Strigil: A Framework for Data Extraction

Jakub Stárka

The Department of Software Engineering
Charles University in Prague
Malostranské nám. 25
118 00 Praha 1
Czech Republic

Overview

- Motivation
- Problems
- Strigil
- Scraping language
- Conclusion

Motivation

Web Crawling Problems

- Web is growing quickly
- Servers have limitated resources
  - Politeness policy is required
- Web is changing
  - And pages are more dynamic
- Specialized crawlers are needed
  - Limited to one domain

Motivation

- Analysis
  - Find new information
- Correction
  - Get data from different sources and correct them
- Aggregation
  - Find new trends
- Linking
  - Allow users to get context

Existing Works

- Scrapers
  - Well examined
  - Many solutions
  - Universal
  - Distributed
- Distributed scrapers
- Program driven scrapers
- „Home“ scrapers
  - Only for few pages with GUI

Other Tools

- Scraper wiki [1]
  - No GUI, scripts created in code editors
  - Ruby, Python, PHP
- Visual Web Ripper [2]
  - Commercial application
  - Human emulation, GUI editor
- Screen-Scraper [3]
  - Java, JavaScript, Python, VBScript
  - Limited Free version

Scraping Languages Overview

- Source format
  - XML, XHTML, HTML, ...
- Transformation language
  - XSL, XQuery, OXPath[4]
- Output format
  - XML
  - Relational DB
  - Objects

