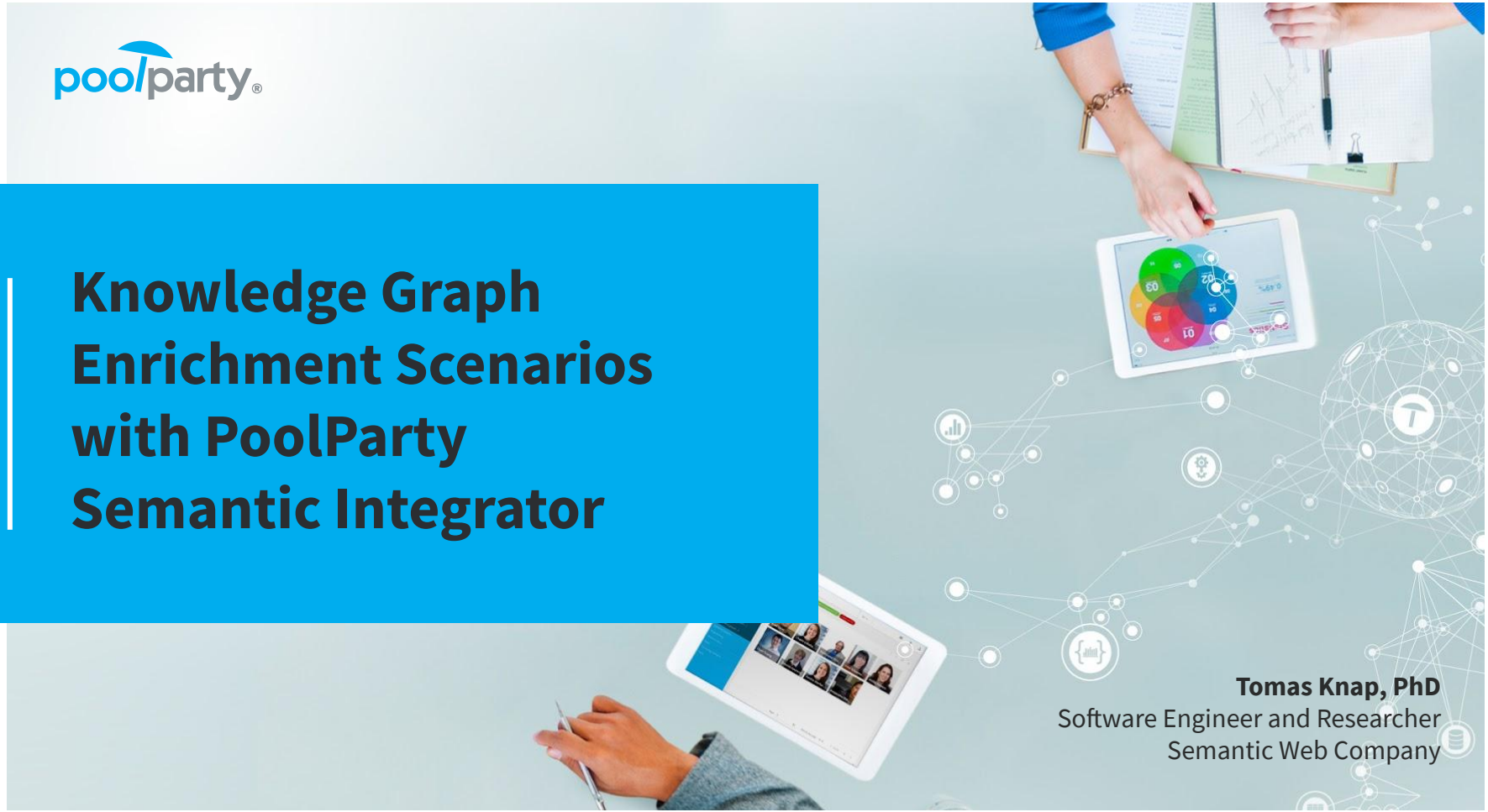
Knowledge Graph Enrichment Scenarios with PoolParty Semantic Integrator



Tomas Knap, PhD
Software Engineer and Researcher
Semantic Web Company

Introducing Semantic Web Company

Founded in **2004**

Based in Vienna

Privately held

50+ **FTE**

Software Engineers & Consultants for **NLP**, **Semantics** and **Machine learning**

