Eksploracja sieci Internet z zastosowaniem analizy semantycznej

Autor inż. Grzegorz Polek
Promotor dr inż. Krzysztof Regulski
Recenzent dr inż. Andrzej Opaliński

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
AGH University of Science and Technology

14/07/2016

Agenda

- Linked Data
- Model danych i formaty RDF
- Ontologie
- Model systemu
- Wdrożenie systemu
- Eksploracja Sieci Web
- Analiza Semantyczna
- Konfiguracja głównego komponentu
- Wynik działania

Linked Data

Linked Data odnosi się do zbioru najlepszych praktyk dla publikowania i linkowania ustrukturowanych danych w sieci.

- Używaj URIs jako nazw dla zasobów;
- Uzywaj HTTP URIs tak, aby ludzie mogli przeglądać te zasoby;
- Gdy ktoś przegląda URI, przedstaw cenne informacje w ustandaryzowanych formatach (RDF, SPARQL).
- Zamieść odnośniki do innych zasobów, aby można było odkryć inne rzeczy.

Model danych RDF

Aby umożliwić wielu różnym aplikacjom możliwość przetwarzania zawartości stron internetowych, bardzo ważne jest wypracowanie ustandaryzowanych formatów danych. Publikując dane Linked Data w sieci, dane są reprezentowane przy pomocy **Resource Description Framework (RDF)**.

W standardzie RDF opis zasobu jest reprezentowany trójką wartości tzw. triple. Zasób składa się z podmiotu, predykatu i obiektu.

Aby opublikować dane w standardzie RDF, muszą one być zserializowane do odpowiedniego formatu RDF

Formaty RDF

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">
    <rdf:Description rdf:about="http://grzegorzpolek.com/ -
grzegorz-polek">
        <rdf:type
rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
        <foaf:name>Grzegorz Polek</foaf:name>
    </rdf:Description>
</rdf:RDF>
```

RDF/XML

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://grzegorzpolek.com/#grzegorz-polek>
    rdf:type foaf:Person ;
    foaf:name "Grzegorz Polek" .
```

Turtle

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
"http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:foaf="http://xmlns.com/foaf/0.1/">

    <head>
        <meta http-equiv="Content-Type"
content="application/xhtml+xml; charset=UTF-8" />
        <title>Grzegorz Polek</title>
    </head>

    <body>
        <div about="http://grzegorzpolek.com/#grzegorz-polek"
typeof="foaf:Person">
            <span property="foaf:name">Grzegorz Polek</span>
        </div>
    </body>
</html>
```

RDFa