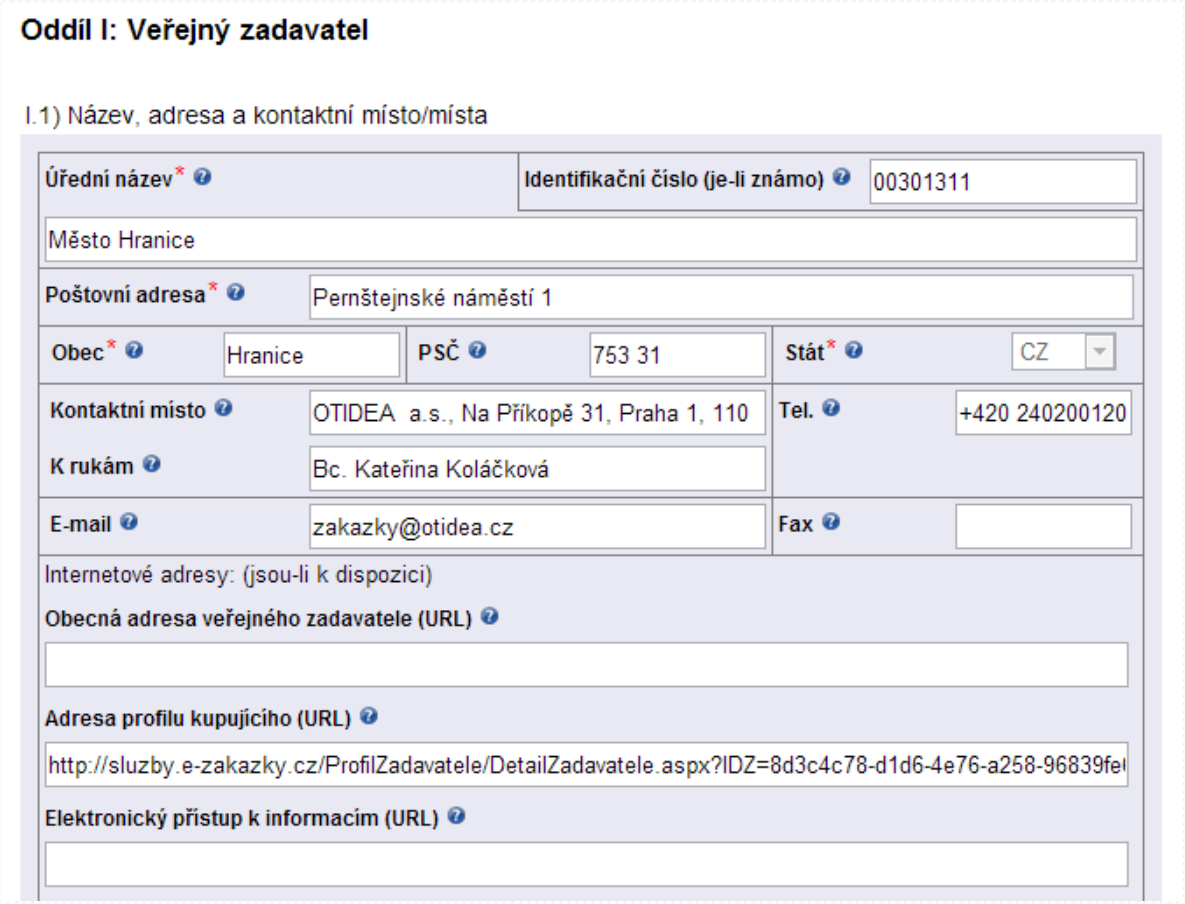
Problems

- Connection speed
  - Distributed solution
- Processing speed
  - JavaScript
- Bot blocking features
  - Human emulation and politeness

Problems

- Deep web
  - Web forms
  - JavaScript
  - AJAX
- Redirection
  - Bot recognition
- Cookies
  - Login
- Flexibility in data extraction

Web Page Fragment 1



Oddíl I: Veřejný zadavatel

I.1) Název, adresa a kontaktní místo/místa

Úřední název*			Identifikační číslo (je-li známo)	00301311	
Město Hranice					
Poštovní adresa*	Pernštejnské náměstí 1				
Obec*	Hranice	PSČ	753 31	Stát*	CZ
Kontaktní místo	OTIDEA a.s., Na Příkopě 31, Praha 1, 110			Tel.	+420 240200120
K rukám	Bc. Kateřina Koláčková				
E-mail	zakazky@otidea.cz			Fax	

Internetové adresy: (jsou-li k dispozici)

Obecná adresa veřejného zadavatele (URL)

Adresa profilu kupujícího (URL)

http://sluzby.e-zakazky.cz/ProfilZadavatele/DetailZadavatele.aspx?IDZ=8d3c4c78-d1d6-4e76-a258-96839fe

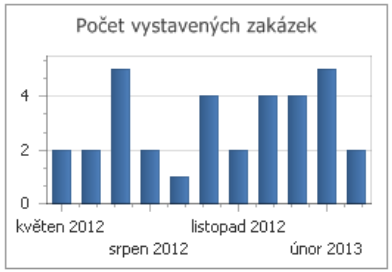
Elektronický přístup k informacím (URL)

- http://www.vestnikverejnychzakazek.cz

Web Page Fragment 2

PROFIL ZADAVATELE: MĚSTO ZNOJMO

Adresa	Obroková 10/12 669 02 Znojmo
IČ	00293881
DIČ	
E-Mail	info@muznojmo.cz
WWW	www.znojmocity.cz
Elektronické tržiště	
Evidenční číslo na ISVZUS	215901



Pro zobrazení detailnější statistiky klikněte na jeden z ukázkových grafů.

VYPSANÉ ZAKÁZKY - AKTIVNÍ

Název zakázky	Číslo zakázky	Specifikace zadávacího řízení	Typ veřejné zakázky	Zakázka před zahájením	Zakázka neukončená	Zakázka byla zadána
Adrenalinové trasy Znojemského podzemí	VZ2012-013-VYB-ISN	Zjednodušené podlimitní řízení	Stavební práce	0	0	2
Komplexní organizační a věcné zajištění zadávacích řízení na veřejné zakázky pro Město Znojmo a příspěvkové organizace	VZ2012-055-ANT-ISN	Otevřené řízení	Služby	1	26	0

DETAIL ZAKÁZKY

Název zakázky	Číslo zakázky	Počátek lhůty pro podání nabídek	Konec lhůty pro podání nabídek	Druh řízení	Typ veřejné zakázky	Status	Předpokládaná cena zakázky
Adrenalinové trasy Znojemského podzemí	VZ2012-013-VYB-ISN			Zjednodušené podlimitní řízení	Stavební práce	Probíhá	7 120 136,00 Kč

- http://www.e-zakazky.cz

Web Page Fragment 3

Oddíl I: Veřejný zadavatel

I.1) **Název, adresa a kontaktní místo/místa**
EVN Makedonija AD, Skopje
11 Oktomvri, Nr. 9
Kontaktní místo: Vesna Nushkova, Atanasovska Liljana, Mile Dabeski
1000 Skopje
BÝVALÁ JUGOSLÁVSKÁ REPUBLIKA MAKEDONIE
Tel.: +389 23205000/42097
E-mail: vesna.nushkova@evn.mk
Fax: +389 23205000/45936
Internetové adresy:
Obecná adresa veřejného zadavatele: http://www.evn.mk
Další informace lze získat: na výše uvedená kontaktní místa
Zadávací dokumentaci a další dokumenty (včetně dokumentů k soutěžnímu dialogu a k dynamickému nákupnímu systému) lze získat: na výše uvedená kontaktní místa
Nabídky nebo žádosti o účast musí být zaslány: na výše uvedená kontaktní místa