Developer & Vendor of **PoolParty Semantic Suite**

Participating in projects with €2.5 million funding for **R&D**

~35%

revenue growth/year

SWC named to KMWorld's **'100 Companies That Matter in Knowledge Management'** in 2016, until 2019

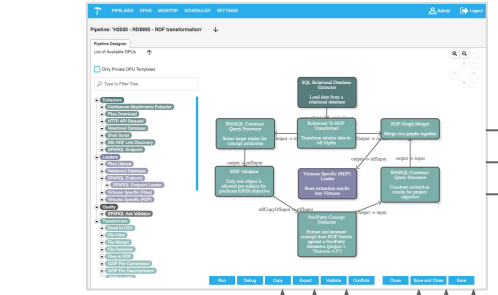
ISO 27001:2013 certified

4,500 followers

© Semantic Web Company 2019

Agenda

- PoolParty Semantic Suite
- Knowledge Graphs (KG)
- KG Enrichment Scenarios
  - Unstructured Data
    - Discover concepts in the documents
    - Link concepts to external knowledge graphs
    - ...
  - Structured Data - CSV, JSON, relational tables
    - Align schemas of entities
    - Link attributes to the entities in the KG
      - Discover duplicates in the KG
    - Fuse different representations of entities
    - ...
- Summary



Bain Capital is a venture capital company based in Boston, MA. Since inception it has invested in hundreds of companies including AMC Entertainment, Brookstone, and Burger King. The company was co-founded by Mitt Romney.

Vendor City	Vendor State	Vendor Zip Code
Fairfax	VA	22030
Washington	DC	20016
Mount Airy	MD	21771
Washington	DC	20022
Rockville	MD	20851
Landover Hills	MD	20784
Potomac	MD	20854

© Semantic Web Company 2019

PoolParty Suite

© Semantic Web Company 2019

Fact sheet: PoolParty Semantic Suite

Most complete and secure **Semantic Middleware** on the Global Market

Semantic AI: Fusion Graphs, NLP, and Machine Learning

ISO 27001:2013 certified

W3C standards compliant

First release in 2009

Current version **7.1**

On-premises or **cloud-based**

Over **200** **installations world-wide**

Named as Sample Vendor in **Gartner's Hype Cycle for AI 2019**

Named as Representative Vendor in **Gartner's Market Guide for Hosted AI Services 2018**

KMWorld listed PoolParty as **Trend-Setting Product** 2015, 2016, 2017, 2018, and 2019

© Semantic Web Company 2019

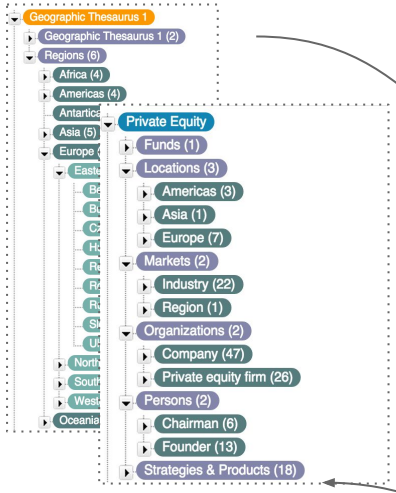
How does it work?

The Knowledge Engineer's perspective

poolparty

https://www.poolparty.biz/

Taxonomy & Ontology Server



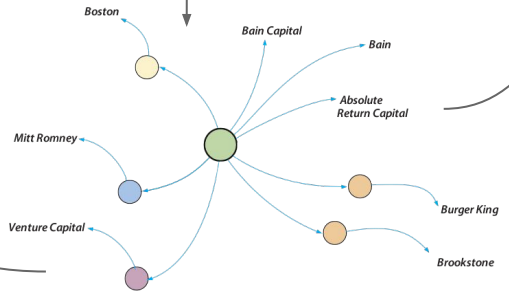
Identify new candidate concepts to be included in a controlled vocabulary