```
<http://grzegorzpolek.com/#grzegorz-polek>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person> .
<http://grzegorzpolek.com/#grzegorz-polek>
<http://xmlns.com/foaf/0.1/name> "Grzegorz Polek" .
```

N-Triples

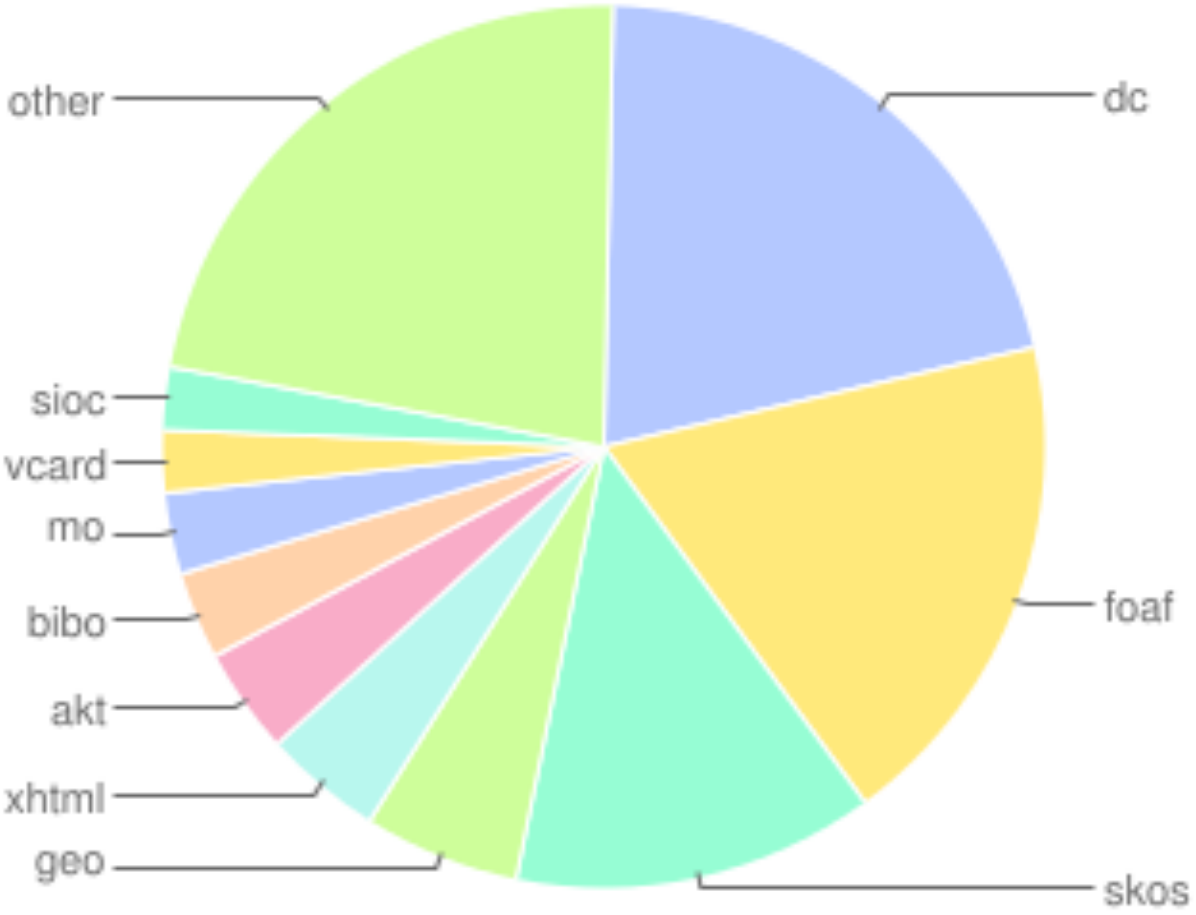
Ontologia

Ontologia to hierarchiczny system kategorii i relacji o uniwersalnym bądź ograniczonym zakresie. Na ontologię składa się zbiór pojęć i relacji między nimi.

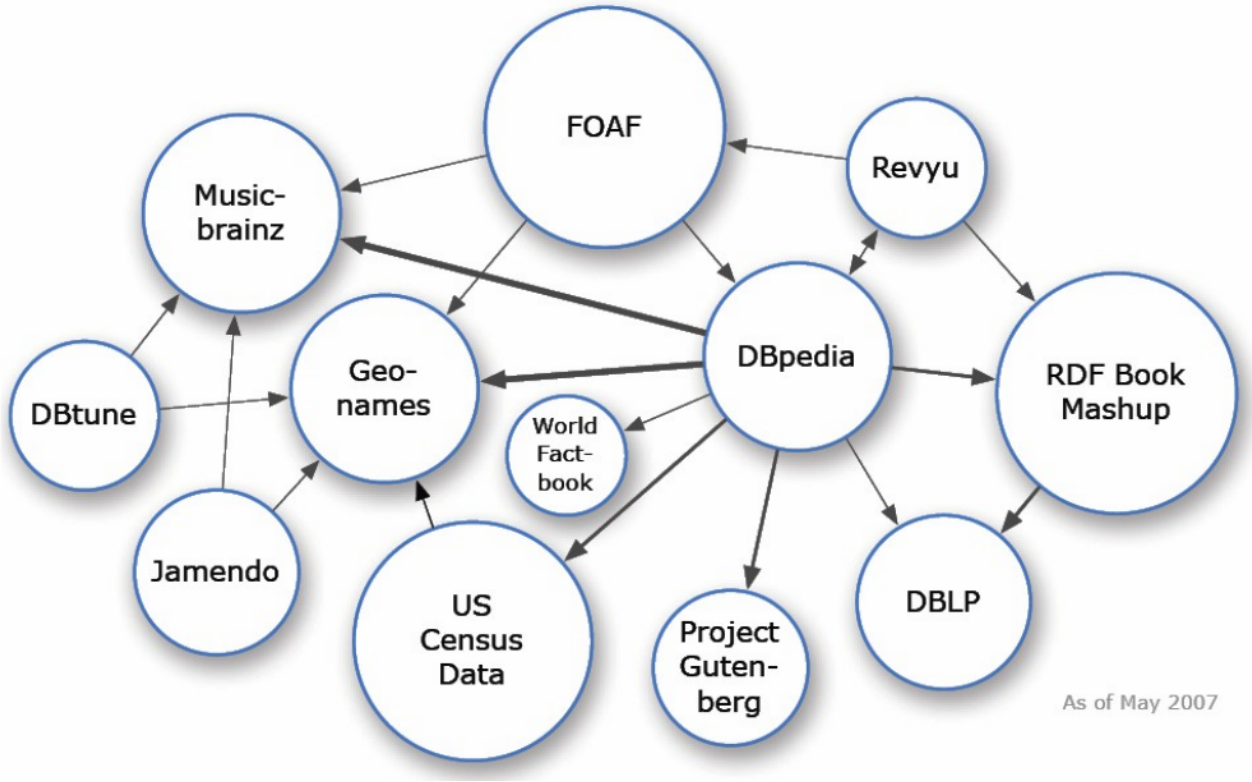
Ontologia różni się jednak od innych sposobów reprezentowania wiedzy tym, że nie tylko dostarcza schematu czy opisu danej dziedziny, ale za pomocą formalnych narzędzi logiki (aksjomatów, definicji, reguł) pozwala ściśle określać hierarchię jej elementów oraz kryteria ich klasyfikacji.

Do opisu ontologii wykorzystuje się języki RDF (ang. Resource Description Framework) lub bardziej rozbudowany OWL (Web Ontology Language), znajdujący zastosowanie w Sieci Web

