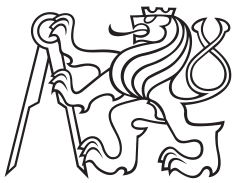
Master's thesis



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science and Engineering

Platform for semantic extraction of the web

Jakub Podlaha

January 2015

/ Declaration

Prohlašuji, že jsem se neflákal.

Abstrakt / Abstract

Tento dokument je pouze pro potřeby testování.

Překlad titulu: Platforma pro sémantickou extrakci webu

This document is for testing purpose only.

Contents

Tables /

Chapter 1
Knowledge base, principles and technologies

1.1 Technology of Semantic Web

Wikipedia defines Semantic Web as a collaborative movement led by international standards body the World Wide Web Consortium (W3C). [??] W3C itself defines Semantic Web as a technology stack to support a `Web of data,` as opposed to `Web of documents,` the web we commonly know and use (XXX `http://www.w3.org/standards/semanticweb/`). Just like with `Cloud` or `Big Data` the propper definition tends to vary, but the notion remains the same. It is collaborative movement led by W3C and it does define a technology stack. It also includes users and companies using this technology and the data itself. Technologies and languages of Semantic Web such as RDF, RDFa, OWL, SPARQL (XXX) are well standardized and will be described in following chapters.

As a general logical concept of the technology... (XXX) Technology of Semantic Web is used to take data and metadata, give them unique identifiers and form them into oriented graphs. The metadata part define a schema of types and properties that can be assigned to data and also relations between this types and properties possibly in a form of ontology. When some data are anotaded by resources from such an ontology we gain power to reason on this data, i.e. resolve new relations based on known ones, and also to query on our data along with any data annotated using the same ontology.

In RDF(S) the is defined in a form of triples. Triple consists of subject, predicate and object, wich all are simply `resources` listed by their identifiers. In this very general form we can express basically any relationship between two resources. On a level of classes and properties, we can for example assign a type to an individual, or set a class as a domain of some property. On a level of ontologies we can specify author and date it was released. (XXX DELME)

1.2 RDF and RDFS

RDF is a family of specifications for syntax notations and data serialization formats, meta data modeling, and vocabulary used for it.

XXX `https://en.wikipedia.org/wiki/Resource_Description_Framework`

We will look closely on URI, the resource identifier, vocabularies and semantics defined by RDF, RDFS, and OWL, and serialization into Turtle and RDF/XML formats.

1.2.1 URI

In order to give each resource an unique identifier a Uniform Resource Identifier is used. This is mostly in a form of URL as we commonly know it as `web address` (e.g. http://www.example.org/some/place#something). In some cases URI can be a URN as well. URN is a complementary syntax for URL that allow us to identify resources without specifying their location. This way we can for example use ISBN codes when

working with books and records, or UUID identifier a Universally Uniqe Identifier widely used to identify technically any data instance.

XXX `https://en.wikipedia.org/wiki/Uniform_resource_identifier`

1.2.2 RDF and RDFS vocabulary

In order to work with data properly (XXX) RDF(S) vocabulary defines several basic URIs along with their semantics.

resource	description
rdf:type	a property used to state that a resource is an instance of a class; a commonly accepted qname for this property is `a`
rdfs:Resource	the class of everything; all things described by RDF are resources.

Table 1.1. Pocet absolventu FEL CVUT. Tabulka je prevzata zË-[??].

rdfs:Class - declares a resource as a class for other resources.

rdfs:Literal – literal values such as strings and integers. Property values such as textual strings are examples of RDF literals. Literals may be plain or typed. rdfs:Datatype – the class of datatypes. rdfs:Datatype is both an instance of and a subclass of rdfs:Class. Each instance of rdfs:Datatype is a subclass of rdfs:Literal. rdf:XMLLiteral – the class of XML literal values. rdf:XMLLiteral is an instance of rdfs:Datatype (and thus a subclass of rdfs:Literal).

rdf:Property – the class of properties. rdfs:domain of an rdf:predicate declares the class of the subject in a triple whose second component is the predicate. rdfs:range of an rdf:predicate declares the class or datatype of the object in a triple whose second component is the predicate.

rdfs:subClassOf allows to declare hierarchies of classes. rdfs:subPropertyOf is an instance of rdf:Property that is used to state that all resources related by one property are also related by another.

rdfs:label is an instance of rdf:Property that may be used to provide a human-readable version of a resource's name. rdfs:comment is an instance of rdf:Property that may be used to provide a human-readable description of a resource.

These are the basic building blocks of our future RDF graphs. The semantics defined in the specification and slightly described here alow us to specify class hierarchy, properties with domain and range as well as use this structure on individuals and literals.

1.3 OWL

- `http://www.w3.org/TR/owl2-primer/`
- `https://en.wikipedia.org/wiki/Web_Ontology_Language`
- `http://www.w3.org/TR/2012/REC-owl2-quick-reference-20121211/`

1.4 Linked Data

Wikipedia defines Linked Data as `a term used to describe a recommended best practice for expo`

- `http://linkeddata.org/guides-and-tutorials`
- `http://linkeddatabook.com/editions/1.0/`
- `http://lov.okfn.org/dataset/lov/`

1.5 Ontology repositories

- http://www.w3.org/wiki/Ontology_repositories

1.6 RDFa

- https://www.sio2.cz/web/psiotwo/publications
- http://rdfa.info/play/

1.7 dalsi

- https://en.wikipedia.org/wiki/SPARQL
- https://en.wikipedia.org/wiki/Turtle_(syntax)

1.8 automatická extrakce dat

TODO in next section