Entity Extractor & Semantic Classifier

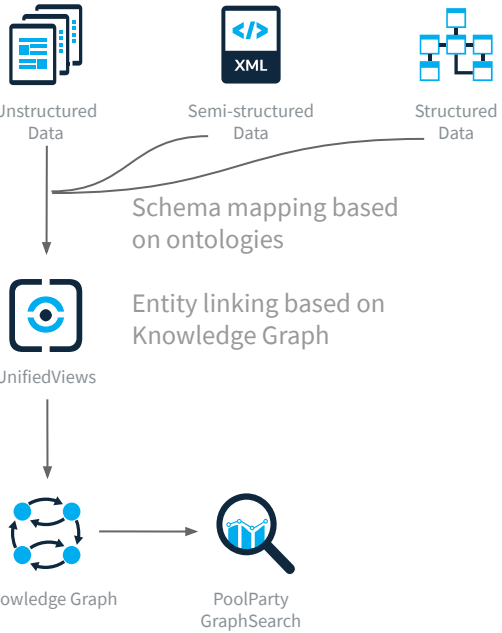
Controlled vocabularies as a basis for highly precise knowledge extraction and text classification

Bain Capital is a venture capital company based in Boston, MA. Since inception it has invested in hundreds of companies including AMC Entertainment, Brookstone, and Burger King. The company was co-founded by Mitt Romney.

Factsheet



Data Integration & Data Linking



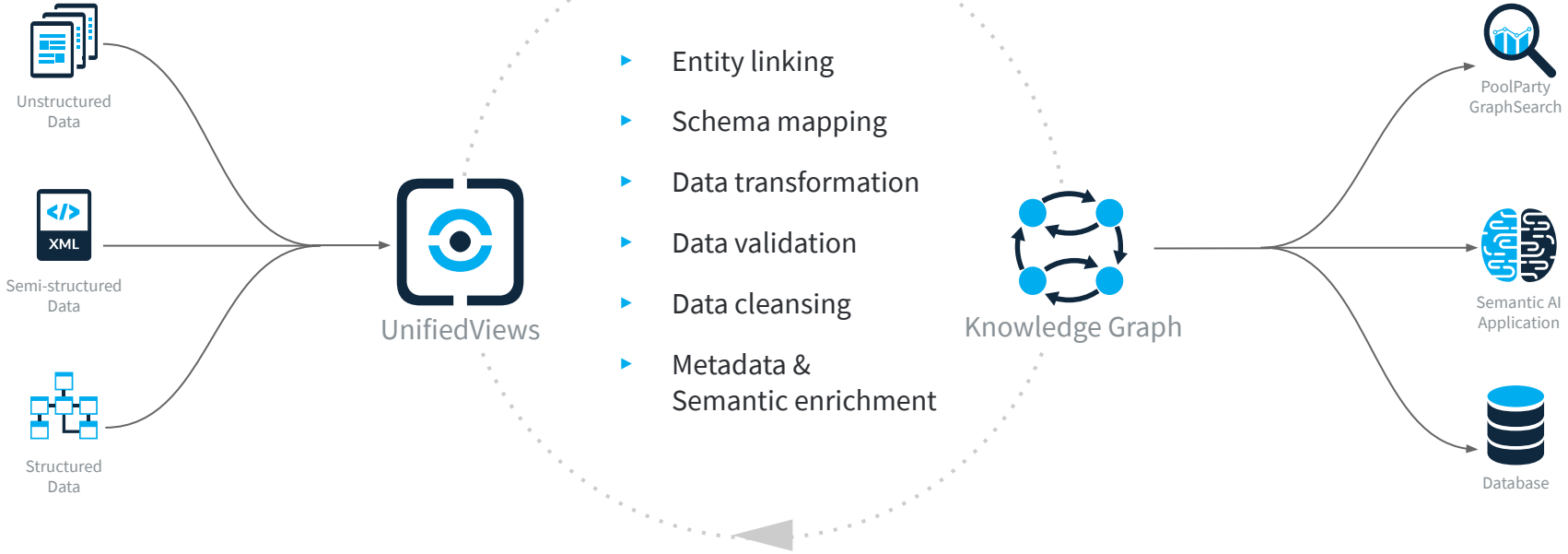
© Semantic Web Company 2019

How does it work?

