Linking the Deep Web to the Linked Data Web

Rahul Parundekar, Craig A. Knoblock and José Luis Ambite
{parundeck, knoblock, ambite}@isi.edu
University of Southern California/Information Sciences Institute
Motivation

• Large amount of data is present on the traditional Web in the form of *Deep Web* and the *Surface Web* data sources

• Automatically generate Semantic Web Services from these traditional Web sources

• Huge potential for structured knowledge can be realized from linking this RDF data to the Linked Data Cloud

• Contribution: Information integration between the LDW and the *Deep Web*
Sources on the Web

- Have well-defined inputs and outputs or produce a result page on accepting specific input

- HTML Forms
Sources on the Web

- Structured data needs to be extracted from HTML result pages
Automatically Constructing Semantic Web Services from Online Sources

[Ambite et al. ISWC’09]

Modeling the Newly Discovered Source for the Input “RBCGX”

Yahoo Finance result

Reynolds Blue Chip Growth (RBCGX)

REYNOLDS BLUE CHIP GROWTH FUND,

- Net Asset Value: 46.60
- Trade Time: Mar 15
- Change: ↓0.17 (0.36%)
- Prev Close: 46.60
- YTD Return*: -5.53%
- Net Assets*: 62.81M
- Yield*: N/A

* As of 31 Jan 10

Google Finance result

Reynolds Blue Chip Growth (MUTF:RBCGX) Watch this

46.60  -0.17 (-0.36%)

Mar 15, 8:00PM EDT  Overall Morningstar Rating™ ★★★★★

Key statistics

- Total assets 67.79M
- Front load -
- Deferred load -
- Expense ratio 2.00%
- Management fee -
- Fund family Reynolds

Funds category statistics on Morningstar »
Modeling the Newly Discovered Source for the Input “RBCGX”

**Yahoo Finance result**

**FundName**

Reynolds Blue Chip Growth (RBCGX)

**CurrentValue**

46.60

**ChangeValue**

-0.17 (0.36%)

**ChangePercentage**

5.53%

**Google Finance result**

Reynolds Blue Chip Growth (MUTF:RBCGX) [Watch this]

46.60 -0.17 (-0.36%)

Mar 15, 8:00PM EDT Overall Morningstar Rating™ ★★☆☆☆

**Key statistics**

- Total assets: 67.79M
- Expense ratio: 2.00%
- Management fee: -
- Fund family: Reynolds

[Funds category statistics on Morningstar »](#)
Modeling the Newly Discovered Source for the Input "RBCGX"

Yahoo Finance result

Reynolds Blue Chip Growth (RBCGX)

REYNOLDS BLUE CHIP GROWTH FUND,

Net Asset Value: 46.60
Trade Time: Mar 15
Change: -0.17 (0.36%)
Prev Close: 46.60
YTD Return*: -5.53%
Net Assets*: 62.81M
Yield*: N/A

Google Finance result

Reynolds Blue Chip Growth (MUTF:RBCGX)

46.60 -0.17 (-0.36%)
Mar 15, 8:00PM EDT Overall Morningstar Rating™ ★★★★★

Key statistics
- Total assets: 67.79M
- Front load: -
- Deferred load: -
- Expense ratio: 2.00%
- Management fee: -
- Fund family: Reynolds

Funds category statistics on Morningstar »
Modeling the Newly Discovered Source for the Input “RBCGX”

Yahoo Finance result

Reynolds Blue Chip Growth (RBCGX)

REYNOLDS BLUE CHIP GROWTH FUND,

Net Asset Value: 46.60
Trade Time: Mar 15
Change: ↓ 0.17 (0.36%)
Prev Close: 46.60
YTD Return*: -5.53%
Net Assets*: 62.81M
Yield*: N/A

* As of 31 Jan 10

Google Finance result

Reynolds Blue Chip Growth (MUTF:RBCGX) Watch this m

46.60 -0.17 (-0.36%)
Mar 15, 8:00PM EDT Overall Morningstar Rating™ ★★★☆☆

Key statistics

- Total assets: 67.79M
- Front load: -
- Deferred load: -
- Expense ratio: 2.00%
- Management fee: -
- Fund family: Reynolds

Funds category statistics on Morningstar »


googlefinance(FundSymbol, FundName, ...)
:-yahoo finance (FundSymbol, ..., FundName)
Generating Triples in the Semantic Web Service

Seed source definition

```
yahooofinance($FundSymbol, NetValue, ChangeDirection, 
  ChangeAmount, ChangePercent, PreviousClose, 
  YTDReturn, NetAssets, Yield, FundName) :- 

Contract (@Ct), 
hasSymbol (@Ct, @Sy), Symbol (@Sy), 
hasValue (@Sy, FundSymbol), 
hasName (@Ct, @N), Name (@N), hasValue (@N, FundName), 
hasNetValue (@Ct, @Net), NetValue (@Net), 
hasValue (@Net, NetValue), 
hasChangeAmount (@Ct, @ChA), ChangeAmount (@ChA), 
hasValue (@ChA, ChangeAmount), 
hasChangePercent (@Ct, @ChP), ChangePercent (@ChP), 
hasValue (@ChP, ChangePercent),
hasChangeDirection (@Ct, @ChD), ChangeDirection (@ChD), 
hasValue (@ChD, ChangeDirection), 
hasPreviousClose (@Ct, @Pre), PreviousClose (@Pre), 
hasValue (@Pre, PreviousClose).
```

