

Platform for semantic extraction of the web



Jakub Podlaha

Problem definition

- Design a tool for extracting data from web
 - Data in semi-structured form (HTML)
- Target structure in a form suitable for Semantic Web
- Focus on dynamic, simple solution
- Use Cases...

Problem definition



NÁRODNÍ
PAMÁTKOVÝ
ÚSTAV

MonumNet

Nemovitě památky

pro tisk: stránka celý výběr do Excelu: stránka celý výběr

Nalezeno: 40203 je chráněno, přírůsky od 03.05.1958 do 10.12.2013

Stránka 1 / 1609 ⇨ ⇩

1 2 3 4 5 6 7 8 9 10 11


Číslo rejstříku	uz	Název okresu	Sídelní útvar	Část obce	čp.	Památka	Ulice,nám./um
20339 / 1-1971	S	Praha hl.m.	Praha	Běchovice	čp.1	zájezdní hostinec Na Staré poště	Praha 9, Českobrodská
104764	P	Praha hl.m.	Praha	Benice		zvonička	
40604 / 1-1569	S	Praha hl.m.	Praha	Bohnice		kostel sv. Petra a Pavla	Praha 8, Bohnice
54973 / 1-1628	R	Praha hl.m.	Praha	Bohnice		výšinné opevněné sídliště - hradiště Zámka, archeologické stopy	Praha 8, na ostrohu nad Vltavou
54974 / 1-1571	S	Praha hl.m.	Praha	Bohnice	čp.1	venkovská usedlost Vraných	Praha 8, Bohnická
44366 / 1-1572	S	Praha hl.m.	Praha	Bohnice	čp.4	fara	Praha 8, Bohnická
54975 / 1-1573	S	Praha hl.m.	Praha	Bohnice	čp.12	činžovní dům - hospoda Štrasburk	Praha 8, Bohnická
40605 / 1-1570	R	Praha hl.m.	Praha	Bohnice	čp.91	nemocnice - psychiatrická léčebna	Praha 8, Ústavní, Bohnická
44368 / 1-1347	S	Praha hl.m.	Praha	Braník		kostel sv. Prokopa	Praha 4, Školní, Nad kostelem
44369 / 1-1713	S	Praha hl.m.	Praha	Braník	čp.15	Maroldova vila	Praha 4, Stará cesta

```

</rdf:Description>
<rdf:Description rdf:about="http://kub1x.org/onto/dip/t/npu-201412300182342414/indiv201412300182458207">
  <rdf:type rdf:resource="http://onto.mondis.cz/resource/npu/District"/>
  <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">R</rdfs:label>
</rdf:Description>
<rdf:Description rdf:about="http://kub1x.org/onto/dip/t/npu-201412300182342414/indiv201412300182502529">
  <rdf:type rdf:resource="http://onto.mondis.cz/resource/npu/MonumentRecord"/>
  <j.0:hasDistrict rdf:resource="http://kub1x.org/onto/dip/t/npu-201412300182342414/indiv201412300182502774"/>
</rdf:Description>
<rdf:Description rdf:about="http://onto.mondis.cz/resource/npu/District">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
</rdf:Description>

```

Problem definition

 NÁRODNÍ
PAMÁTKOVÝ
ÚSTAV

MonumNet

Nemovitě památky

pro tisk: stránka celý výběr do Excelu: stránka celý výběr

Nalezeno: 40203 je chráněno, přírůsky od 03.05.1958 do 10.12.2013

Stránka 1 / 1609 ⇌ ↺

1 2 3 4 5 6 7 8 9 10 11

Číslo rejstříku	uz	Název okresu	Sídelní útvar	Část obce	čp.	Památka	Ulice,nám./um
20339 / 1-1971	S	Praha hl.m.	Praha	Běchovice	čp.1	zájezdní hostinec Na Staré poště	Praha 9, Českobrodská
104764	P	Praha hl.m.	Praha	Benice		zvonička	
40604 / 1-1569	S	Praha hl.m.	Praha	Bohnice		kostel sv. Petra a Pavla	Praha 8, Bohnice

54973 / 1-1628

<rdf:RDF

54974 / 1-1571

xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"

44366 / 1-1572

xmlns:j.0="http://onto.mondis.cz/resource/npu/"

54975 / 1-1573

xmlns:owl="http://www.w3.org/2002/07/owl#"

40605 / 1-1570

xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"

44368 / 1-1347

xmlns:xsd="http://www.w3.org/2001/XMLSchema#" >

44369 / 1-1713

<rdf:Description rdf:about="http://kublx.org/onto/dip/t/npu-201412300182342414/indiv201412300182458526">

