User Assisted Creation of Open-Linked Data for Training Web Information Extraction in a Social Network

Martin Necasky
Charles University, Czech Republic

Dominik Fiser
Charles University, Czech Republic

Ivo Lasek
Charles University, Czech Republic

Ladislav Peska
Charles University, Czech Republic

Peter Vojtas
Charles University, Czech Republic

EXECUTIVE SUMMARY

In this chapter we describe our project under development and proof of concept for creating large Open-Linked Data repositories. The main problem is twofold: (1) Who will create (annotate) Open-Linked Data and in which vocabularies? (2) What will be the usage and profit of it?

For the first problem we propose several procedures on how to create Open-Linked data, including assisted creation of annotations (serving as base line or training set for Web Information Extraction tools), employing the social network, and also specific approaches to creating Open-linked data from governmental data resources. We describe some cases where such data can be used (e.g., in e-commerce, recommending systems, and in governmental and public policy projects).

INTRODUCTION

Increasingthe Web size and automation of its processing is a challenge for the IT community. There are several approaches on how to tackle this problem. Most remarkable are, of course, search engines. We would like to go beyond key word search and also support web scale applications–services, based on semantics added to web pages (e.g., in a form of RDFa annotations). So that technology and standards are ready, we describe how to start the process.

We focus on two aspects of this problem. The first problem is mainly sociological. To enable machines to understand web pages like humans, an initial human effort is necessary. The problem is: who (and maybe also why, how, when, where, etc.) will create semantic content? There are several possibilities, and we will discuss some of them. One possibility is to convince publishers to annotate their web resources by some vocabulary (ontology). A big impulse for this is the schema.org and sitemaps.org initiatives of Bing, Google, and Yahoo! with aim to improve search. Large human effort is/was invested into Wikipedia and Linked data are already extracted to DBPedia. Our approach is to use a social network and a specialized tool for third party annotation of web resources. We also touch on the problem of vocabulary for annotation. Our system enables both to create its own vocabulary and to use shared vocabularies such as schema.org andGoodRelations.The second problem is assessing what will be the use and profit of such Open-Linked data. We describe several use cases.

In this chapter we describe our project under development and a proof of concept for creating large Open-Linked Data repositories and applications which use them.

For the first problem, we propose a user assisted creation of a base line of Open-Linked Data. Motivation for doing this is supported by a social network. We present a tool enabling this. Further, this base line can be used for training Web Information Extraction tools.

We describe some cases where such data can be used (e.g., in aggregated web shops, news recommendations, and in governmental and public policy projects).

HUMAN ASSISTED CREATION OF SEMANTIC CONTENT

As we already mentioned in the introduction, the main problem of Semantic Web (web of data) is a sociological problem (and only afterwards it is a managerial problem and then also a technological problem). The problem is: Who (and also why, how, when, where, etc.) will create semantic content? This is the main goal of our project of Web Semantization (Dědek, Eckhardt, & Vojtáš, 2009). Here, web semantization is understood as a process of gradual enrichment of the web