USC Viterbi
School of Engineering

Linking the Deep Web to the Linked Data Web

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

USC

USC Viterbi School of Engineering

Motivation

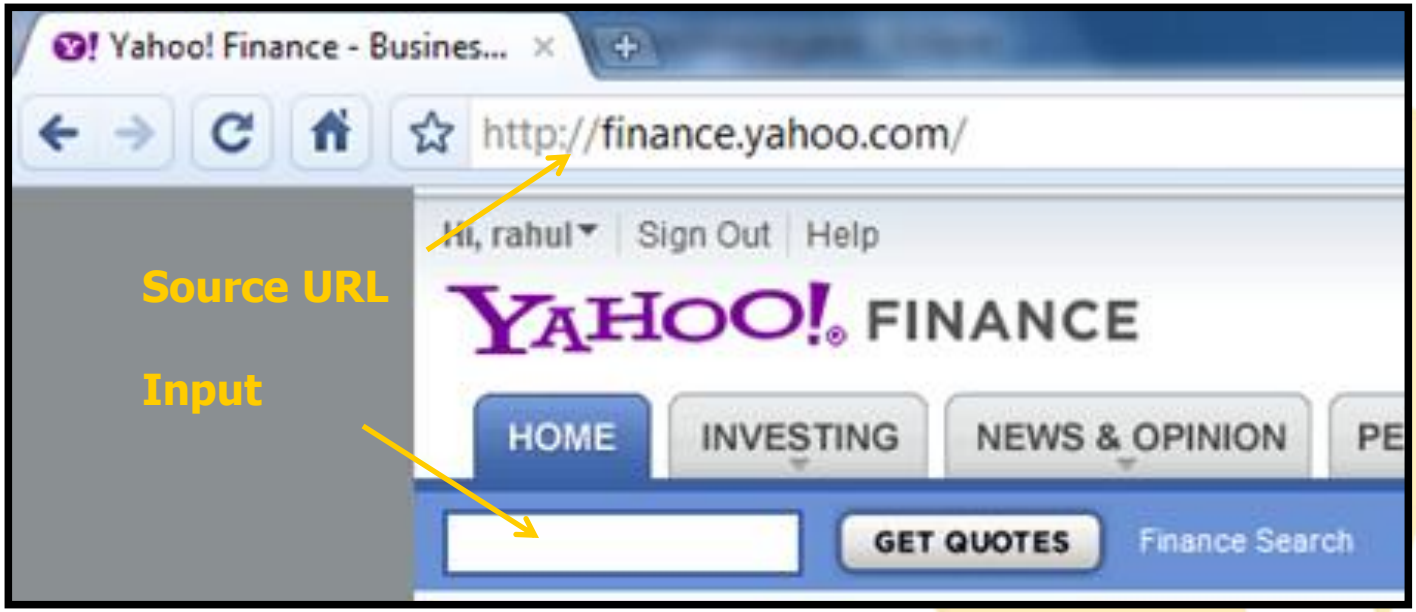
- Large amount of data is present on the traditional Web in the form of *Deep Web* and the *Surface Web* data sources
- Automatically generate Semantic Web Services from these traditional Web sources
- Huge potential for structured knowledge can be realized from linking this RDF data to the Linked Data Cloud
- Contribution: Information integration between the LDW and the *Deep Web*