The Data Engineer's perspective

poolparty

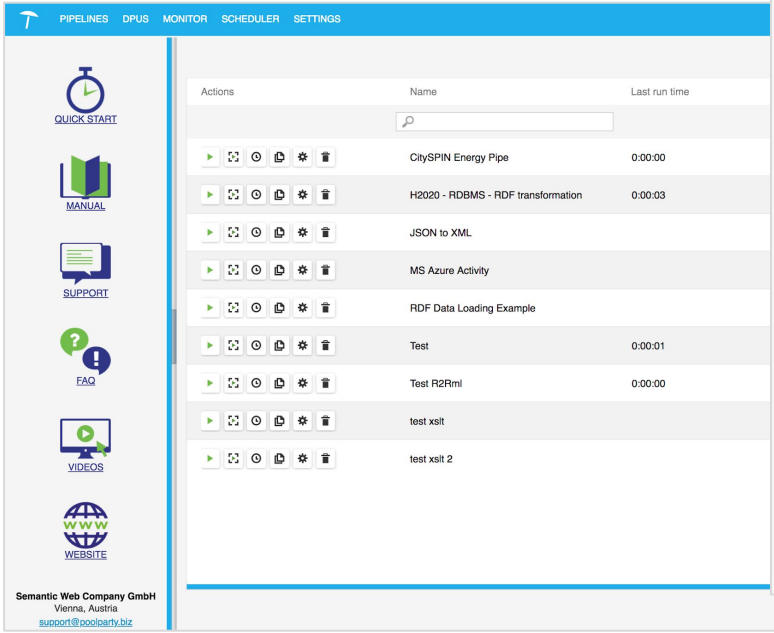
https://www.poolparty.biz/

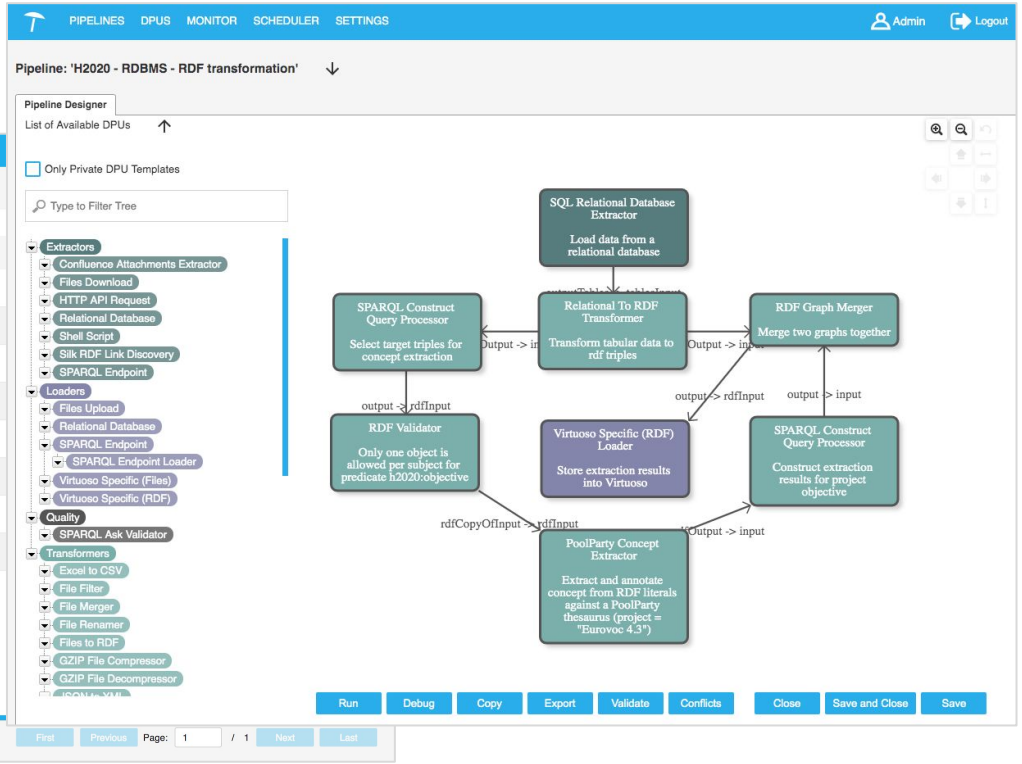


© Semantic Web Company 2019

PoolParty UnifiedViews for Data Engineers

Knowledge Graph Automation and Linked Data Orchestration





© Semantic Web Company 2019

Knowledge Graphs

© Semantic Web Company 2019

Knowledge Graphs (KG)

- “Network of all kinds of things which are relevant to the organization”
  - created by subject-matter experts with the help of tools and machine learning algorithms
- KG provides common interface to query the data

