agency	Geography and Environment

Figure 3. Conventional Single-word TagCloud generated from Data.gov dataset titles

4.2.2 Results

Among a selection of 400 Data.gov datasets, we discovered that 205 datasets mentioned at least one US state (either by name or postal code) and that California was mentioned the most. In this case, the hidden links are exposed via co-referenced US states.

CONCLUSION

The above cases have yielded interesting results and a greater understanding of the hidden cross-dataset links, which could help users to leverage government data effectively and efficiently. Our current work started with a few simple semantic analyses, and we will try more semantic technologies and tools to expand this framework in the future.

ACKNOWLEDGMENTS

We wish to extend our gratitude to the Tetherless World Constellation and the Linking Open Government Data project at RPI for providing the infrastructure, resources and expertise which allowed us to perform these experiments in data linking. This work is partially supported by DARPA's Transformational Convergence Technology Office, and gifts from Lockheed Martin, Fujitsu Labs America, LGS, and Microsoft.

REFERENCES

[1] Berners-Lee,T. (2009). Putting government data online. http://www.w3.org/DesignIssues/GovData.html

[2] Bhattacharya, I. and Getoor, L.. (2007). Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data* 1(1), Article 5.

[3] Ding, L., DiFranzo, D., Graves, A., Michaelis, J., Li, X., McGuinness, D., and Hendler, J. (2010). TWC data-gov corpus: incrementally generating linked government data from data.gov. in WWW'2010 (developer track).

[4] Wang, K., Thrasher, C., Viegas, E., Li, X., and Hsu, B.. (2010). An overview of Microsoft web N-gram corpus and applications. In *Proceedings of the NAACL HLT 2010 Demonstration Session.*

[5] Flores, J. (2011). Automated implicit linking of open government data. Master's thesis, Rensselaer Polytechnic Institute, May 20.