USC

USC Viterbi School of Engineering

Sources on the Web

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

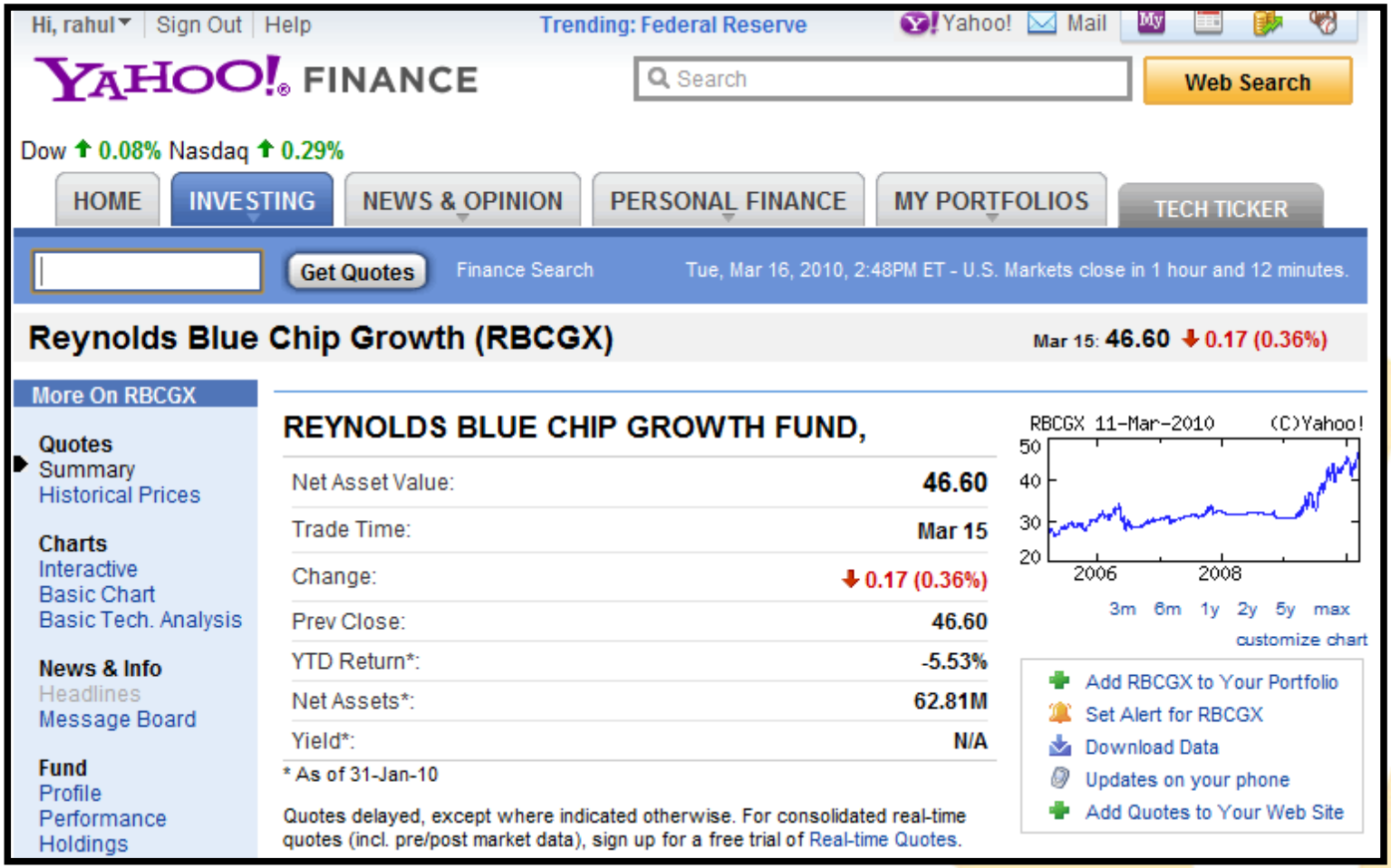


USC

USC Viterbi School of Engineering

Sources on the Web

- **Structured data needs to be extracted from HTML result pages**

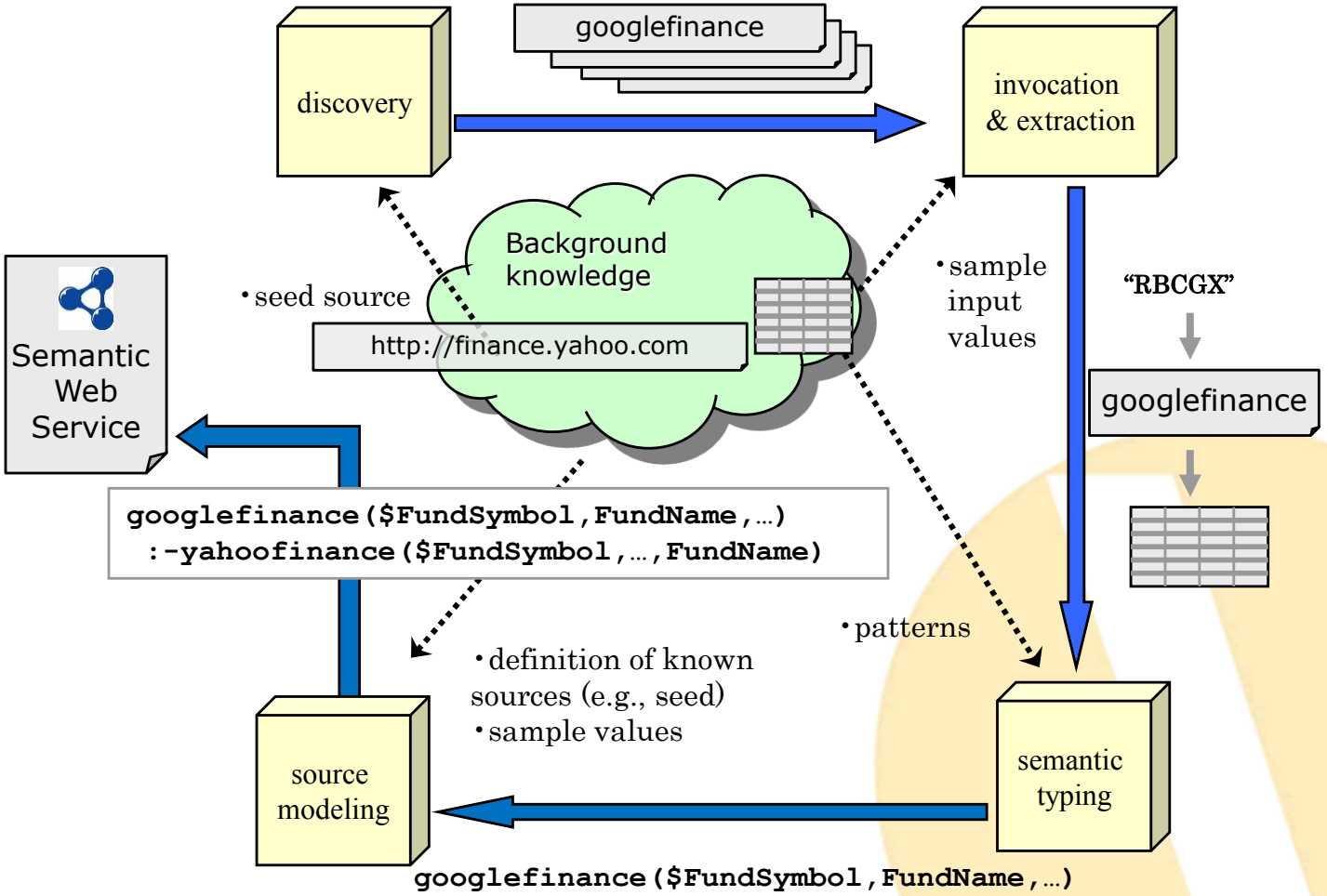


USC

USC Viterbi School of Engineering

Automatically Constructing Semantic Web Services from Online Sources

[Ambite et al. ISWC'09]

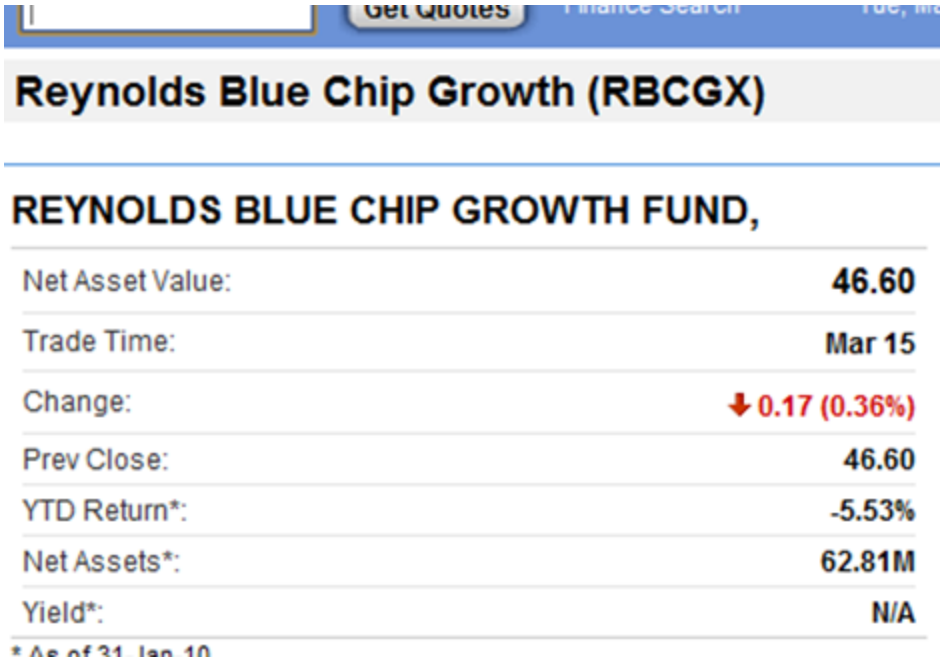


Ambite, J.L. and Darbha, S. and Goel, A. and Knoblock, C.A. and Lerman, K. and Parundekar, R. and Russ, T. - Automatically Constructing Semantic Web Services from Online Sources – *Presented at the International Semantic Web Conference 2009*