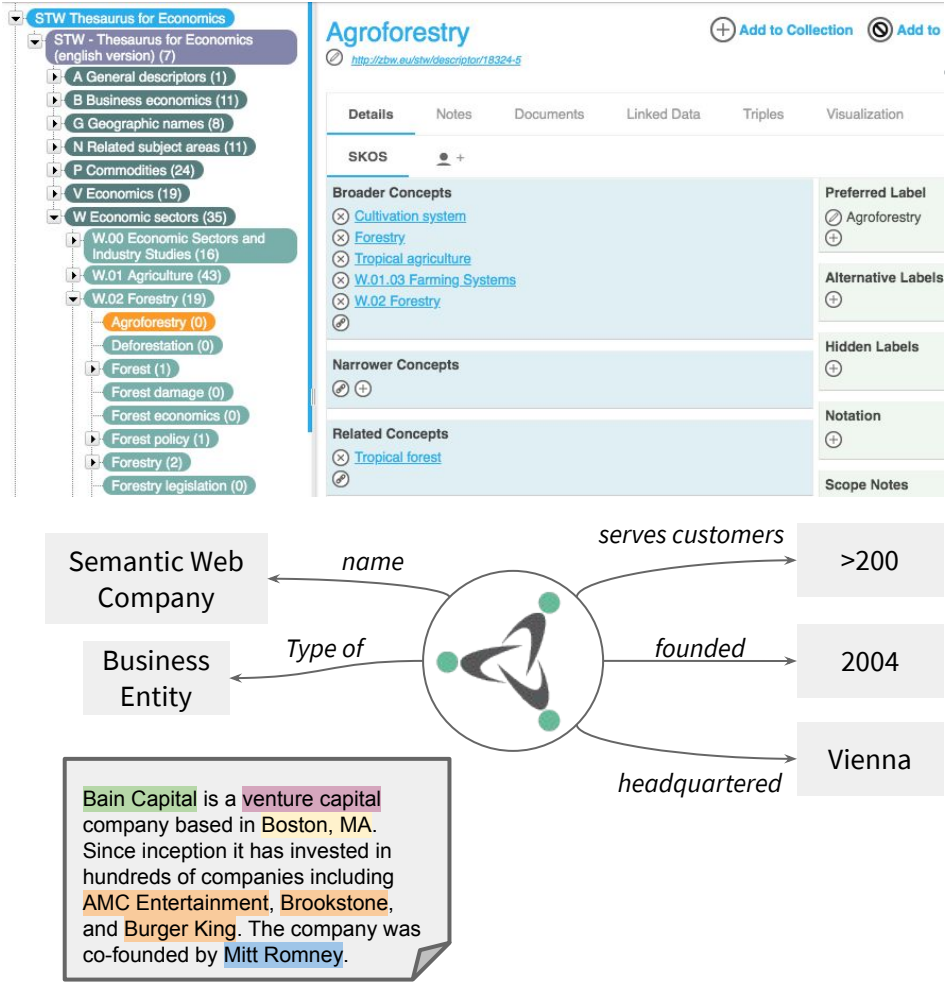
https://graphviews.poolparty.biz/

https://semantic-web.com/2014/07/15/from-taxonomies-over-ontologies-to-knowledge-graphs/

© Semantic Web Company 2019

Knowledge Graphs

- **Abstract concepts** from thesauri/taxonomies, together with **related** concepts
- Real world **entities,** with its **attributes and relations, classes of entities**
  - Semantic Web Company, Business Entity, Projects
- **Documents**
  - links to concepts/entities discovered



© Semantic Web Company 2019

KB Enrichment Scenarios

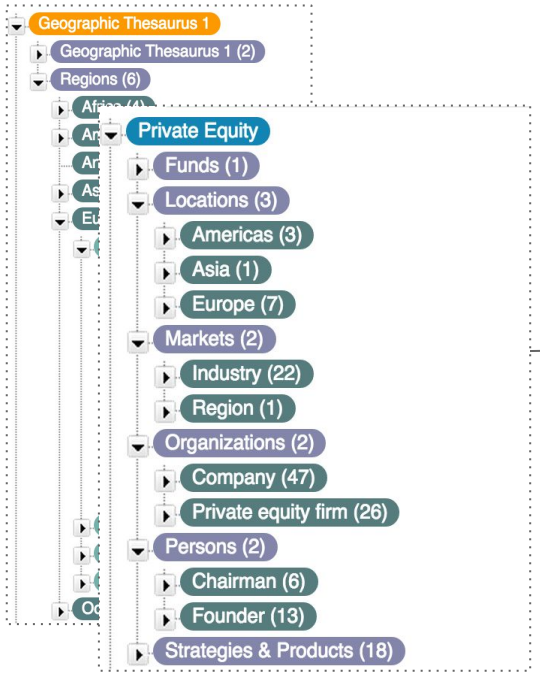
Unstructured Data

© Semantic Web Company 2019

KB Enrichment Scenarios for Unstructured Data

- Discover concepts in the documents based on the KG, concept co-occurrences
- Suggest new concepts
- **Link to concepts in external knowledge bases (DBpedia)**
  - Fetch attributes/relations
- Recognize (named) entities
- Classify documents

