Dealing with the Messiness of the Web of Data

Stefan Schlobach and Craig A. Knoblock

Research on the Semantic Web, which is now in its second decade, has been successful in encouraging people to publish data on the Web in structured, linked, and standardized ways. The success of what has now become the Web of Data can be seen from the sheer number of triples available within the Linked Open Data, Linked Life Data and Open Government initiatives. However, this growth in data makes many of the established assumptions inappropriate and offers a number of new research challenges.

In stark contrast to early Semantic Web applications that dealt with small, hand-crafted ontologies and data-sets, the new Web of Data comes with a plethora of contradicting world-views and contains incomplete, inconsistent, incorrect, fast-changing, and opinionated information. This information not only comes from academic sources and trustworthy institutions, but is often community built, scraped or translated. In short, the data is messy and difficult to use. This special issue is devoted to this messiness and how to deal with it. The approaches in this paper can broadly be classified into two classes: first, to provide guidelines or best practices for avoiding the messiness in the first place, and secondly, by giving users an infrastructure and techniques for building useful applications in spite of the messiness.

The paper "Emerging practices for mapping and linking life sciences data using RDF - a case series" (by Marshall, Boyce, Deus, Zhao, Willighagen, Samwald, Pichler, Hajagos, Prud’hommeaux and Stephens) is an excellent example of the first type. The authors introduce a general data workflow for mapping health care and life science (HCLS) data to RDF, and linking it to other data sources. Along with this workflow comes a number of recommended best practices for creating and publishing Linked Data in the HCLS domain. Although the workflow and practices are thoroughly evaluated in four specific case studies out of this particular domain, the impact of this paper is far wider, as the findings are applicable and relevant to many other application domains. For practitioners interested in data publication, this paper will be an excellent starting point and can be instrumental in producing high quality Linked Data that can be effectively consumed by others on the Web of Data.

Although the importance of such guidelines and best-practices is commonly acknowledged, their practical impact is difficult to measure and remains an object of study itself. "An empirical survey of Linked Data conformance" (by Hogan, Umbrich, Harth, Cyganiak, Polleres and Decker) goes to the core of publishing Linked Data and takes some well accepted concrete guidelines as the starting point of an empirical study of the relation between compliance to those guidelines and impact. For a huge corpus of several billion triples from over 800 data providers, they study conformance to 14 individual guidelines and then relate this conformance to the PageRank of the respective data providers. Apart from individual results for each guideline and provider, there are some important generic findings: such as a general adherence to the core principles (e.g. of dereferenceability and HTTP URIs), but also that there might be very good reasons for occasional nonconformance. More generally, there is no general correlation between guidelines and PageRank in all cases but rather for particular guidelines, such as external linkage
or vocabulary re-use. While simple conformance to guidelines is not the only descriptor of the quality of a data set on the Web of Data, this paper clearly shows that there are strong, though complex, relations between the two.

The final paper, "Integrating Open Government Data with Stratosphere for more Transparency" (by Heise and Neumann) falls into the second category of how people deal with the messiness of data. The authors start out from the very concrete problem that the heterogeneity of Open Government Data hinders meaningful search, analysis and integration. To overcome this problem they develop new operators for data cleaning for the Stratosphere data analysis framework, allowing them to integrate several well-known sources and data-sets at technical, structural, and semantic levels. An interesting aspect touched in this paper is the fact that the messiness of the Web of Data is not only in its heterogeneity and data quality, but also in its size. The author’s answer to the scalability of integration of messy data is parallelization using a generalization of Map/Reduce.

Common to all papers in this special issue is that human experts retain crucial roles in dealing with messiness: such as the data journalist having to clean messy data before being able to analyze it (as in the paper by Heise and Neumann), or the authors of guidelines who support the development of high quality data with their best practices (as in the papers by Marshall et al.). The contribution of Hogan et al. links these two ends of the spectrum by showing both limitations and potential of such guidelines in the context of potential applications.

The papers in this special issue discuss complementary approaches to dealing with the messiness of the Web of Data. There is no magic wand turning messy into high-quality data, but there are now more and more well understood and broadly accepted best practices which can guide data publishers towards better data. Nevertheless, the Web of Data will retain some of its current properties, such as heterogeneity and scale, and users who want to use this rich source of information will have to cope with its messiness in the future.