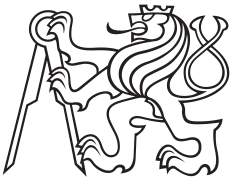
Diplomová práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra kybernetiky

Minimální dokument

Jakub Podlaha

November 2013

/ Prohlášení

Prohlašuji, že jsem se neflákal.

Abstrakt / Abstract

Tento dokument je pouze pro potřeby testování.

This document is for testing purpose only.

/ Obsah

Kapitola 1
Zadání SW Projektu

1. Seznamte se technologiemi pro automatickou extrakci dat z webových stránek a s jazyky sémantického webu RDF, RDFS a OWL.
2. Navrhněte a implementujte vhodný datový formát pro popis scénářů extrakce dat, které bude možné zpracovat vhodným open-source crawlerem (např. [1]). Vytvořte jednoduché uživatelské rozhraní ve vhodném webovém prohlížeči, sloužící k tvorbě scénářů ve vámi navrženém datovém formátu pro následnou extrakci sémantických data z webových stránek.

Kapitola 2
Knowledge base, principles and technologies

Seznamte se technologiemi pro automatickou extrakci dat z webových stránek a s jazyky sémantického webu RDF, RDFS a OWL.

2.1 RDF and RDFS

- https://en.wikipedia.org/wiki/Resource_Description_Framework

2.2 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/

2.3 Linked Data

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

2.4 Ontology repositories

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

2.5 RDFa

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

2.6 dalsi

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

Kapitola 3
research - existující řešení

3.1 InfoCram 2000 - Jirka

- zalozeny na Aardwark [1])

3.2 iMacros

- `http://wiki.imacros.net/Command_Reference`
- `http://wiki.imacros.net/iMacros_for_Firefox`
- `http://wiki.imacros.net/iMacros_for_Chrome`

3.3 Sahi

Yet another web automation project. `http://sourceforge.net/projects/sahi/`

3.4 Selenium IDE

- IDE - `http://www.seleniumhq.org/projects/ide/`
- plugins - `http://www.seleniumhq.org/projects/ide/plugins.jsp`
- current commands - `http://release.seleniumhq.org/selenium-core/1.0.1/reference.html`
- documentation - `http://docs.seleniumhq.org/docs/index.jsp`

[1]) `https://addons.mozilla.org/en-US/firefox/addon/aardvark/`

Kapitola 4
crOWLer

4.1 zavislosti

- maven - apache project managing tool
  - https://maven.apache.org
  - https://maven.apache.org/run-maven/index.html
  - https://maven.apache.org/guides/mini/guide-ide-eclipse.html
- sesame
  - http://www.openrdf.org/download.jsp ??
- jena
  - https://github.com/ansell/JenaSesame !!
  - or https://github.com/afs/JenaSesame ??
  - or http://jena.apache.org/ ???
  - or http://sjadapter.sourceforge.net/ ????
  - or http://sourceforge.net/projects/jenasesamemodel/
  - might help http://www.iandickinson.me.uk/articles/jena-eclipse-helloworld/
  - little hint http://spqr.cerch.kcl.ac.uk/?page_id=130
  - another hit http://answers.semanticweb.com/questions/20865/how-to-get-the-jena-sesame-adapter
  - wiki https://en.wikipedia.org/wiki/Jena_(framework)
  - jena vs. sesame flame http://answers.semanticweb.com/questions/1638/jena-vs-sesame-is-there-a-serious-complete-up-to-date-unbiased-well-informed-side-by-side-comparison-between-the-two

4.2 Implementation

4.2.1 Classes

- ImmovableHeritageConfiguration extends MonumnetConfiguration implements ConfigurationFactory
  - implements Configuration, which is parameter for FullCrawler.run() method
- FullCrawler
  - implements the whole crawling algorithm
  -

4.3 notes

- http://onto.mondis.cz/resource/page/npu/

4.3.1 Run configuration

```
  crowler cz.sio2.crowler.configurations.npu.ImmovableHeritageConfiguration
file jena_con
```

- Class ImmovableHeritageConfiguration implements Configuration class.
- Folder jena_con will be created and all the rdf's will be stored in int with names derived from ontology uri

Kapitola 5
Data

5.1 Pamatky

- http://monumnet.npu.cz/pamfond/list.php?hledani=1&KrOk=&HiZe=&VybUzemi=1&sNazSidOb=&Adresa=&Cdom=&Pamatka=&CiRejst=&Uz=B&PrirUbytOd=3.5.1958&PrirUbytDo=10.12.2013
- http://dominanty.cz/pamatky-cihana.php

Kapitola 6
Implementace

6.1 SelectOWL - Plugin pro Selenium IDE - Firefox

- https://developer.mozilla.org/en-US/Add-ons/Setting_up_extension_development_environme
- http://kb.mozillazine.org/Getting_started_with_extension_development
- http://code.google.com/p/selenium/source/browse/
- http://docs.seleniumhq.org/download/maven.jsp
- http://repo1.maven.org/maven2/org/seleniumhq/selenium/ide/selenium-ide/1.0.2/

6.2 Snippets