PatchR: A Framework for Linked Data Change Requests

Magnus Knuth, Hasso Plattner Institute for Software Systems Engineering, University of Potsdam, Potsdam, Germany

Harald Sack, Hasso Plattner Institute for Software Systems Engineering, University of Potsdam, Potsdam, Germany

ABSTRACT

Incorrect or outdated data is a common problem when working with Linked Data in real world applications. Linked Data is distributed over the web and under control of various dataset publishers. It is difficult for data publishers to ensure the quality and timeliness of the data all by themselves, though they might receive individual complaints by data users, who identified incorrect or missing data. Indeed, the authors see Linked Data consumers equally responsible for the quality of the datasets they use. PatchR provides a vocabulary to report incorrect data and to propose changes to correct them. Based on the PatchR ontology a framework is suggested that allows users to efficiently report and data publishers to handle change requests for their datasets.

Keywords: Change Management, Crowd-Sourcing, Data Curation, Linked Data, RDF

INTRODUCTION

With the continuous growth of Linked Data on the World Wide Web and the increase of web applications that consume Linked Data, the quality of Linked Data resources has become a relevant issue. Recent initiatives, as the Pedantic Web group\(^1\) and DBpedia Data Quality Evaluation Campaign\(^2\) uncovered various defects and flaws in Linked Data resources. Apart from structural defects, semantic flaws and factual mistakes are hard to detect by automatic procedures and require updates on the schema level as well as on the data level.

It is in fact a problem that erroneous data is distributed and reused in various semantic web applications, but it also opens up opportunities for joint efforts such as crowdsourcing to improve data quality. Indeed, we see Linked Data consumers equally responsible for the quality of the datasets they use within their applications. For example, a semantic web application might offer the possibility of user feedback to signalize facts, which need to be revised. Then, detected errors could be shared with the original data publisher and other users of the dataset. Both would be

DOI: 10.4018/IJSWIS.2015010102
able to correct the identified defects. While the need of error correction and data cleansing has reached the interest of the Linked Data community there exists no generally accepted method to expose, advertise, and retrieve suitable updates for Linked Data resources. In order to reuse curation efforts and to realize the vision of a collaborative method for error detection and effective exchange of corresponding corrections the following requirements have to be considered:

1. The description of defects and their corresponding fixes for Linked Data resources should be facilitated combined with various criteria, e.g. the scope of a fix, provenance information, and the type of defect to select fixes efficiently;
2. The realization of an appropriate workflow that covers guidelines to publish detected errors has to notify the original publishers as well as other users of a particular dataset. To encourage acceptance the application of updates has to be as convenient as possible;
3. Quality improvements for Linked Data resources should also be published as Linked Data to ease their exchange and to make them available for rating, discussions, and reuse.

In this paper we propose an approach that allows users to report Linked Data change requests (patches) within datasets and respective data publishers to effectively process such reports in order to pick up improvement suggestions from the community. The approach consists of the PatchR ontology, a framework implementation, and an appropriate workflow.

The arguments for the presented framework start with an overview of related work in the area of Linked Data curation. Next, the overall workflow of requesting Linked Data changes is explained. It allows to expose, rate, and select updates for particular Linked Data resources with a specialized ontology that is described in detail thereafter. Then, the internals of the framework and general usage guidelines are discussed in more detail. The feasibility and technical opportunities of this approach are illustrated exemplary for large knowledge bases, such as DBpedia, where flaws have been detected with the help of human users, in particular with a collaborative data cleansing game (WhoKnows?) and a fact ranking tool (FRanCo), as well as heuristic data cleansing tools (namely RDFUnit and SDType), for single file Linked Data publications, and for ontology evolution scenarios. The created patches are exposed and shared using the herein described ontology. Finally, the conclusion of the paper and an outlook on future work is given.

RELATED WORK

In order to raise quality in Linked Data published on the web multiple efforts aim to assure data consistency. On the one hand tools have been developed to identify erroneous data mainly on syntactic level and on the other hand an increasing number of efforts concentrate on the correction of broken or incomplete data. In this section, first related work on Linked Data validation and error detection is discussed, followed by a discussion of recent efforts on Linked Data correction and enhancement.

Validation

Various work has already focused on checking syntactical and logical data consistency by providing validators for the Semantic Web languages RDF and OWL, e.g. W3C’s RDF Validator, Vapour, and OWL Validator. Poor data quality in Linked Data has been identified, as for instance Hogan et al. analyzed typical errors and set up the RDF:Alerts Service that detects syntax and datatype errors to assist Linked Data publishers (Hogan, Harth, Passant, Decker, &
Related Content

Semantics Discovery via Human Computation Games
[www.igi-global.com/article/semantics-discovery-via-human-computation/62561?camid=4v1a](www.igi-global.com/article/semantics-discovery-via-human-computation/62561?camid=4v1a)

Towards Controlled Natural Language for Semantic Annotation
Brian Davis, Pradeep Dantuluri, Siegfried Handschuh and Hamish Cunningham (2010). *International Journal on Semantic Web and Information Systems* (pp. 64-91).
[www.igi-global.com/article/towards-controlled-natural-language-semantic/50499?camid=4v1a](www.igi-global.com/article/towards-controlled-natural-language-semantic/50499?camid=4v1a)
Adaptive Hybrid Semantic Selection of SAWSDL Services with SAWSDL-MX2
[www.igi-global.com/chapter/adaptive-hybrid-semantic-selection-sawsdl/64021?camid=4v1a](www.igi-global.com/chapter/adaptive-hybrid-semantic-selection-sawsdl/64021?camid=4v1a)

A Preventive Approach for Consistent OWL 2 DL Ontology Versions
[www.igi-global.com/article/a-preventive-approach-for-consistent-owl-2-dl-ontology-versions/217013?camid=4v1a](www.igi-global.com/article/a-preventive-approach-for-consistent-owl-2-dl-ontology-versions/217013?camid=4v1a)