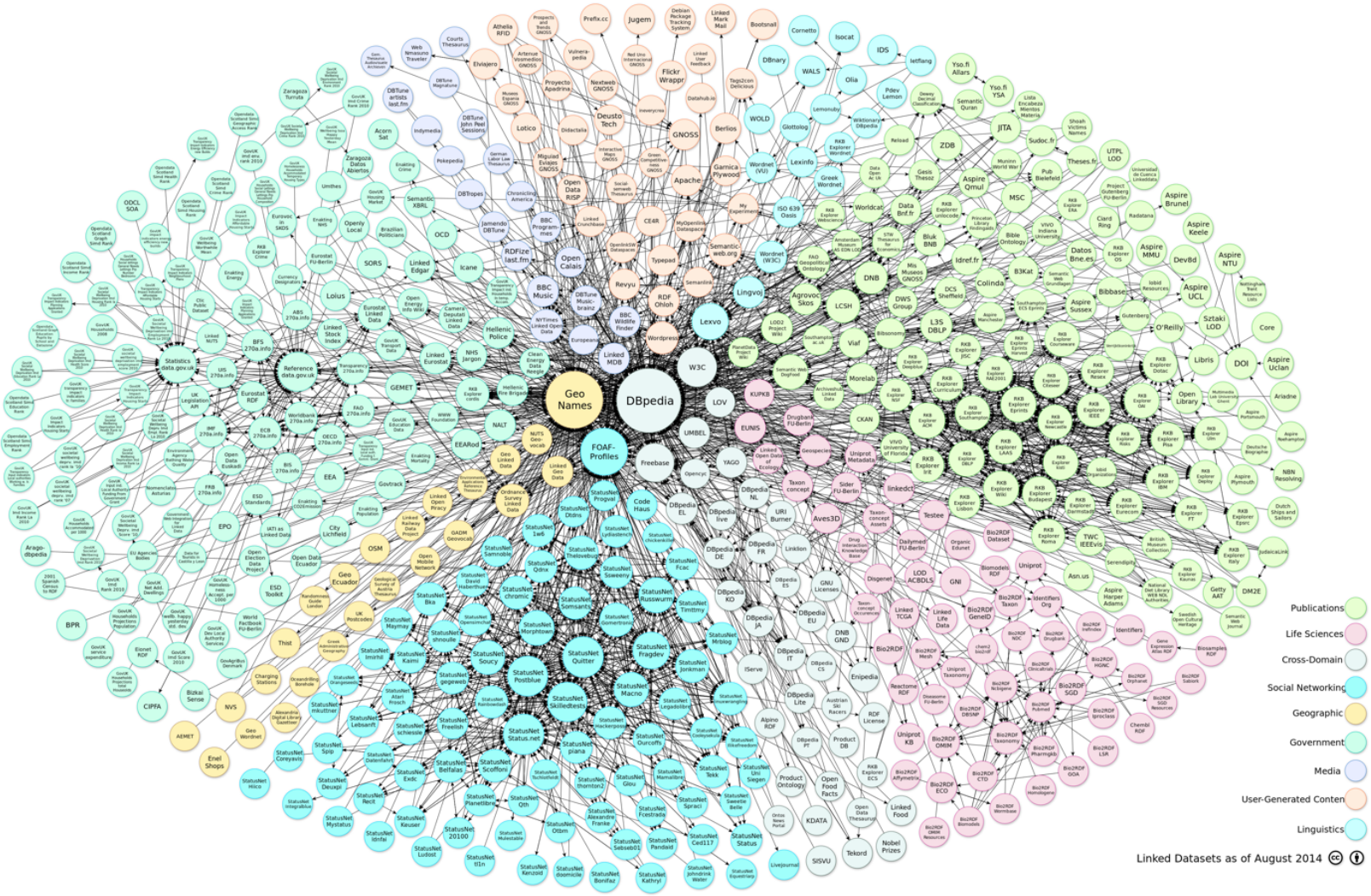
Ontologie



Zasób słowników projektu Linking Open Data

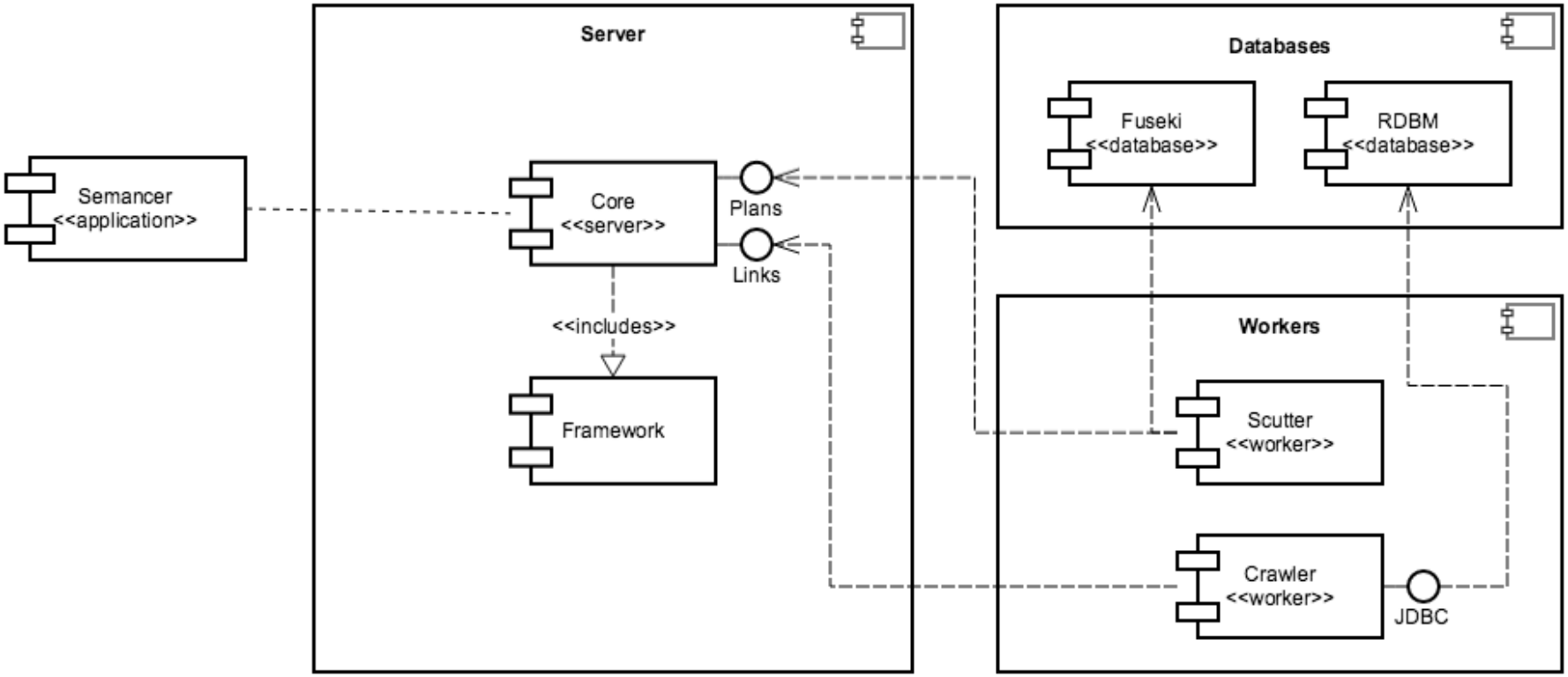


Stan Linked Data na Maj 2007

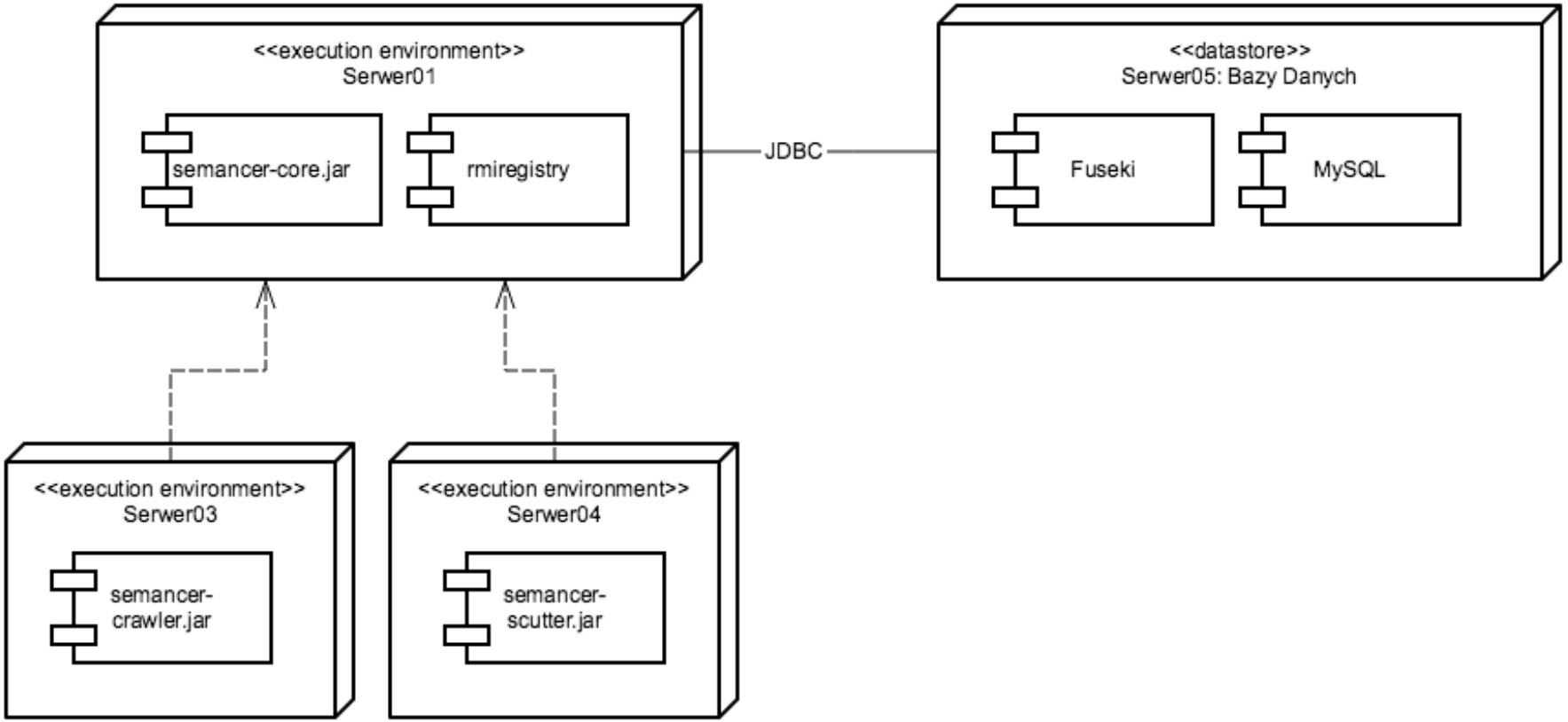


Stan Linked Data na Sierpień 2014

Model systemu



Wdrożenie systemu



Eksploracja Sieci Web

W trybie eksploracji realizowanym przez klienta **crawler**, przeszukujemy sieć w poszukiwaniu semantycznych danych czy to w postaci całych dokumentów RDF, czy stron internetowych zawierających dane w postaci np. RDFa. W tym trybie system nie jest świadomy danych, które eksploruje, a jedynie tworzy graf połączeń pomiędzy zasobami, co ułatwia późniejszą analizę zasobów.

Crawler podąża za wszystkimi wystąpieniami odniesień do innych zasobów czy to w postaci linków do innych stron internetowych, czy przy pomocy semantycznych odniesień rdf:seeAlso.

