Chapter 97
The Importance of Authoritative URI Design Schemes for Open Government Data

Alexei Bulazel
Tetherless World Constellation, Rensselaer Polytechnic Institute, USA

Dominic DiFranzo
Tetherless World Constellation, Rensselaer Polytechnic Institute, USA

John S. Erickson
Tetherless World Constellation, Rensselaer Polytechnic Institute, USA

James A. Hendler
Tetherless World Constellation, Rensselaer Polytechnic Institute, USA

ABSTRACT

A major challenge when working with open government data is managing, connecting, and understanding the links between references to entities found across multiple datasets when these datasets use different vocabularies to refer to identical entities (i.e.: one dataset may refer to Microsoft as "Microsoft", another may refer to the company by its SEC filing number as "0000789019", and a third may use its stock ticker "MSFT".) In this paper the authors propose a naming scheme based on Web URLs that enables unambiguous naming and linking of datasets and, more importantly, data elements, across the Web. They further describe their ongoing work to demonstrate the implementation and authoritative management of such schemes through a class of web service they refer to as the "instance hub". When working with linked government data, provided either directly from governments via open government programs or through other sources, the issue of resolving inconsistencies in naming schemes is particularly important, as various agencies have disparate conventions for referring to the same concepts and entities. Using

DOI: 10.4018/978-1-5225-5191-1.ch097

linked data technologies the authors have created instance hubs to assist in the management and linking of entity references for collections of categorically and hierarchically related entities. Instance hubs are of particular interest to governments engaged in the publication of linked open government data, as they can help data consumers make better sense of published data and can provide a starting point for development of linked data applications. In this paper the authors present their findings from the ongoing development of a prototype instance hub at the Tetherless World Constellation at Rensselaer Polytechnic Institute (TWC RPI). The TWC RPI Instance Hub enables experimentation and verification of proposed URI design schemes for open government data, especially those developed at TWC in collaboration with the United States Data.gov program. They discuss core principles of the TWC RPI Instance Hub design and implementation, and summarize how they have used their instance hub to demonstrate the possibilities for authoritative entity references across a number of heterogeneous categories commonly found in open government data, including countries, federal agencies, states, counties, crops, and toxic chemicals.

INTRODUCTION

Motivated by the Obama administration's government transparency initiatives, in May 2009 the United States launched the Data.gov web portal with a catalog of 47 datasets containing government information that had previously not been easily available online. (Kirkpatrick, 2009) The aggregation of datasets in a single online portal at Data.gov made them easier to find and search, and has led to the publication of datasets previously not available online. During its first year Data.gov grew to include more than 250,000 datasets and has inspired the creation many hundreds of applications and services. (Kundra 2010)

The launch of Data.gov was a key event near the beginning of a worldwide movement of open government data[1] publication as governments, NGOs, and many other institutions began to make their data openly accessible to interested parties. During Data.gov's first year, US municipalities including San Francisco and New York City, and states including California, Utah, Michigan, and Massachusetts launched open data portals; around the world, countries including the UK, Canada, and Australia, as well as organizations such as the World Bank followed the Data.gov lead. By early 2013, the International Open Government Dataset Search[2] project of the Tetherless World Constellation at Rensselaer Polytechnic Institute (TWC RPI) had recorded nearly 200 catalogs from over 40 countries, totaling more than a million datasets spanning a vast array of topics. (Erickson et al, 2011)

The significant growth in number and size of open government data catalogs since 2009 has been made possible by the emergence of an open government data ecosystem consisting of policy makers, agencies (as providers and consumers), data experts, independent software developers and service providers, academia, and citizen stakeholders. The publication of widely varied data has inspired a wide range of applications and services, has provided essential data for journalists, bloggers and activists, and has fueled academic research. In turn, demand from stakeholders has increased the quantity, quality and diversity of this data.

The growth in the availability of open government data from providers around the world is encouraging, but for potential users of this data including developers of applications and services, the variety of formats and practices used to publish the data can cause interoperability, scalability, and usability problems. Government datasets are typically published "as is" (i.e., using a variety of structures and formats), requiring substantial human workload to clean them up for machine processing and to make