Oddíl II: Předmět zakázky

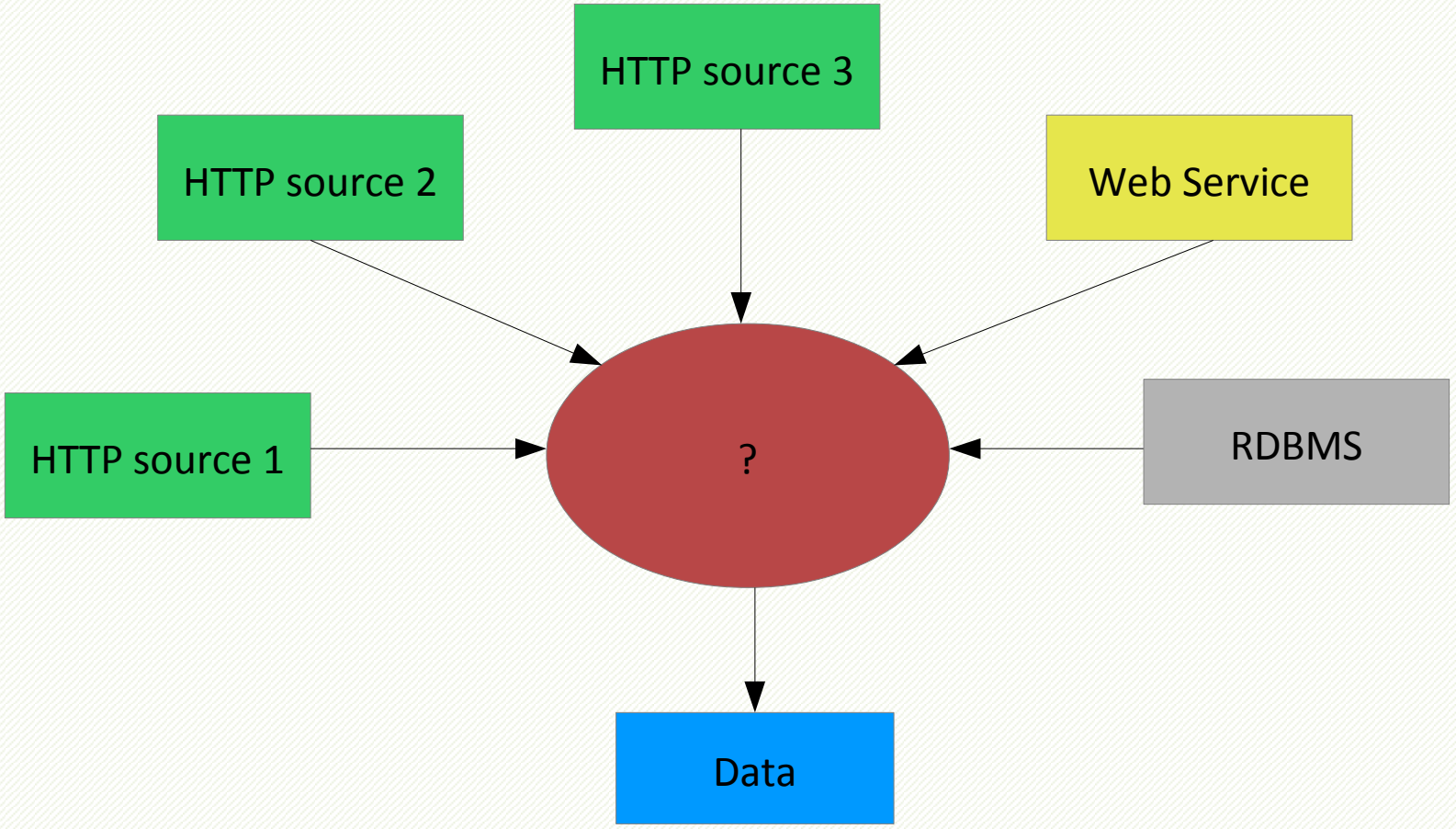
II.1) **Popis**

II.1.6) **Společný slovník pro veřejné zakázky (CPV)**
31681410

Popis
Elektrické materiály.