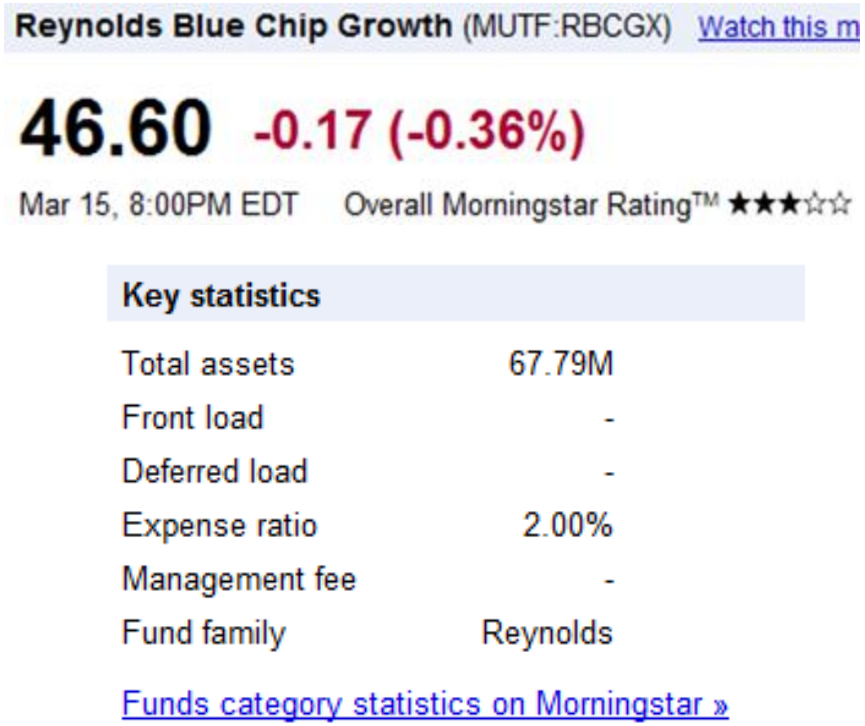
USC Viterbi School of Engineering

Modeling the Newly Discovered Source for the Input "RBCGX"

Yahoo Finance result



Google Finance result



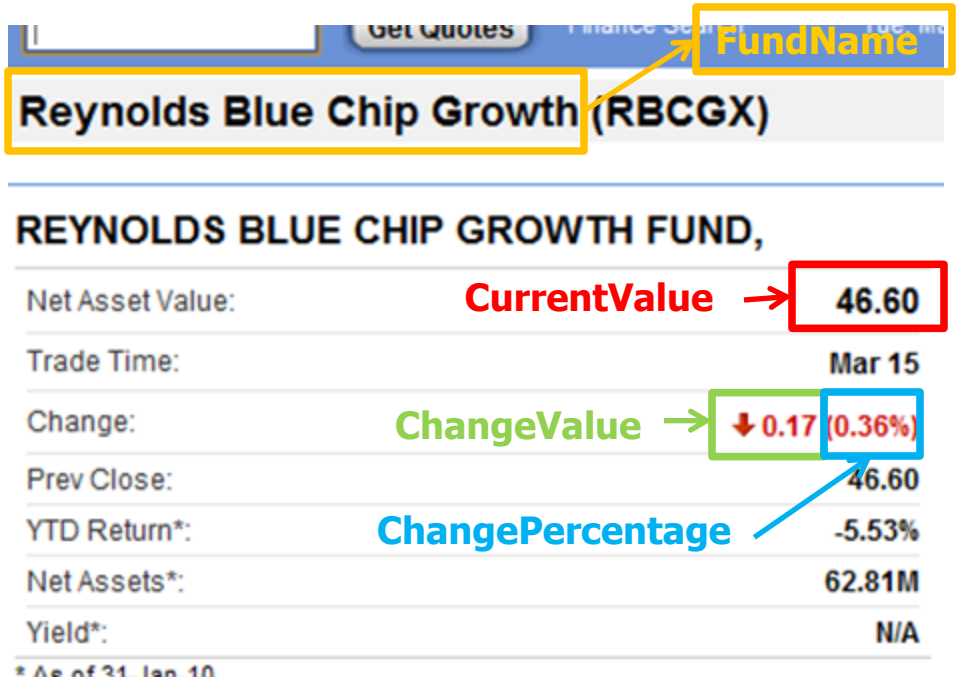
USC

USC Viterbi School of Engineering

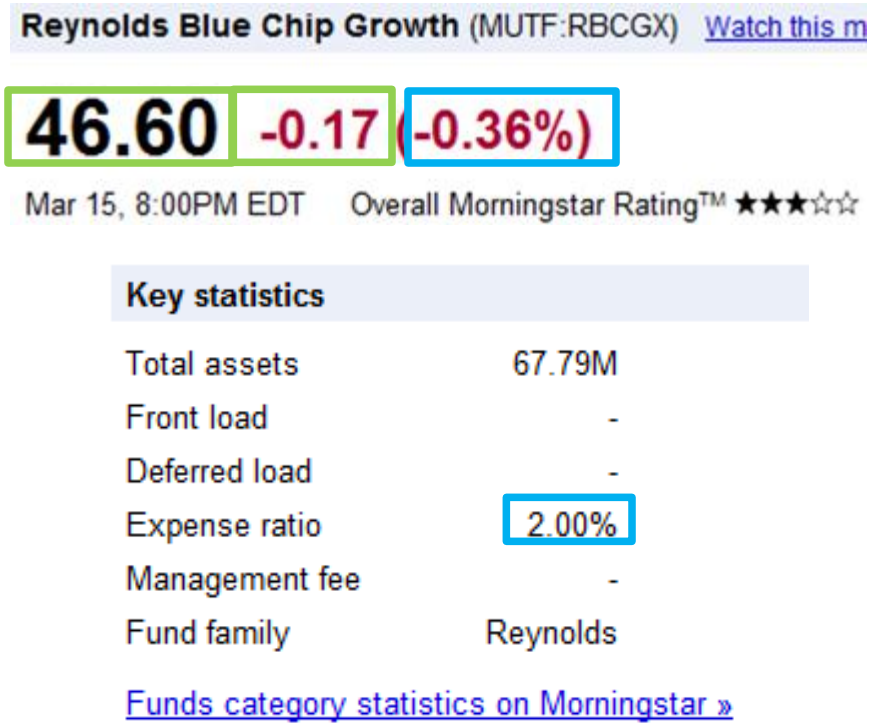
Modeling the Newly Discovered Source for the Input "RBCGX"