44369 / 1-1713

<rdf:type rdf:resource="http://onto.mondis.cz/resource/npu/MonumentRecord"/>

44369 / 1-1713

<j.0:hasDistrict rdf:resource="http://kublx.org/onto/dip/t/npu-201412300182342414/indiv201412300182458765"/>

44369 / 1-1713

</rdf:Description>

44369 / 1-1713

<rdf:Description rdf:about="http://kublx.org/onto/dip/t/npu-201412300182342414/indiv201412300182458207">

44369 / 1-1713

<rdf:type rdf:resource="http://onto.mondis.cz/resource/npu/District"/>

44369 / 1-1713

<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">R</rdfs:label>

44369 / 1-1713

</rdf:Description>

44369 / 1-1713

<rdf:Description rdf:about="http://kublx.org/onto/dip/t/npu-201412300182342414/indiv201412300182502529">

44369 / 1-1713

<rdf:type rdf:resource="http://onto.mondis.cz/resource/npu/MonumentRecord"/>

44369 / 1-1713

<j.0:hasDistrict rdf:resource="http://kublx.org/onto/dip/t/npu-201412300182342414/indiv201412300182502774"/>

44369 / 1-1713

</rdf:Description>

44369 / 1-1713

<rdf:Description rdf:about="http://onto.mondis.cz/resource/npu/District">

44369 / 1-1713

<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>

44369 / 1-1713

</rdf:Description>

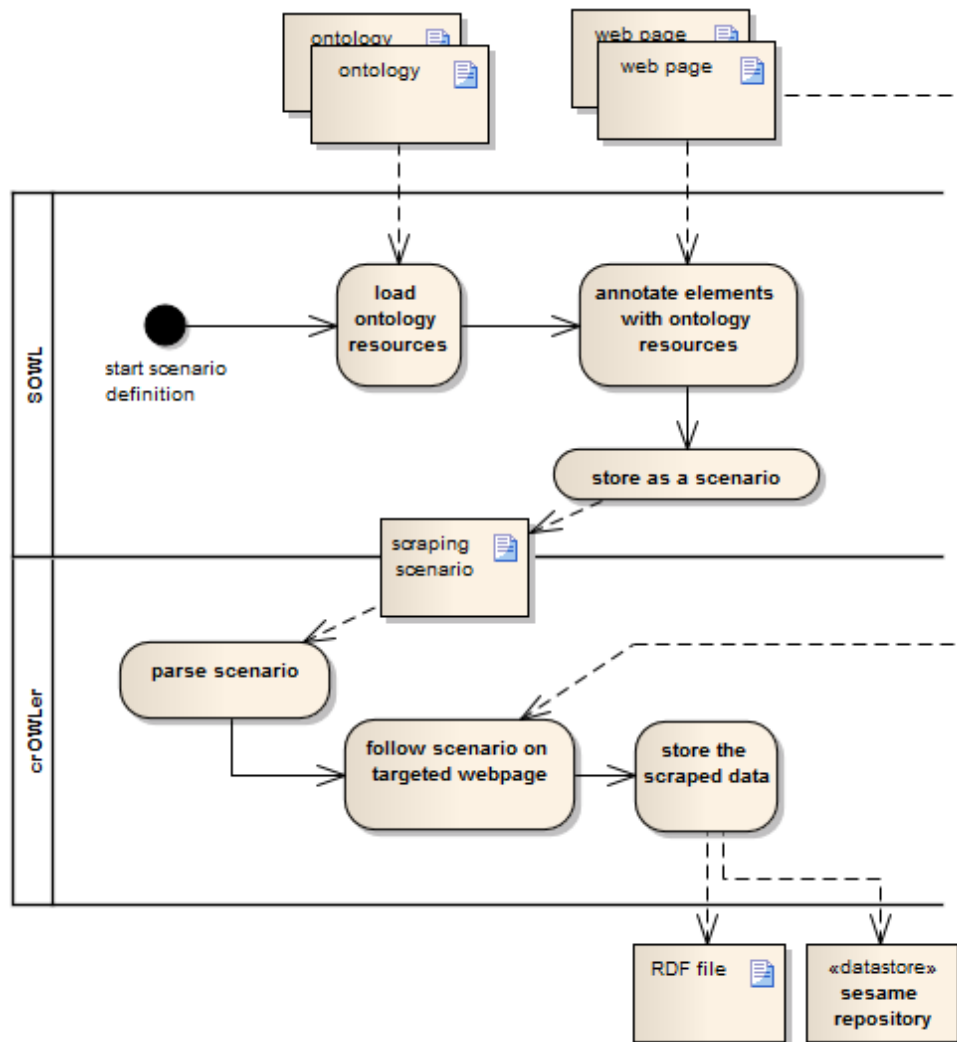
Existing solutions

- Rapid prototyping to explore suitable platforms and libraries
- Semantic
 - Strigil, crOWLer, jOWL, rdfquery
- Non-semantic
 - InfoCram 6000, Aardvark, Selenium IDE and Builder, WebDriver