- http://ted.europa.eu

Data

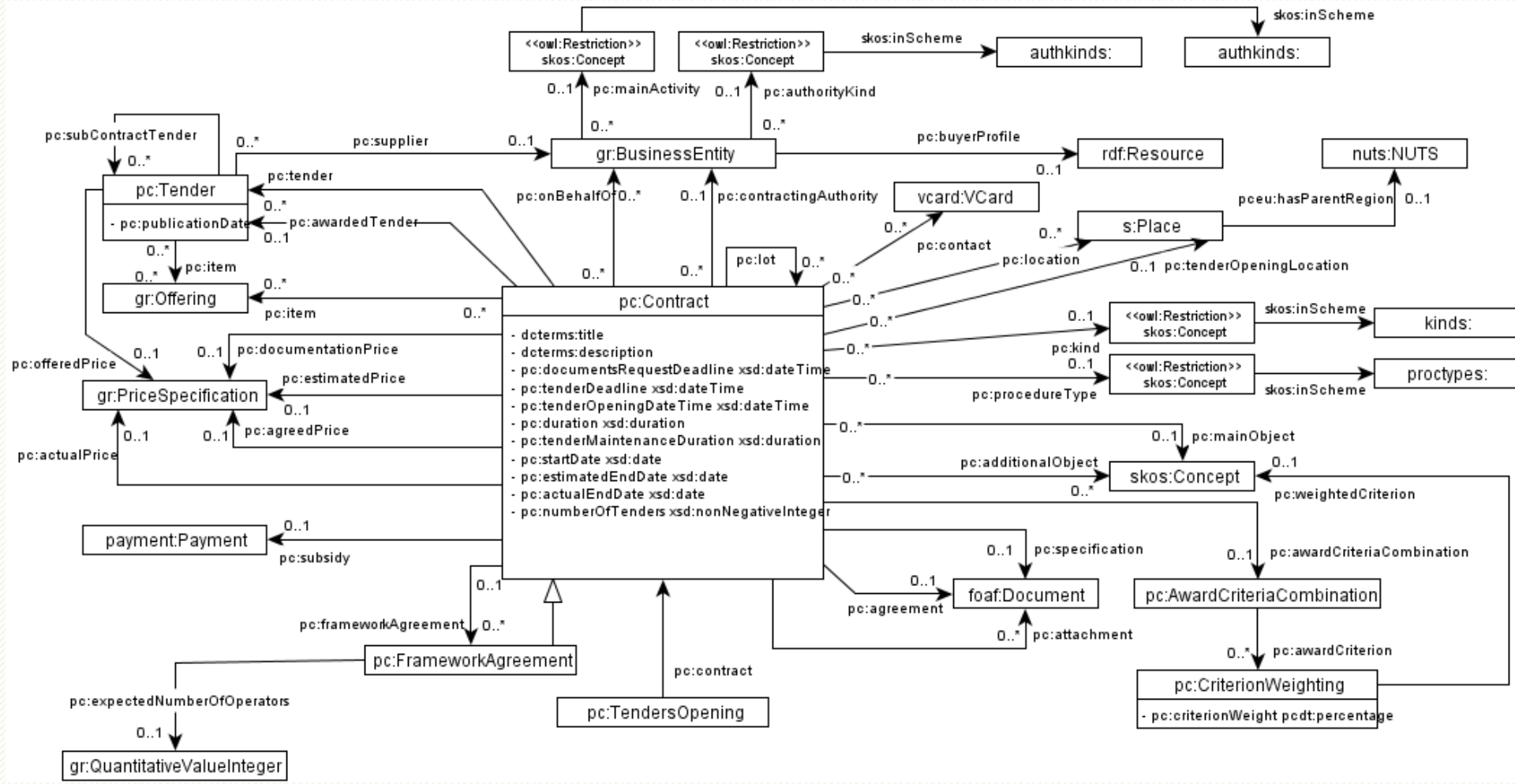

HTTP source 3
HTTP source 2
Web Service
HTTP source 1
?
RDBMS
Data

Strigil

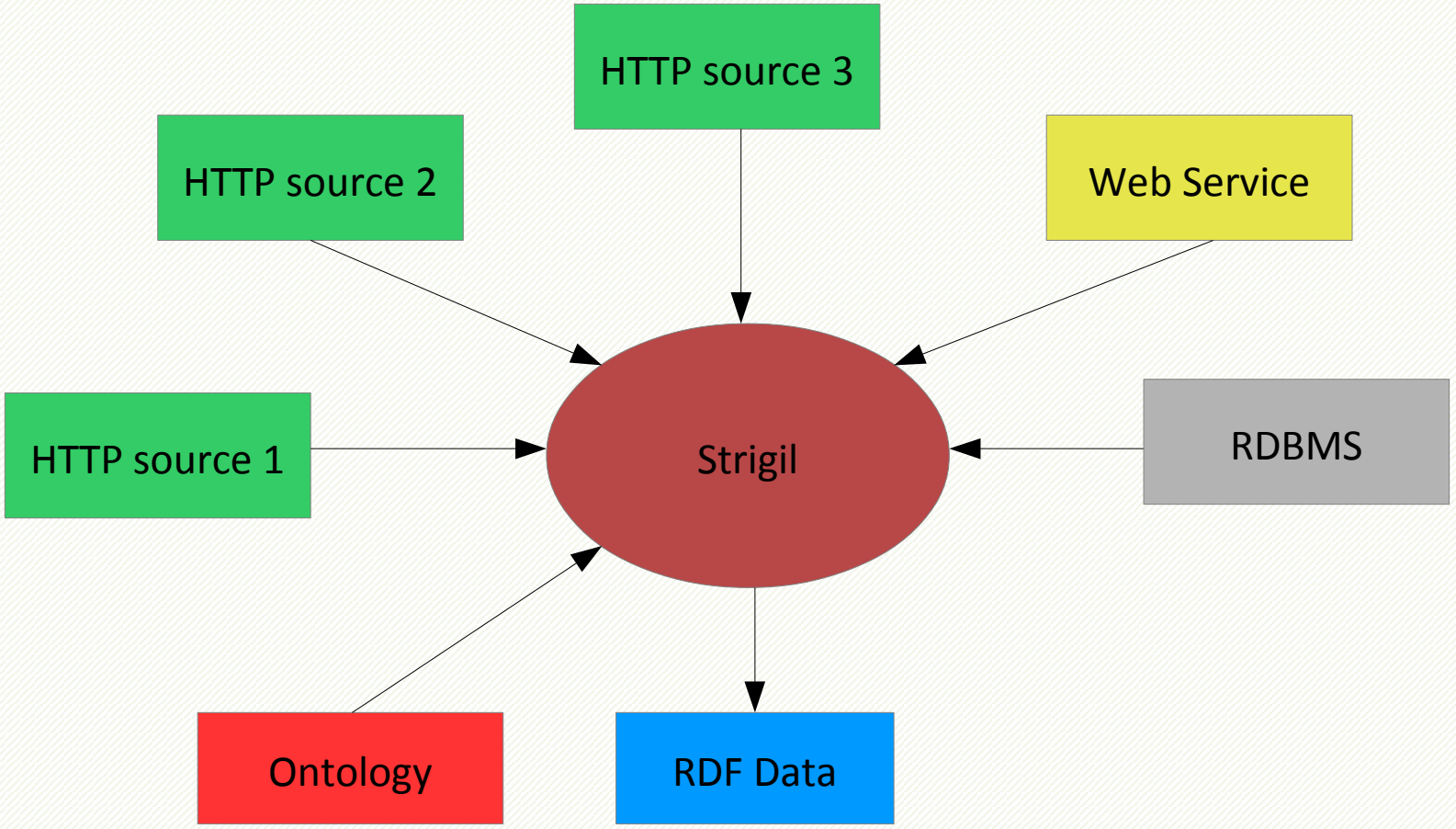
Strigil

- Scheduling
- Politeness
- Universal transformation
  - Variable selectors
- Scalability
  - Multiple downloaders
  - Proxy servers
- Ontology integration