Semantic Typing

Yahoo Finance result



Google Finance result



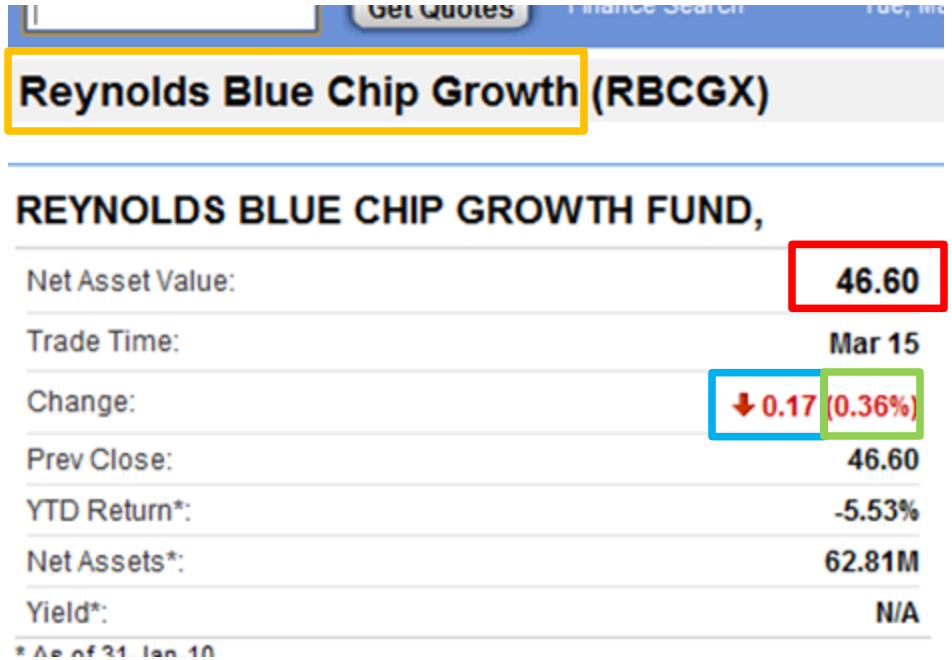
USC

USC Viterbi School of Engineering

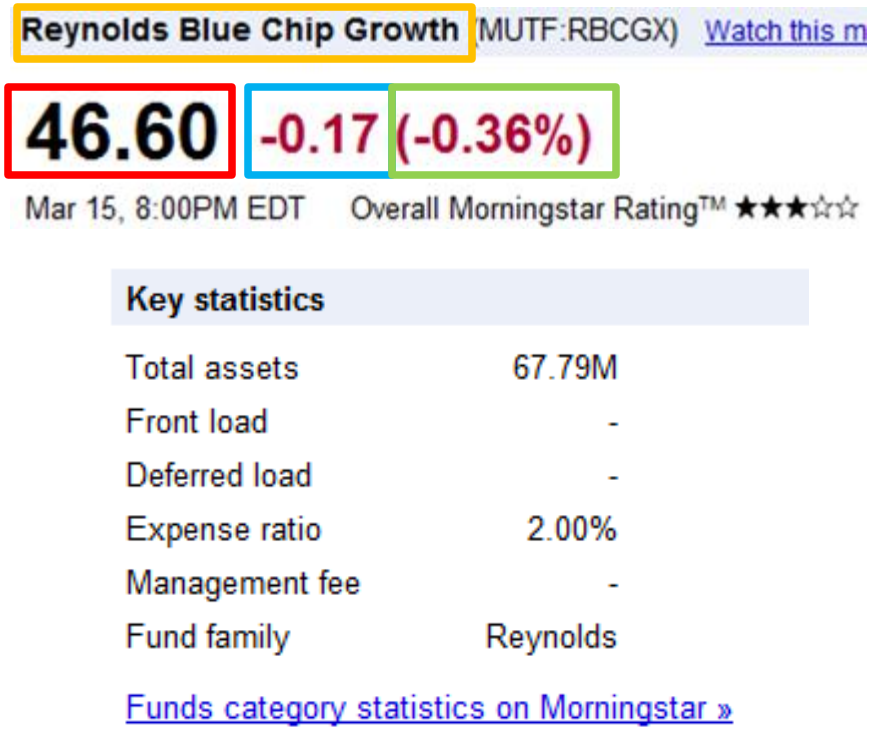
Modeling the Newly Discovered Source for the Input "RBCGX"

