Abstract

The Web is evolving into a global space with more and more content published with semantic markup. Recent studies show that almost 20% of the web pages contain some structured data. Government agencies are releasing datasets, in heterogeneous data formats which generally make hard to reuse or to interlink with other datasets. Linked Data (LD) is a promising set of technologies aiming at harmonizing a common graph-based data model (Resource Description Framework or RDF) and access protocol. Datasets published in RDF can be easily interconnected, discovered and consumed. The process of publishing and consuming datasets as L(O)D requires different techniques and tools which are not always familiar to publishers and lay users. Therefore, there is a need for a comprehensive framework and set of guidelines to help government agencies to publish their data into RDF in order to be fully compliant with the Linked Data principles. In this thesis, we focus on geospatial datasets where we aim to support the French National mapping agency (IGN) to publish geospatial data in the Semantic Web.

IGN produces different but complementary geographic vector reference databases delivered in traditional GIS formats. However, linked data consumers have different expectations and habits, such as the need to browse and explore RDF data using the "follow-your-nose" principle. Besides, traditional GIS data formats are not interoperable with RDF. Yet, all these geographic datasets could be used with benefits on the Web of data, either with direct georeferencing through geographic primitives, or indirectly through postal addresses. We have contributed to the georeferencing of datasets published on the Web of data by providing such resources for the French territory. Firstly, we propose two vocabularies designed for representing structured geometries defined with coordinates expressed in any Coordinates Reference System (CRS). Secondly, we reuse these vocabularies and the CRSs' dataset to publish a reference dataset on administrative units that can also be reused for indirect georeferencing purposes. Finally, we also propose two vocabularies for describing geographic feature types. In addition to these resources, we also present a comprehensive workflow for easily publishing geographic data on the Web of data. We interlink those datasets with other popular linked datasets in the geographical domain and we describe extensive experiments and evaluation results regarding the problems for interlinking such data.

Datasets published in the LOD cloud can often be accessed by different means such as API access, bulk download or as linked data fragments, while most of the time, a SPARQL endpoint is also provided. While the LOD cloud keeps growing, having a quick glimpse of those datasets is getting harder and harder, and there is a need to develop new methods enabling to detect automatically what an arbitrary dataset is about in order to recommend visualizations for excerpts of data. We consider that *"a visualization is worth a million triples"*, and we propose a novel approach that mines the content of datasets and automatically generates visualizations. Our approach

is directly based on the usage of SPARQL queries that will detect the important categories of a dataset and that will specifically consider the properties used by the objects which have been interlinked via `owl:sameAs` links. We then propose to associate type of visualization for those categories. We have implemented this approach into a so-called Linked Data VIzualization Wizard (LDVizWiz).
Finally, it is widely accepted that by controlling metadata, it is easier to publish high quality data on the web. Metadata, in the context of Linked Data, refers to vocabularies and ontologies used for describing data. With more and more data published on the web, the need for reusing controlled taxonomies and vocabularies is becoming more and more a necessity. Catalogues of vocabularies are generally a starting point to search for vocabularies based on search terms. Some recent studies recommend that it is better to reuse terms from "popular" vocabularies [1]. However, there is not yet an agreement on what makes a popular vocabulary since it depends on diverse criteria such as the number of properties, the number of datasets using part or the whole vocabulary, etc. We propose a method for ranking vocabularies based on information content metrics applied to the Linked Open Vocabulary (LOV) catalogue.