Ontology


<<owl:Restriction>>
skos:Concept
<<owl:Restriction>>
skos:Concept
skos:inScheme
authkinds:
skos:inScheme
authkinds:
0..1
pc:mainActivity
0..1
pc:authorityKind
0..*
0..*
pc:subContractTender
0..*
0..*
pc:supplier
0..1
gr:BusinessEntity
pc:buyerProfile
rdf:Resource
0..1
nuts:NUTS
pc:tender
pc:Tender
- pc:publicationDate
0..*
pc:awardedTender
0..1
pc:onBehalfOf
0..*
0..1
pc:contractingAuthority
vcard:VCard
pceu:hasParentRegion
0..1
0..*
pc:item
0..*
pc:lot
0..*
0..*
pc:contact
0..*
s:Place
pc:location
0..1
pc:tenderOpeningLocation
0..*
gr:Offering
0..*
pc:item
0..*
pc:Contract
- dcterms:title
- dcterms:description
- pc:documentsRequestDeadline xsd:dateTime
- pc:tenderDeadline xsd:dateTime
- pc:tenderOpeningDateTime xsd:dateTime
- pc:duration xsd:duration
- pc:tenderMaintenanceDuration xsd:duration
- pc:startDate xsd:date
- pc:estimatedEndDate xsd:date
- pc:actualEndDate xsd:date
- pc:numberOfTenders xsd:nonNegativeInteger
0..1
<<owl:Restriction>>
skos:Concept
skos:inScheme
kinds:
pc:kind
0..1
<<owl:Restriction>>
skos:Concept
skos:inScheme
proctypes:
pc:procedureType
pc:offeredPrice
0..1
0..1
pc:documentationPrice
pc:estimatedPrice
gr:PriceSpecification
0..1
0..1
0..1
pc:agreedPrice
pc:actualPrice
0..*
0..1
pc:mainObject
pc:additionalObject
0..*
skos:Concept
0..1
0..*
pc:weightedCriterion
payment:Payment
0..1
pc:subsidy
0..1
pc:specification
0..1
pc:awardCriteriaCombination
0..1
foaf:Document
pc:AwardCriteriaCombination
pc:agreement
0..*
pc:attachment
0..1
pc:frameworkAgreement
0..*
pc:FrameworkAgreement
0..*
pc:awardCriterion
pc:CriterionWeighting
- pc:criterionWeight pcdt:percentage
pc:contract
pc:TendersOpening
pc:expectedNumberOfOperators
0..1
gr:QuantitativeValueInteger