Source Modeling

Yahoo Finance result



Google Finance result

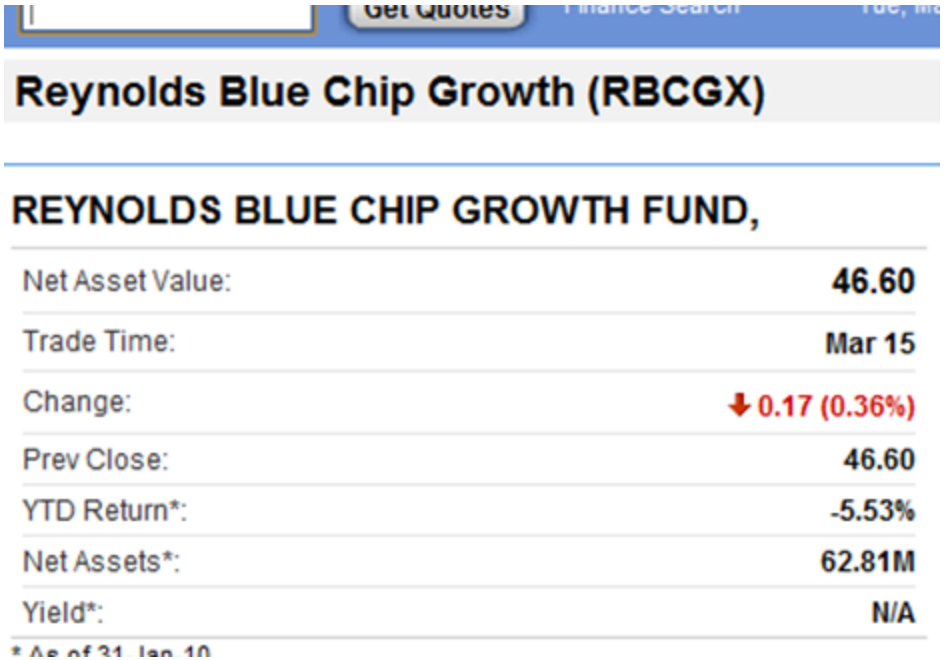


USC

USC Viterbi School of Engineering

Modeling the Newly Discovered Source for the Input "RBCGX"

Yahoo Finance result



Get Quotes

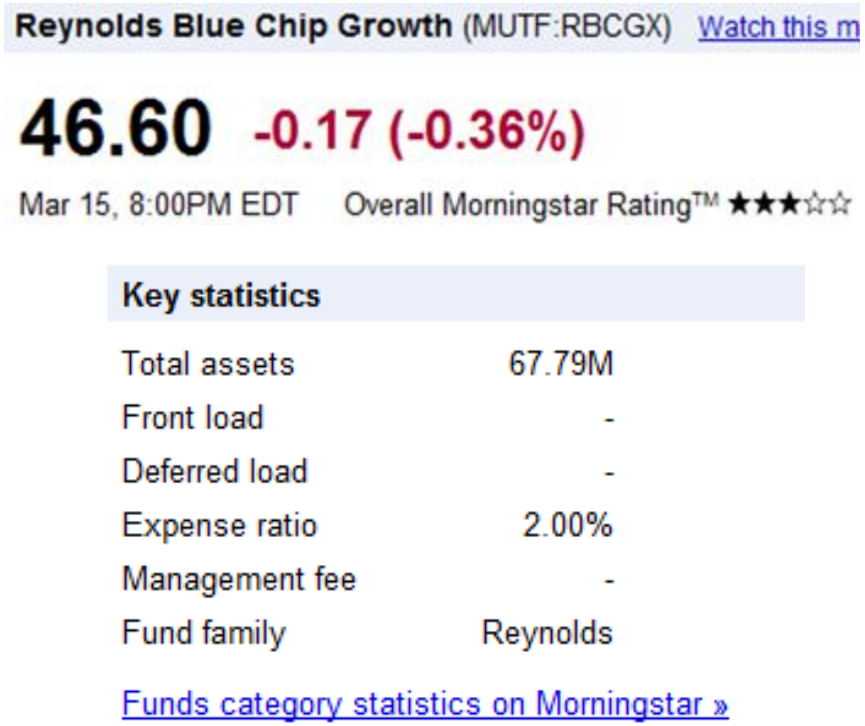
Reynolds Blue Chip Growth (RBCGX)

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 »

USC

```
googlefinance(FundSymbol,FundName,…)
 :-yahoofinance(FundSymbol,…,FundName)
```

USC Viterbi School of Engineering

Generating Triples in the Semantic Web Service

Seed source definition

```
yahoofinance($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

USC

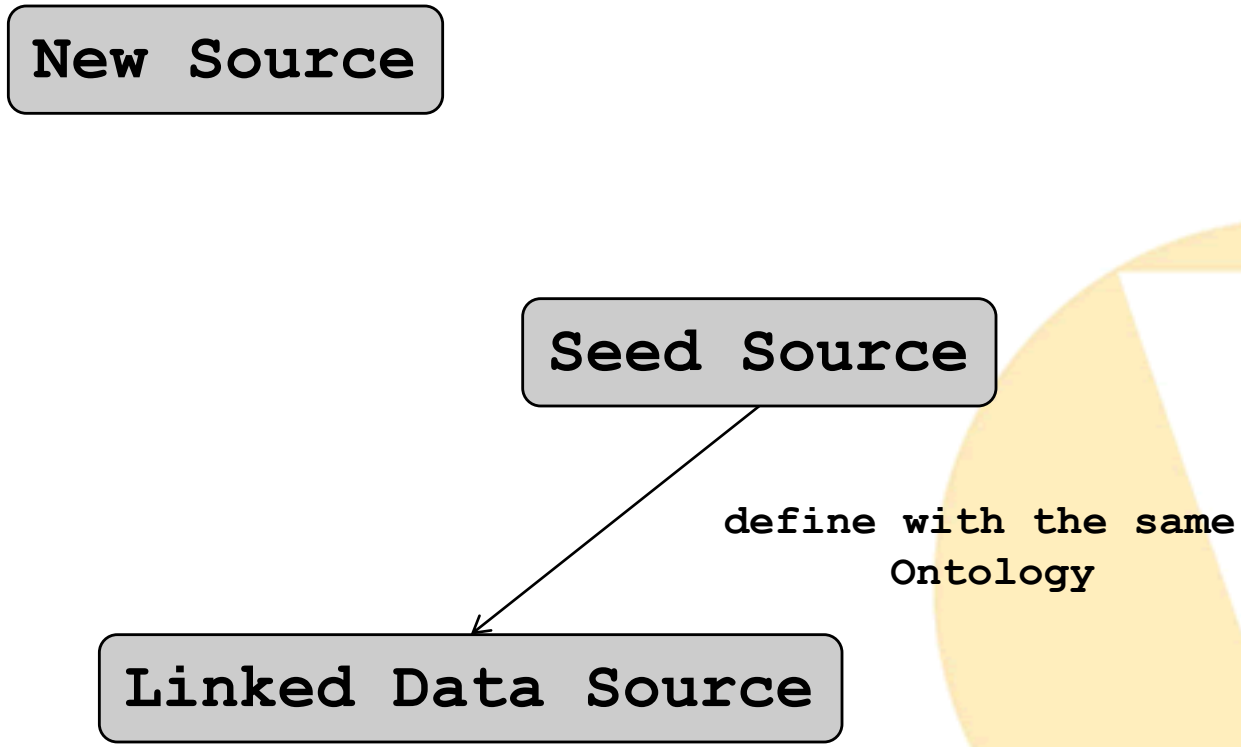
Definition of the discovered Source