Analiza Semantyczna

W trybie analizy, który jest realizowany przez klienta **scutter**, system ma za zadanie przetworzyć dane znajdujące się w sieci do semantycznego formatu, np. trójek turtle, który może być w prosty sposób zaimportowany do przeznaczonej do tego celu bazy danych.

W tym trybie scutter operuje tylko i wyłącznie na dokumentach rdf, bądź dokumentach zawierających częściowe dane semantyczne i świadomie podąża za odniesieniami rdf:seeAlso.

Konfiguracja komponentu scutter

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:slug="http://purl.org/NET/schemas/slug/config/">

    <rdf:Description rdf:about="">
        <dc:title>Scutter Configuration File</dc:title>
        <dc:description></dc:description>
    </rdf:Description>

    <slug:Scutter rdf:about="default">
        <dc:description>A default Scutter
configuration</dc:description>

        <!-- configure global memory -->
        <slug:hasMemory rdf:resource="memory" />

        <slug:file>/tmp/output.rdf</slug:file>

        <!-- how many worker threads -->
        <slug:workers>4</slug:workers>

        <slug:userAgent></slug:userAgent>

        <!-- configures consumers for incoming data -->
        <slug:consumers>
            <rdf:Seq>
                <rdf:li rdf:resource="storer"/>
                <rdf:li rdf:resource="rdf-parser"/>
                <rdf:li rdf:resource="rdf-transformer"/>
                <rdf:li rdf:resource="rdf-follower"/>
            </rdf:Seq>
        </slug:consumers>

        <!-- configures filter pipeline for controller -->
        <slug:filters>
            <rdf:Seq>
                <rdf:li rdf:resource="depth-filter"/>
            </rdf:Seq>
        </slug:filters>

    </slug:Scutter>

</rdf:RDF>
```

