Improving Curated Web-Data Quality with Structured Harvesting and Assessment

Kevin Chekov Feeney, Trinity College Dublin, Dublin, Ireland
Declan O’Sullivan, Trinity College Dublin, Dublin, Ireland
Wei Tai, Trinity College Dublin, Dublin, Ireland
Rob Brennan, Trinity College Dublin, Dublin, Ireland

ABSTRACT

This paper describes a semi-automated process, framework and tools for harvesting, assessing, improving and maintaining high-quality linked-data. The framework, known as DaCura', provides dataset curators, who may not be knowledge engineers, with tools to collect and curate evolving linked data datasets that maintain quality over time. The framework encompasses a novel process, workflow and architecture. A working implementation has been produced and applied firstly to the publication of an existing social-sciences dataset, then to the harvesting and curation of a related dataset from an unstructured data-source. The framework’s performance is evaluated using data quality measures that have been developed to measure existing published datasets. An analysis of the framework against these dimensions demonstrates that it addresses a broad range of real-world data quality concerns. Experimental results quantify the impact of the DaCura process and tools on data quality through an assessment framework and methodology which combines automated and human data quality controls.

Keywords: Data Curation, Data Quality, Digital Humanities, Linked Data, Web Data

INTRODUCTION

The ‘publish first, refine later’ philosophy associated with Linked Data (LD) has resulted in an unprecedented volume of structured, semantic data being published in the Web of Data. However, it has also led to widespread quality problems in the underlying data: some of the data is incomplete, inconsistent, or simply wrong. These problems affect every application domain and serve as significant impediments in building real-world LD applications (Zaver, 2013a): it is inherently difficult to write robust programs which depend upon incomplete and inconsistently structured data. In order for real-world applications to emerge which fully leverage the Web of Data, there is a need for higher-quality datasets – more complete, more consistent and more correct - and this implies the need for tools, processes and

DOI: 10.4018/ijswis.2014040103
methodologies which can monitor, assess, improve and maintain data quality over time.

The focus of the research described in this paper is the maintenance of data quality in a locally managed dataset. This includes the maintenance of inter-links to remote datasets but does not address the quality of those linked datasets directly. The emphasis is on curating a Linked Dataset in such a way that it can be easily and reliably consumed by third parties and linked to from other datasets. Within these limits, we are concerned with managing the full life-cycle of the local dataset. The life-cycle starts with the creation or collection of the data, which could be fully automated, scraped from existing web sources, require manual user input, or anything in between. It continues through review, assessment and compilation to the publication of the dataset for consumption by third parties (Auer, 2012). Our goal is to design a process, and a technical framework to support that process, which maximizes dataset quality over time, even as both the dataset and the underlying schema changes.

The basic research question that we are addressing is how one can build a linked data platform which allows people to harvest, curate and publish datasets in such a way that maximizes data quality over time. In order to answer this question, we need to address several sub-problems:

- How can data quality be evaluated and measured? Our aim is to produce a system to harvest and curate datasets which maintain a sufficiently high quality that they can serve as a basis for reliable application development. Thus a broad, multi-dimensional view of data quality is required to capture the variety of factors that are important to achieve this goal.
- What is a suitable workflow to integrate data quality checks into a linked data management life-cycle? Our goal is to produce a platform that can be used by non-knowledge engineers to harvest and curate datasets and our workflow must reflect this.
- What is a suitable architecture to allow us to efficiently instantiate this workflow?
- What tools and user interfaces can assist dataset managers, contributors and users in improving dataset quality? Once again, the answer to this question is influenced by our goal of supporting non-knowledge engineers.
- How can we evaluate the effectiveness of our process, architecture and tools in improving data quality?

We assume that our linked dataset is encoded as an RDF graph and stored in a triplestore. RDF is, by itself, a very flexible data model and RDF datasets can be self-describing. RDFS and OWL are languages, built on top of RDF, which provide support for the definition of vocabularies and ontologies. By defining schemata in these languages, automated validation algorithms can be applied to ensure that instance data is consistent with the schema defined for the relevant class. However, even with exhaustively specified RDFS / OWL schemata, there are still many degrees of freedom which present individuals with opportunities for making different, incompatible choices (Hogan, 2010). In practice, Linked Data datasets tend to employ various conventions which are amenable to automated validation but are not easily expressed as semantic constraints (Hyland, 2013). For example, RDF specifies that every resource is identified by a URI, and Linked Data URIs should dereference to resource representations (Bizer, 2009) but such constraints are beyond the scope of OWL and RDFS.

Data quality depends upon more than fidelity to a schema – in most domains it is also important that data be correct, that it is an accurate model of whatever real world entities it purports to represent (Zaver, 2013b). Validation of data correctness with respect to the real world entities it purports to model is often beyond the reach of automated mechanisms, as it depends upon domain expertise. For example, a dataset which represents a historical era may contain dates of events and the accuracy of such information cannot be properly ascertained.
Related Content

Data Integration in the Geospatial Semantic Web

Visualising Social Networks in Collaborative Environments

A Distributed Patient Identification Protocol Based on Control Numbers with Semantic Annotation
Automatic Sentiment Analysis on Web Texts for Competitive Intelligence
www.igi-global.com/chapter/automatic-sentiment-analysis-web-texts/71863?camid=4v1a