```
googlefinance(FundSymbol,FundName,…)
  :-yahoofinance(FundSymbol,…,FundName)
```

USC Viterbi School of Engineering

Linking the *Deep Web* Sources into LDW

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

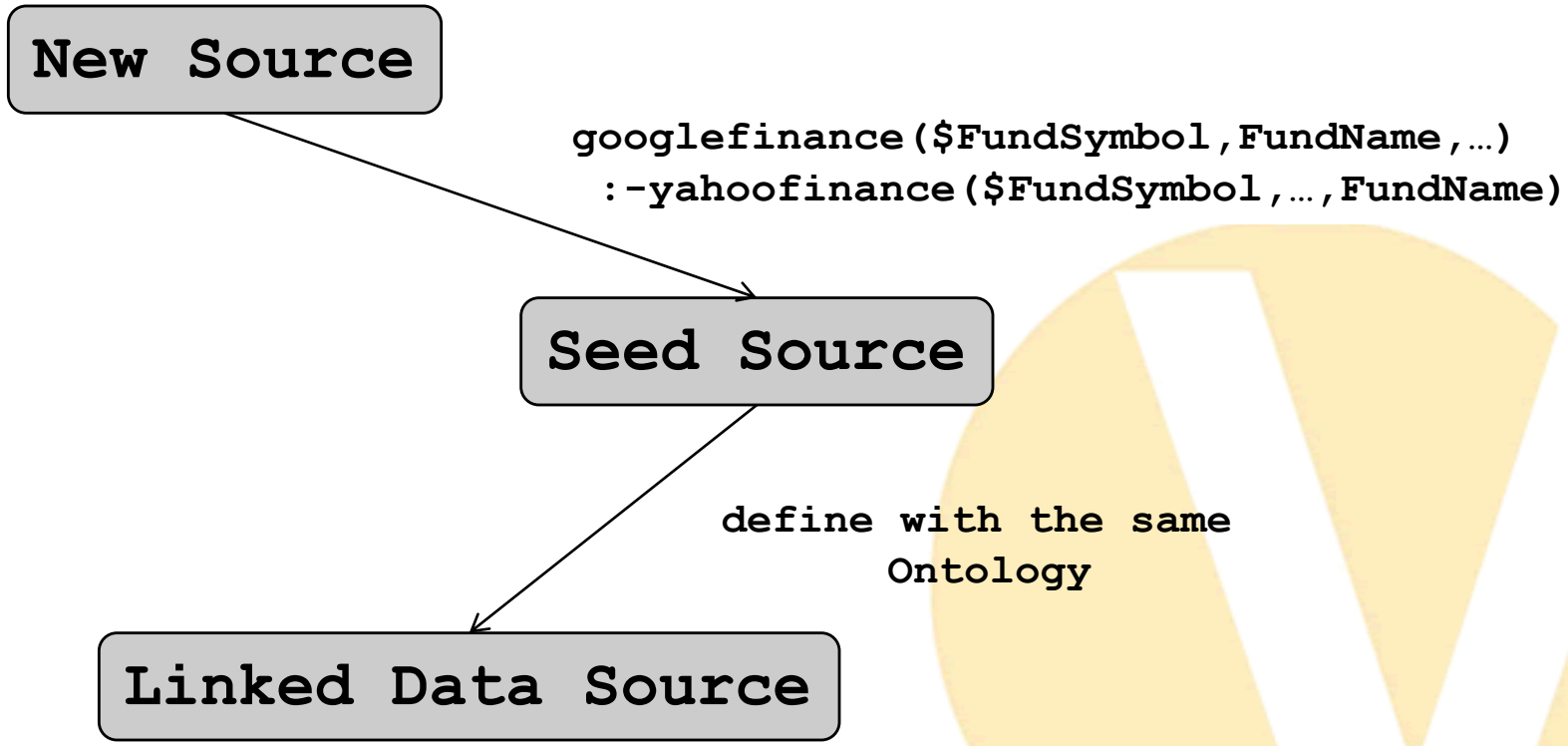


USC

USC Viterbi School of Engineering

Linking the *Deep Web* Sources into LDW

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

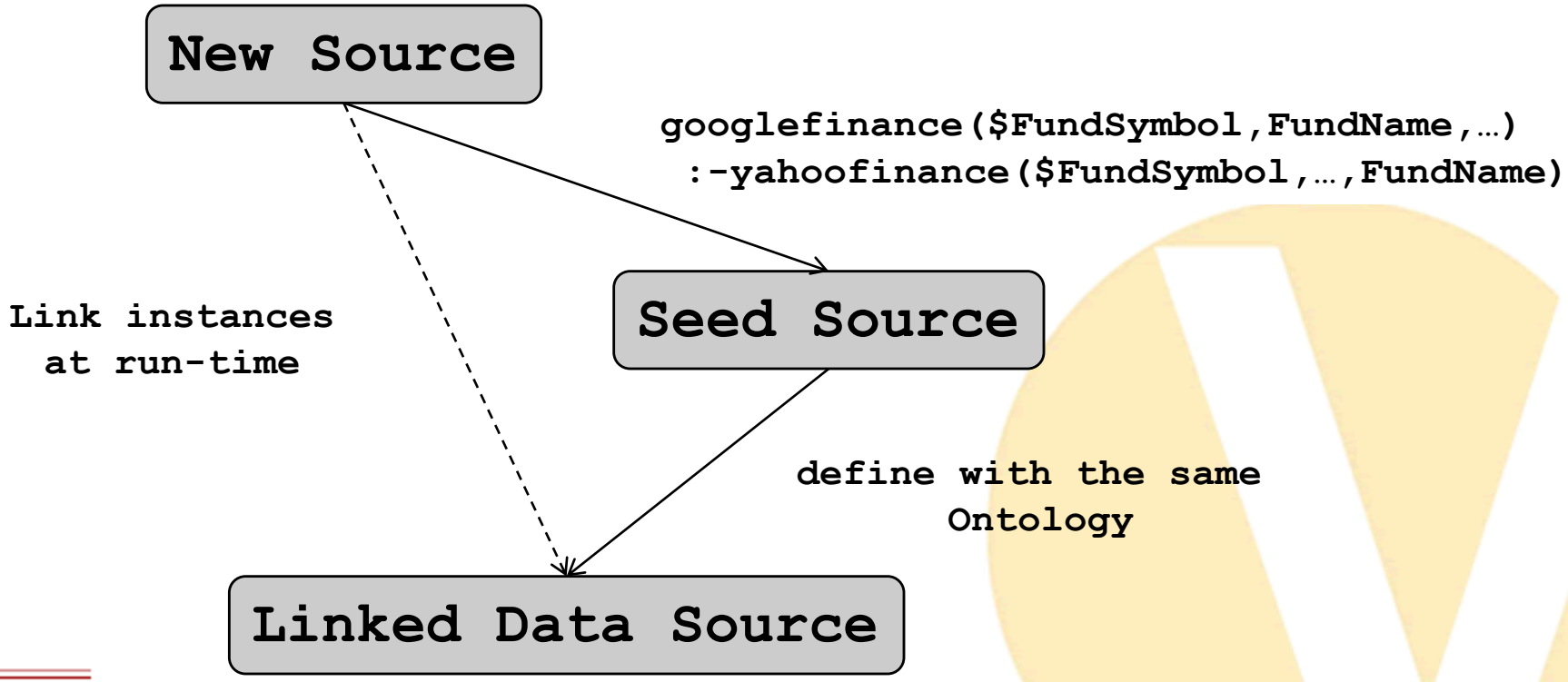


USC

USC Viterbi School of Engineering

Linking the *Deep Web* Sources into LDW

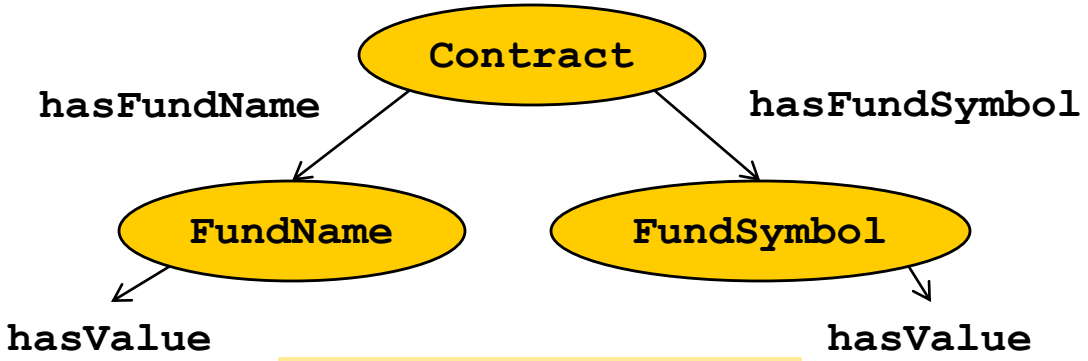
- **Instances generated by the Semantic Web Service need to be linked to existing Individuals in the LDW**



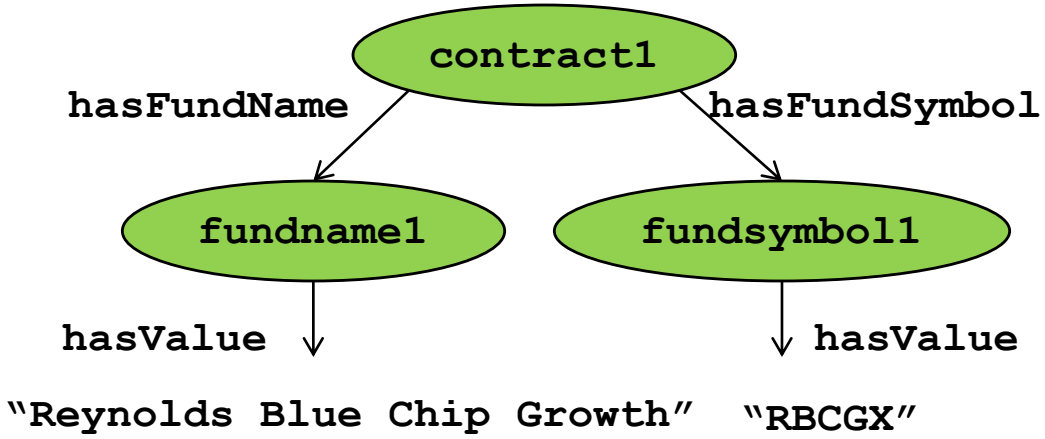
USC

USC Viterbi School of Engineering

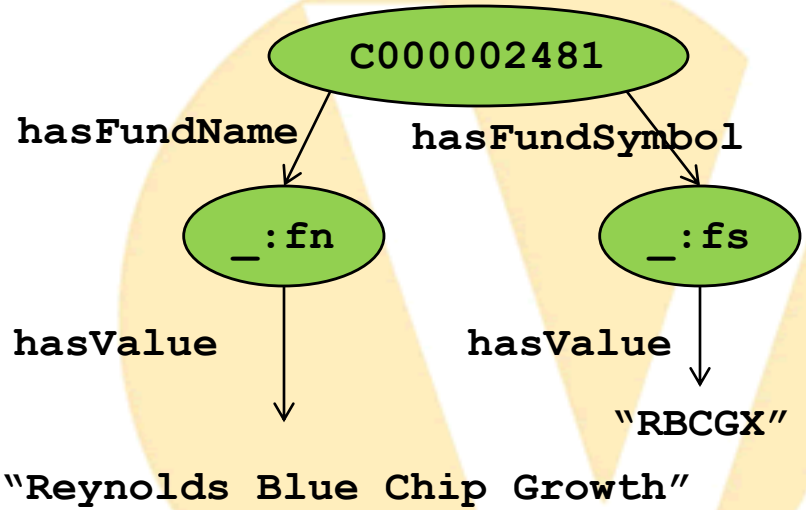