Program design

- Components
 - SOWL – frontend, the tool for scenario creation
 - crOWLer – backend, the semantic crawler
- Scraping scenario
 - Syntax and semantics of scraping scenario commands
 - Mapping of ontological resources on DOM nodes
 - Elemental data handling
 - JavaScript support

Program workflow



Implementation

Accident Reports

www.nts.gov/investigations/AccidentReports/Pages/AccidentReports.aspx

sowl

Scenario: unnamed

Template: init

template init

create http://onto.mondis.cz/resource/npu/MonumentRecord

tbody tr.list

assign http://onto.mondis.cz/resource/npu/hasCHObjectNumber

td:nth-child(0)

create http://onto.mondis.cz/resource/npu/District

td:nth-child(2)

call template: init

td table a.ind:nth-child(0) @href

Filter: uri

http://onto.mondis.cz/resource/npu/hasCHObjectNumber

http://onto.mondis.cz/resource/npu/hasHouseNumber

http://onto.mondis.cz/resource/npu/hasLocation

http://onto.mondis.cz/resource/npu/hasLandRegistry

http://onto.mondis.cz/resource/npu/hasIdentifiableLandRe

http://onto.mondis.cz/resource/npu/hasLocalityName

http://onto.mondis.cz/resource/npu/hasEndangeredScope

http://onto.mondis.cz/resource/npu/hasRegistryNumber


http://onto.mondis.cz/resource/npu/District

http://onto.mondis.cz/resource/npu/hasDistrict

http://onto.mondis.cz/resource/npu/hasState

http://onto.mondis.cz/resource/npu/hasAnnotation

http://onto.mondis.cz/resource/npu/MonumentRecord



NATIONAL TRANSPORTATION SAFETY BOARD

Search this site... Search Site

Advanced Search

HOME NEWS & EVENTS SAFETY ADVOCACY INVESTIGATIONS DISASTER ASSISTANCE LEGAL ABOUT PUBLICATIONS

Home > INVESTIGATIONS > Accident Reports

SHARE

Accident Reports

Accident Reports are one of the main products of an NTSB investigation. Reports provide details about the accident, analysis of the factual data, conclusions and the probable cause of the accident, and the related safety recommendations. Most reports focus on a single accident, though the NTSB also produces reports addressing issues common to a set of similar accidents.

Most Recent Reports

Report Number	NTSB Title	Accident Date	Report Date	City	State	Country	Other	Report
MAB1422	Fire on Board Towing Vessel <i>Shanon E. Settoon</i>	3/12/2013	12/10/2014	Bayou Perot	LA	USA	29°38.03 N, 90°10.63 W	PDF
RAB1414	Collision of BNSF Railway Company and Union Pacific Railroad Trains Near Keithville, Louisiana	12/30/2012	12/1/2014	Keithville	LA			PDF
AIR1401	Auxiliary Power Unit Battery Fire Japan Airlines Boeing 787-8, JA829J	1/7/2013	11/21/2014	Boston	MA			PDF
AAR1404	Crash Following In-Flight Fire Fresh Air, Inc. Convair CV-440-38, N153JR	3/15/2012	11/17/2014	San Juan	PR			PDF
RAR1402	Collision of Union Pacific Railroad Freight Train with BNSF Railway Freight Train	5/25/2013	11/17/2014	Chaffee	MO	USA		PDF
MAB1421	Marine Accident Brief: Breakaway of Tanker <i>Harbour Feature</i> from its Mooring and	4/1/2013	11/12/2014	Portsmouth	NH			PDF

Reports by Mode

- Aviation Accident Reports
- Hazardous Materials Accident Reports
- Highway Accident Reports
- Marine Accident Reports
- Pipeline Accident Reports
- Railroad Accident Reports

Outcomes and future work

- Improvement of current approach in semantic crawling
 - Definition of scraping scenario
 - Architecture and implementation of platform stack
 - Usability and accessibility for end users
- Full implementation of proposed design
- Usability improvements according to users feedback
- Better handling of ontological resources