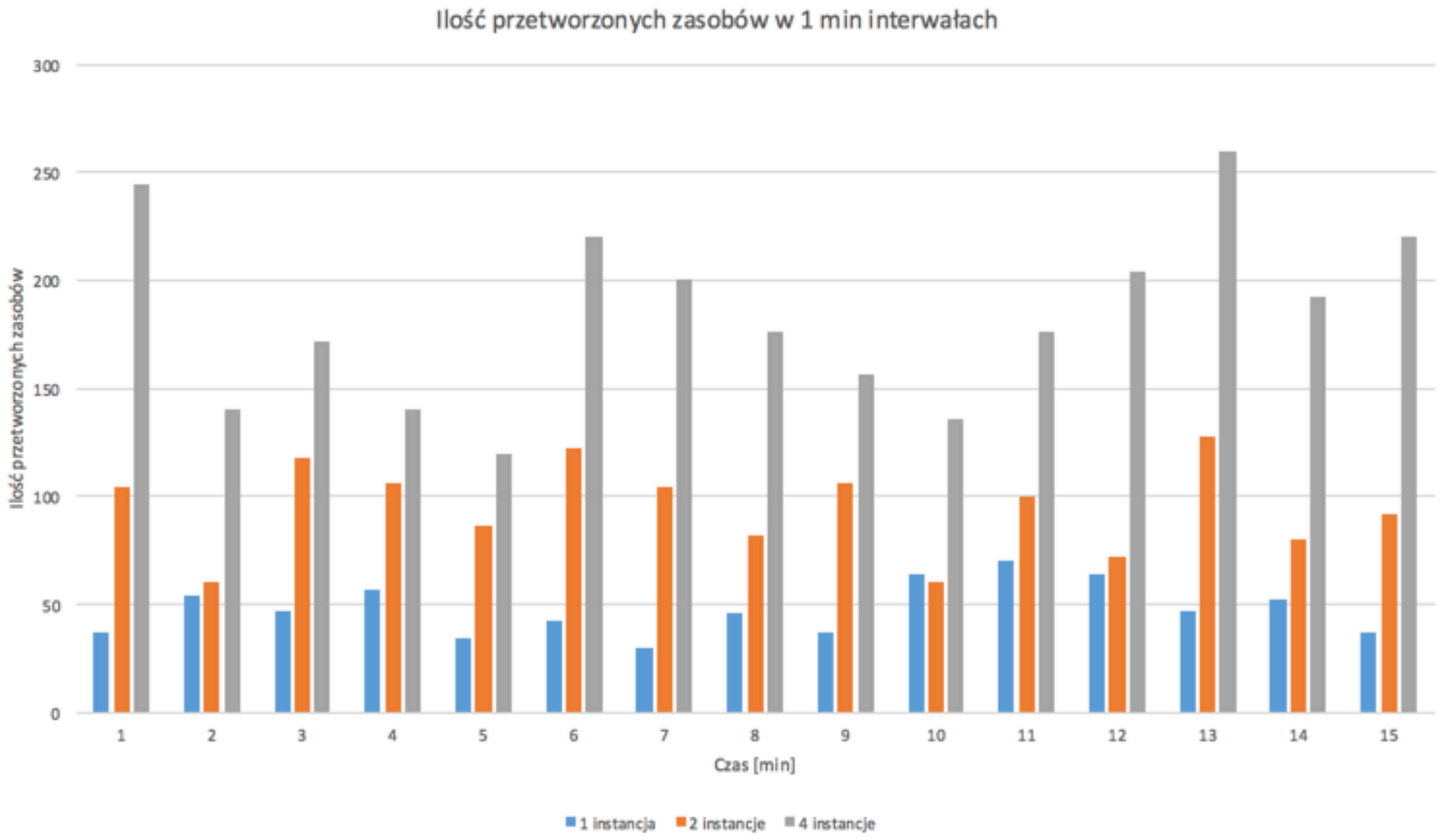
Plan dla komponentu Scutter

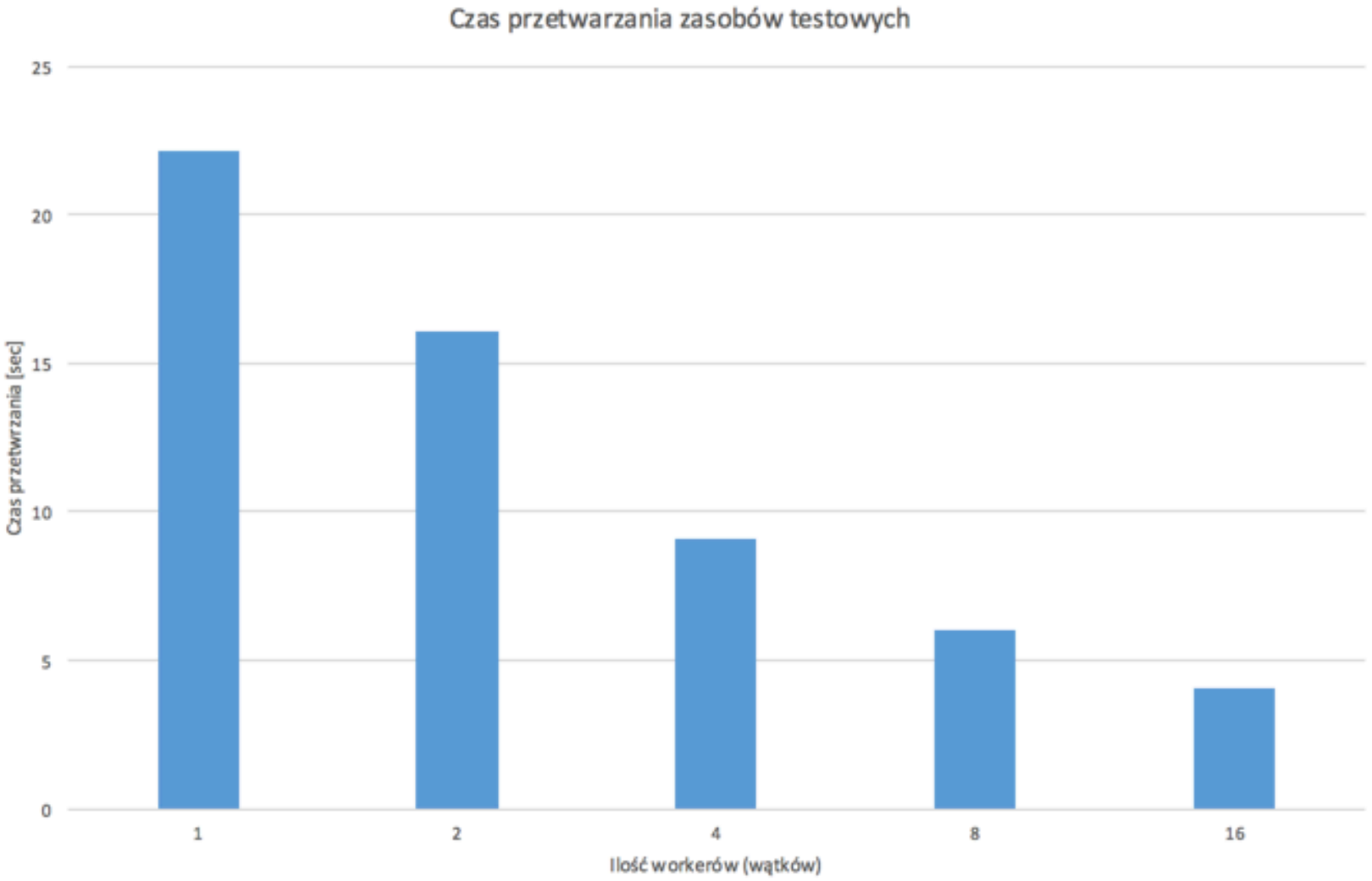
```
<?xml version="1.0" standalone="yes"?>
<rdf:RDF
    xmlns:img="http://jibbering.com/2002/3/svg/#"
    xmlns:wn="http://xmlns.com/wordnet/1.6/"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:an="http://rdf.desire.org/vocab/recommend.rdf#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:wot="http://xmlns.com/wot/0.1/"
    xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#"
    xmlns:air="http://www.megginson.com/exp/ns/airports#">

    <rdf:Description>
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/Cambridge.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/Iceland.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/George_W._Bush.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/Diana_Ross.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/Like_a_Virgin.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/Oliver_Stone.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/Pulp_Fiction.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/The_Lord_of_the_Rings.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/World_of_Warcraft.rdf" />
        <rdfs:seeAlso
rdf:resource="http://dbpedia.org/data/Resource_Description_Framework.r
df" />
    </rdf:Description>


</rdf:RDF>
```