Linking the Seed Source to the LDW



Common Ontology



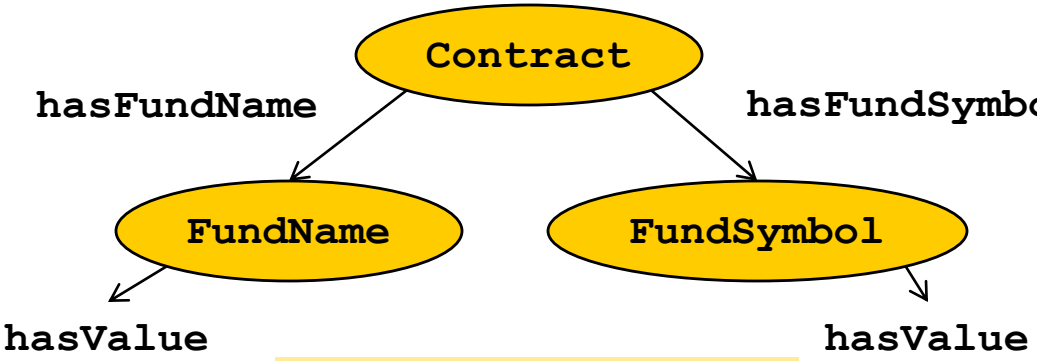
SWS Instances



LDS Instances

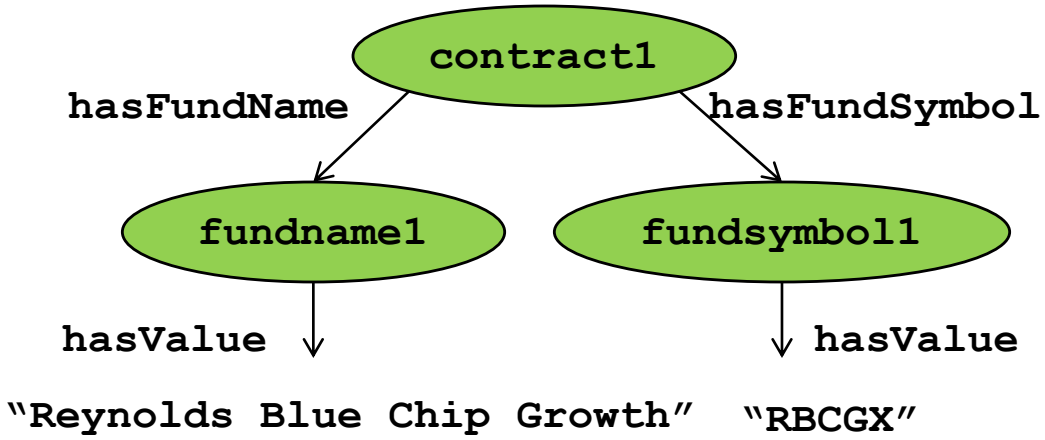
USC Viterbi School of Engineering

Linking the Seed Source to the LDW

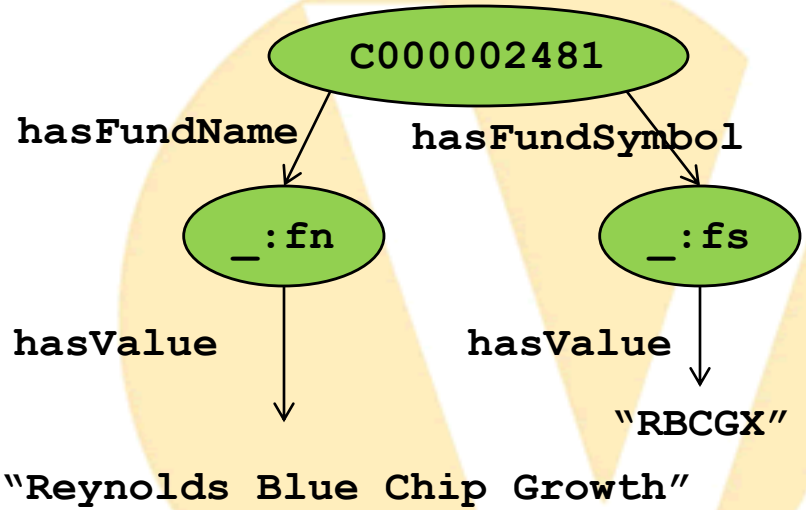


Common Ontology

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



SWS Instances



LDS Instances

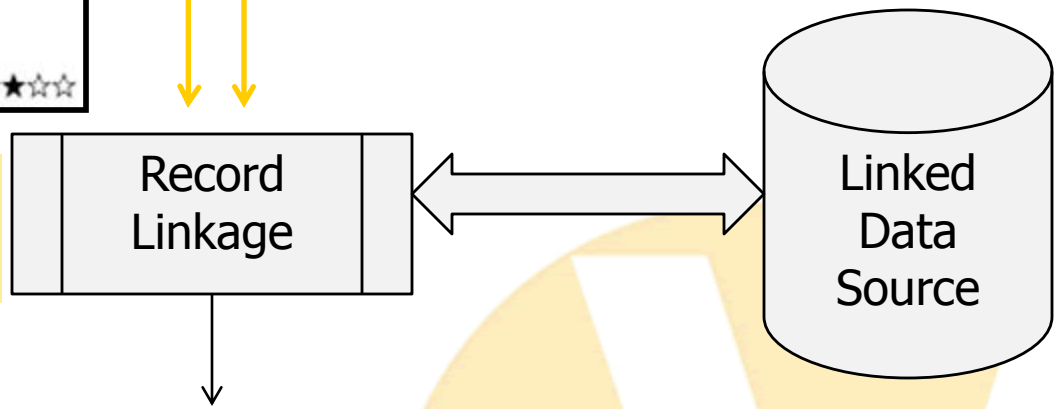
USC Viterbi School of Engineering

Linking the New Source to the LDW

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

```
contract1 rdf:type Contract .
symbol1 rdf:type Symbol .
contract1 hasSymbol symbol1 .
symbol1 hasValue "RBCGX" .
name1 rdf:type Name .
contract1 hasName name1 .
name1 hasValue "Reynolds Blue Chip Growth" .
...
contract1 owl:sameAs
http://www.rdfabout.com/rdf/usgov/sec/id/C000002481.
```

USC

USC Viterbi School of Engineering

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

CIK Series Class/Contract	Name	Ticker Symbol
0000832574	REYNOLDS FUNDS INC	
S000000865	REYNOLDS BLUE CHIP GROWTH FUND	
C000002481	REYNOLDS BLUE CHIP GROWTH FUND	**RBCGX**

  - 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

USC

USC Viterbi School of Engineering

Preliminary Results

- **Sources discovered by the previous work**
  - http://www.google.com/finance
  - http://moneycentral.msn.com/investor/home.asp
  - 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

USC

USC Viterbi
School of Engineering

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

USC

USC Viterbi
School of Engineering

USC