Ontology in terms of unary and binary predicates in a LAV rule to perform lifting and format the results at run time into triples for output

Definition of the discovered Source

```
googleofinance (FundSymbol, FundName, ...) :- yahooofinance (FundSymbol, ..., FundName)
```
Linking the *Deep Web* Sources into LDW

- Instances generated by the Semantic Web Service need to be linked to existing Individuals in the LDW
Linking the Deep Web Sources into LDW

- Instances generated by the Semantic Web Service need to be linked to existing Individuals in the LDW

```
newSource
  googlefinance($FundSymbol,FundName,...)
  :- yahooFinance($FundSymbol,...,FundName)
```

```
seedSource
  define with the same Ontology
```

```
linkedDataSource
```

```
newSource - seedSource - linkedDataSource
```
Linking the *Deep Web* Sources into LDW

- Instances generated by the Semantic Web Service need to be linked to existing Individuals in the LDW

```
Link instances at run-time
```

```
New Source
```

```
Seed Source
```

```
Linked Data Source
```

```
googelfinance($FundSymbol,FundName,...)
:- yahoofinance($FundSymbol,...,FundName)
```
Linking the Seed Source to the LDW

Common Ontology

Contract
  - hasFundName
  - hasFundSymbol

FundName
  - hasValue

FundSymbol
  - hasValue

contract1
  - hasFundName
  - hasFundSymbol

fundname1
  - hasValue
  "Reynolds Blue Chip Growth"

fundsymbol1
  - hasValue
  "RBCGX"

C000002481
  - :fn
  - :fs
  - hasFundName
  - hasFundSymbol

  - hasValue
  "Reynolds Blue Chip Growth"

  - hasValue
  "RBCGX"

SWS Instances

LDS Instances
Linking the Seed Source to the LDW

**Common Ontology**

- **Contract**
  - **FundName**
  - **FundSymbol**

Record Linkage:
"Find an instance in the LDS with Name like 
<FundName> or Symbol like 
<FundSymbol>"

**SWS Instances**

- **contract1**
  - **fundname1**
    - hasFundName: "Reynolds Blue Chip Growth"
    - hasFundSymbol: "RBCGX"

- **fundsymboll1**
    - hasFundName: "Reynolds Blue Chip Growth"
    - hasFundSymbol: "RBCGX"

**LDS Instances**

- **C000002481**
  - _:fn
  - _:fs
  - hasFundName: "Reynolds Blue Chip Growth"
  - hasFundSymbol: "RBCGX"
“Find an instance in the LDS with Name matches ‘REYNOLDS BLUE CHIP GROWTH’ or Symbol matches ‘RBCGX’”

Newly discovered source (googlefinance)

googlefinance SWS instances generated at run-time
Implementation

**Linked Data Source**
- http://www.rdfabout.com/demo/sec/
- Corporate ownership data published as Linked Data.
- We extrapolate the Ontology used to match the structure of the EDGAR database & generate appropriate URIs

<table>
<thead>
<tr>
<th>CIK</th>
<th>Series</th>
<th>Class/Contract</th>
<th>Name</th>
<th>Ticker Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000832574</td>
<td></td>
<td></td>
<td>REYNOLDS FUNDS INC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>REYNOLDS BLUE CHIP GROWTH FUND</td>
<td>RBCGX</td>
</tr>
<tr>
<td></td>
<td>S000000865</td>
<td></td>
<td>REYNOLDS BLUE CHIP GROWTH FUND</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C000002481</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- As the database was not downloadable, we realized the Linking Query as a Wrapper that returns the URI of the Company/Series/Contract instance that we want the instance generated by the Semantic Web Service to be linked to
Preliminary Results

- **Sources discovered by the previous work**
  - http://www.google.com/finance
  - http://www.streetinsider.com/
  - http://money.cnn.com/

- **Instances in the result of the SWS were linked to the LDW**

- **Limitation of the simple Record Linkage: String Equality imposes strong restriction**
  - E.g. streetinsider does not return FundName. Has prefix of ‘MF:’ to the fund code in the result
  - Relies on input value of FundSymbol for linking
Conclusion & Future Work

• We are able publish the extracted data from known as well as unknown sources as structured linked data.

• A potentially large amount of Data can be now be accessible as Linked Data.

• Substantial step in automatically integrating *Deep Web* sources to the Linked Data Web.

• Future Work:
  • Automatically linking Concepts of sources in the LDW.
  • Aligning ontologies present in the LDW using the instance level ‘owl:sameAs’ links.