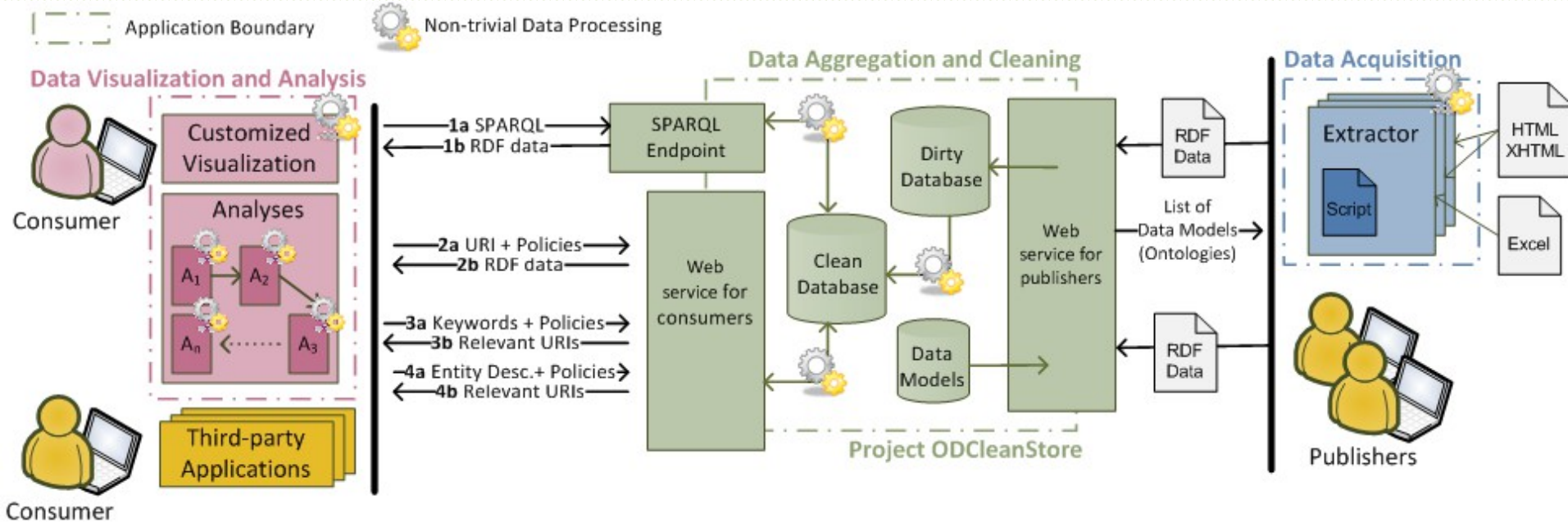
Data + Ontology



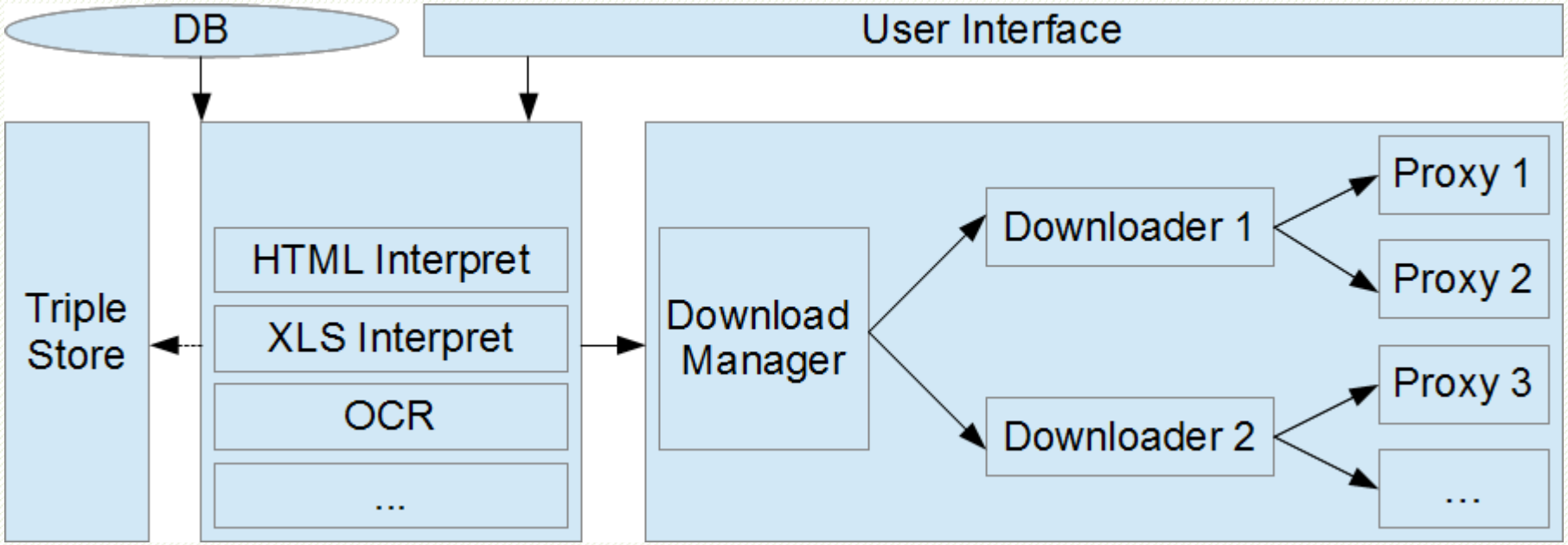
Ontology Connection

- Classes and properties are connected to extracted data
- GUI
  - Shows available properties and classes
  - Automatic download of used ontologies

Framework Architecture



Architecture



Download Manager

- POST, GET parameters
- COOKIE support
- Politeness policy
- Scalable
- Paralelism
- Proxy server support

Scraping Script

Script Creation

- Format specific user interface
  - Web Browser for HTML
  - MS Office for Excel
  - Ontology support
- XML editor
  - XML Schema

Scraping Script

- template
  - Basic building block
  - One page definition
- call-template
  - Call other defined template
- onto-element
  - Ontology connection
  - Subject or object

Scraping Script

- value-of
  - Data extraction
  - Static values
  - Selector values
    - css selectors
  - Regular expressions
- param
  - Script parameters

Script Interpretation

- Paralelism
- Statistics
  - Finished pages
  - Time
  - Ontology properties coverage
- Connection to triple store

Scraping Script

- Functions
  - conc(string, string)
  - convertSpacesToHtmlEntity(string)
  - removeSpaces(string)
  - convertToDate(string)
    - Recognize common date formats and return xsd:Date
  - convertToFloat(string)
  - generateUUID()
  - getCurrentUrl()
  - …

CSS Selectors

- JSoup [5]
- Tagname, classname, id
- Attributes
  - [attr], [attr~=regExp]
- Indexing
  - :lt(3), :eq(2), :gt(1)
- Text search
  - :contains(text), :matches(regExp)