Wynik działania komponentu Scutter

```
Jun 07, 2016 11:50:56 PM com.ldodds.slug.Scutter run
INFO: Crawler memory loaded
INFO: Plan received default /tmp/scutterplan60e8c8af-8cf8-4e8d-a474-
1f6eca7dff74.rdf
INFO: Resources found in plan
[[http://dbpedia.org/data/The_Beatles.rdf]http://dbpedia.org/data/Th
e_Beatles.rdf, depth=0,
[http://dbpedia.org/data/Berlin.rdf]http://dbpedia.org/data/Berlin.r
df, depth=0,
[http://dbpedia.org/data/Category:Cities_in_England.rdf]http://dbped
ia.org/data/Category:Cities_in_England.rdf, depth=0,
[http://dbpedia.org/data/Category:English_musicians.rdf]http://dbped
ia.org/data/Category:English_musicians.rdf, depth=0,
[http://dbpedia.org/data/SPARQL.rdf]http://dbpedia.org/data/SPARQL.r
df, depth=0,
[http://dbpedia.org/data/Paul_McCartney.rdf]http://dbpedia.org/data/
Paul_McCartney.rdf, depth=0,
[http://dbpedia.org/data/Semantic_Web.rdf]http://dbpedia.org/data/Se
mantic_Web.rdf, depth=0,
[http://dbpedia.org/date/Tetris.rdf]http://dbpedia.org/date/Tetris.r
df, depth=0,
[http://dbpedia.org/data/The_Lord_of_the_Rings.rdf]http://dbpedia.or
g/data/The_Lord_of_the_Rings.rdf, depth=0]
Jun 07, 2016 11:50:56 PM com.ldodds.slug.framework.Controller run
INFO: Controller thread starting
Jun 07, 2016 11:51:26 PM com.ldodds.slug.framework.Controller run
INFO: Controller thread stopping. Task list completed
Jun 07, 2016 11:51:26 PM com.ldodds.slug.framework.Controller stop
INFO: Controller thread halted
Jun 07, 2016 11:51:26 PM com.ldodds.slug.Scutter stop
INFO: Stopping scutter
Jun 07, 2016 11:51:26 PM com.ldodds.slug.framework.Controller stop
INFO: Controller thread halted
Jun 07, 2016 11:51:26 PM com.ldodds.slug.Scutter save
INFO: Saving memory
Jun 07, 2016 11:51:26 PM com.ldodds.slug.framework.config.FileMemory
save
INFO: Memory Saved
```

```
data❯ ls -la
total 3920
drwxr-xr-x  10 gpolek  staff   340B Jun  8 00:01 .
drwxr-xr-x   3 gpolek  staff   102B May 28 23:55 ..
-rw-r--r--   1 gpolek  staff   1.3M Jun  8 01:20 Berlin.rdf
-rw-r--r--   1 gpolek  staff   4.6K Jun  8 01:20
Category:Cities_in_England.rdf
-rw-r--r--   1 gpolek  staff    32K Jun  8 01:20
Category:English_musicians.rdf
-rw-r--r--   1 gpolek  staff   236K Jun  8 01:20
Paul_McCartney.rdf
-rw-r--r--   1 gpolek  staff    34K Jun  8 01:20 SPARQL.rdf
-rw-r--r--   1 gpolek  staff    59K Jun  8 01:20 Semantic_Web.rdf
-rw-r--r--   1 gpolek  staff   222K Jun  8 01:20 The_Beatles.rdf
-rw-r--r--   1 gpolek  staff    66K Jun  8 01:20
The_Lord_of_the_Rings.rdf
```


Ilość przetworzonych zasobów w 1 min interwałach
300
250
200
150
100
50
0
Ilość przetworzonych zasobów
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Czas [min]
1 instancja
2 instancje
4 instancje


Czas przetwarzania zasobów testowych
25
20
15
10
5
0
Czas przetwrzania [sec]
1
2
4
8
16
Ilość workerów (wątków)

Eksploracja sieci Internet z zastosowaniem analizy semantycznej

Autor inż. Grzegorz Polek
Promotor dr inż. Krzysztof Regulski
Recenzent dr inż. Andrzej Opaliński

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
AGH University of Science and Technology

14/07/2016