Enabling the Web of (Linked Open) Data

Epaminondas Kapetanios(✉)

Faculty of Science and Technology, University of Westminster, London, UK
e.kapetanios@westminster.ac.uk

Abstract. In this tutorial, we will take a look at the Web of Data and the Linked Open Data (LOD) project, in particular, via the lenses of "emergent semantics", which is quintessential for a much sought after semantic interoperability in highly dynamic environments. We also review the Web of Data via the property of self-organisation, which is an essential property of emergent semantics systems. This approach is promising to tackle some of the challenges in relation with successful publishing, reusability and linking of open data on the Web, since data sets and their descriptions in such a dynamic environment are continuously evolving and, therefore, need to be explored and searched via approximate querying, pattern search and similarity functions.

Keywords: Web of data · Linked Open Data · Semantic interoperability · Semantic web · Pattern search · Information retrieval · Emergent semantics · Distributed semantics · Self-organisation

Synopsis

The World Wide Web changed dramatically the way we attempt to share knowledge by lowering the barrier to publishing and accessing documents as part of a global information space. This functionality has been primarily enabled by the generic, open and extensible nature of the Web, which is also a key feature in the Web's uncompromised growth. Recently, the Web also evolved from a global information space of linked documents to one where not only documents but data are linked too. Underpinning this evolution is a set of best practices for publishing and connecting structured data on the Web known as Linked Data. This, in turn, has led to viewing the Web as a global data space connecting data from diverse domains such as people, companies, books, scientific publications, films, music, television and radio programmes, genes, proteins, drugs and clinical trials, on-line communities, statistical and scientific data, and reviews.

The openness of the Web and the rise in numbers of linked datasets, however, created further issues with (re-)usability, as well as quality, performance, reliability of the infrastructure in the linked data ecosystem. To this extent, automating certain tasks, such as discovery, selection and optimisation, becomes more and more important as it is not enough anymore to argue that URIs and RDF are all one needs to explore the linked datasets. The possible links that can be followed from a starting URI raises both performance and trust issues. In addition, the dynamics of the data-sources

P. Cimiano et al. (Eds.): ICWE 2015, LNCS 9114, pp. 728–729, 2015.
DOI: 10.1007/978-3-319-19890-3_67

also has an impact on the discovery, selection and performance of crawling collections of datasets.

For instance, in order to find the right dataset and to make this dataset accessible for biologists, the developer has to go through the process of locating a dataset that contains information relevant to biologists' research interests, such as information about a specific organism, or more specifically, genomic information about a particular organism. Subsequently, find out how this dataset can be programmatically accessed, as an RDF dump, through SPARQL endpoint, or any other protocol. Mostly important though is to understand the content of the dataset in order to perform an alignment with other datasets. Moreover, a data consumer may have discovered several datasets as a result of an indexer query. The question then arises how to select appropriate datasets from this list of potential candidates, with the emphasis on how to define "appropriateness" along the dimensions of contents, interlinking with other data sets and vocabularies being used.

In this context, the tutorial will provide an overview of the current Linked Open Data (LOD) stack of technologies, with particular emphasis on search engines and technologies tailored to alleviate the task of discovery and selection of appropriate data sets for reusability and linking on the Web. The tutorial, however, will also take a look at the Web of Data (LOD project) via the lenses of "emergent semantics", which is quintessential for a much sought after semantic interoperability in dynamic environments, as well as self-organisation, which is an essential property of emergent semantics systems, and how this view has been embraced by search engines and technologies.

Alongside these considerations, the tutorial is built upon the principles of emergent semantics, e.g., semantic handshaking protocol, evolution from local interactions and agreements towards global ones, and semantic self-organisation, with emphasis on examples found in science and nature, e.g., magnetisation in Physics, or examples from biology and chemistry including the stripped patterns in Zebras, Fish and the ocular dominance columns of the brain. These patterns are produced due to the individual responses of the cells to local conditions and the response of the neighbouring cells. It will also build upon the principles of pattern search based information retrieval with the focus on how these apply to the highly dynamic environment of the Web of (Linked Open) Data for the sake of data reusability and interoperability. Particular emphasis will be given on patterns among high-level features in order to bridge the semantic gap in search.

Some Useful Links

1. http://www.semantic-web.at/LOD-TheEssentials.pdf
2. http://bit.ly/open-data-map
3. http://datacatalogs.org
4. http://okfn.org
5. http://www.opengovdata.org/home/8principles
6. http://opengovernmentdata.org