Example

- European social fund in the czech republic
  - http://www.esfcr.cz/modules/procurements/?&data[is_running]=1&data[find]=less&lang=1

Scraping Script – Example 1

```
<?xml version="1.0" encoding="UTF-8"?>
<scr:script xmlns:scr="http://sourceforge.net/projects/strigil" id="KEG"
prefix="dc: http://purl.org/dc/elements/1.1/ pc:
http://purl.org/procurement/public-contracts# xsd:
http://www.w3.org/2001/XMLSchema#" version="0.1b" type="HTML">
    <scr:meta status="draft" domain="www.esfcr.cz" author="Administrátor"
date="2013-03-28" />
    <scr:call-template name="SeznamZakazek" type="http/GET">
        <scr:value-of text="http://www.esfcr.cz/modules/procurements/?
&amp;data[is_running]=1&amp;data[find]=less&amp;lang=1" />
    </scr:call-template>
    <scr:template name="SeznamZakazek" mime="text/html">
        <scr:samplePage url="http://www.esfcr.cz/modules/procurements/?
&amp;data[is_running]=1&amp;data[find]=less&amp;lang=1" />
        <scr:call-template name="DetailZakazky" type="http/GET">
            <scr:value-of select="body#esf div#wrapper div#all.cs div#main
div.innerwrap div.middle div#content.column-in div#main-content-center
div.blockContent div#procurements-box.procurements div.procurement-item div.item-
header h3 a @href" />
        </scr:call-template>
    </scr:template>
    ...
</scr:script>
```

Scraping Script – Example 2

```
…
<scr:template name="DetailZakazky" mime="text/html">
    <scr:onto-elem>
        <scr:onto-elem rel="http://www.w3.org/1999/02/22-rdf-syntax-ns#type"
about="http://purl.org/procurement/public-contracts#Contract" />
        <scr:value-of select="body#esf div#wrapper div#all.cs div#main
div.innerwrap div.middle div#content.column-in div#main-content-center
div.blockContent div#article h2" property="http://purl.org/dc/elements/1.1/title"
/>
        <scr:function name="convertDate"
property="http://purl.org/procurement/public-contracts#tenderDeadline">
            <scr:with-param>
                <scr:value-of select="body#esf div#wrapper div#all.cs
div#main div.innerwrap div.middle div#content.column-in div#main-content-center
div.blockContent div#article div.article-content div.pr-main-info
p:contains(Lhůta pro podávání nabídek)" />
            </scr:with-param>
        </scr:function>
    </scr:onto-elem>
</scr:template>
…
```

Sample Output

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:pc="http://purl.org/procurement/public-contracts#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
    <rdf:Description rdf:nodeID="A0">
        <dc:title>P&G – jazykové vzdělávání</dc:title>
        <pc:tenderDeadline rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
            2013-03-29
        </pc:tenderDeadline>
        <rdf:type rdf:resource="http://purl.org/procurement/public-contracts#Contract"/>
    </rdf:Description>
    <rdf:Description rdf:nodeID="A1">
        <dc:title>
            Poradenství a vzdělávání při zavádění moderních metod řízení pro město
Klimkovice
        </dc:title>
        <pc:tenderDeadline rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
            2013-04-04
        </pc:tenderDeadline>
        <rdf:type rdf:resource="http://purl.org/procurement/public-contracts#Contract"/>
    </rdf:Description>
…
```

XLS Selectors

- Worksheets
  - wks[name^="L"]
- Rows and columns
  - tr
  - td
- Headers
  - wks[name="L"] tc[hdr^="Person"]

Conclusion

Conclusion

- Data extraction
- Strigil
  - http://strigil.sourceforge.net/
  - Framework for script-based data extraction
  - Network scalability
  - Multiple format support
  - Ontology support
  - Scheduling
  - Modular design

Problems

- Adaptibility
  - Based on samplePage
- Script limitations
- Prototype

Future Work

- Script modification
- New formats
- New protocols
  - FTP, filesystem
  - Web services
- JavaScript support
- Cache

Thank You

References

- [1] https://scraperwiki.com/
- [2] http://www.visualwebripper.com
- [3] http://www.screen-scraper.com
- [4] Exploring the Web with OXPath
  - T. Furche, G. Gottlob, G. Grasso, C. Schallhart and A. Sellers. Proc. of LWDM. 2011.
- [5] http://jsoup.org/