Knowledge Graph

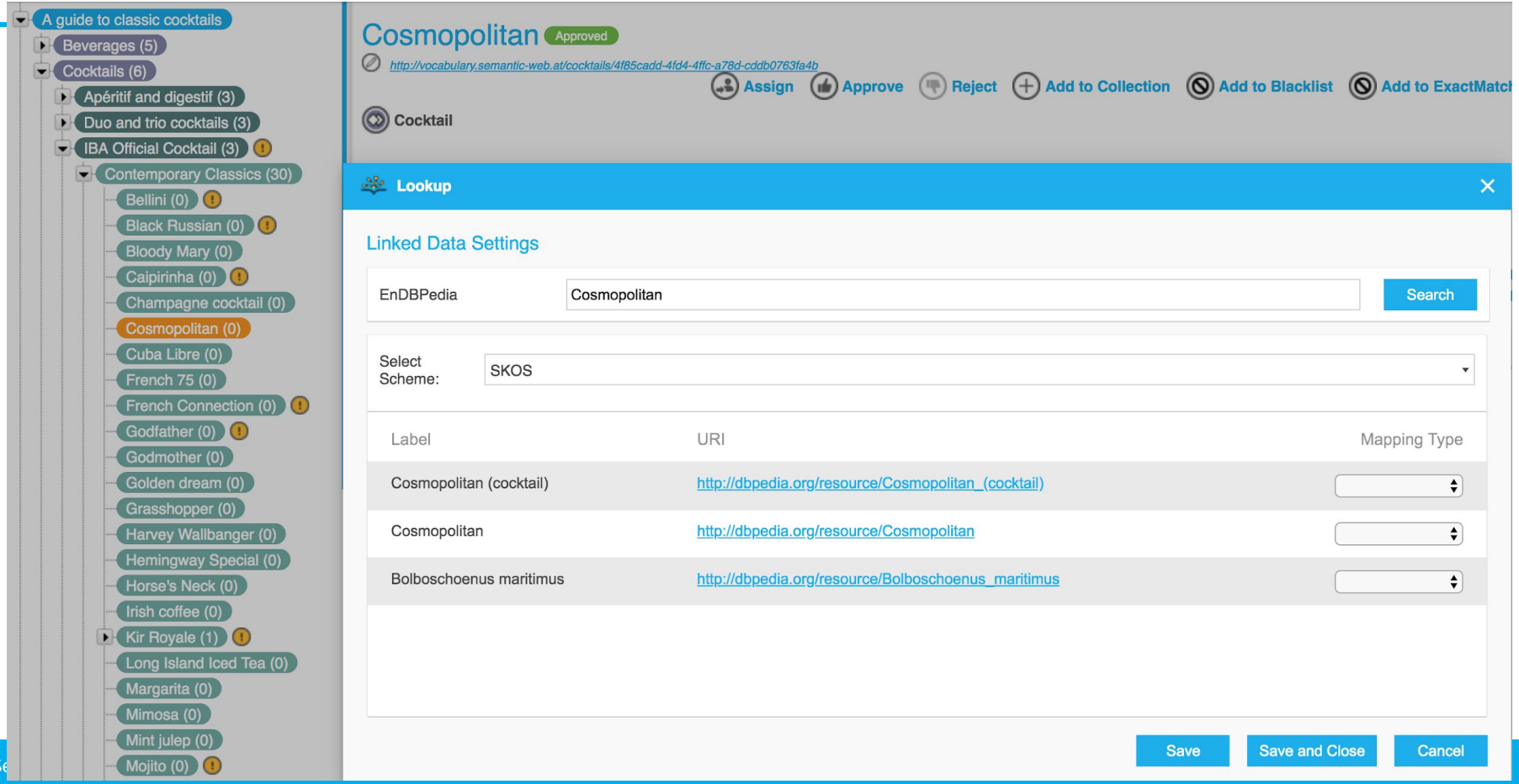


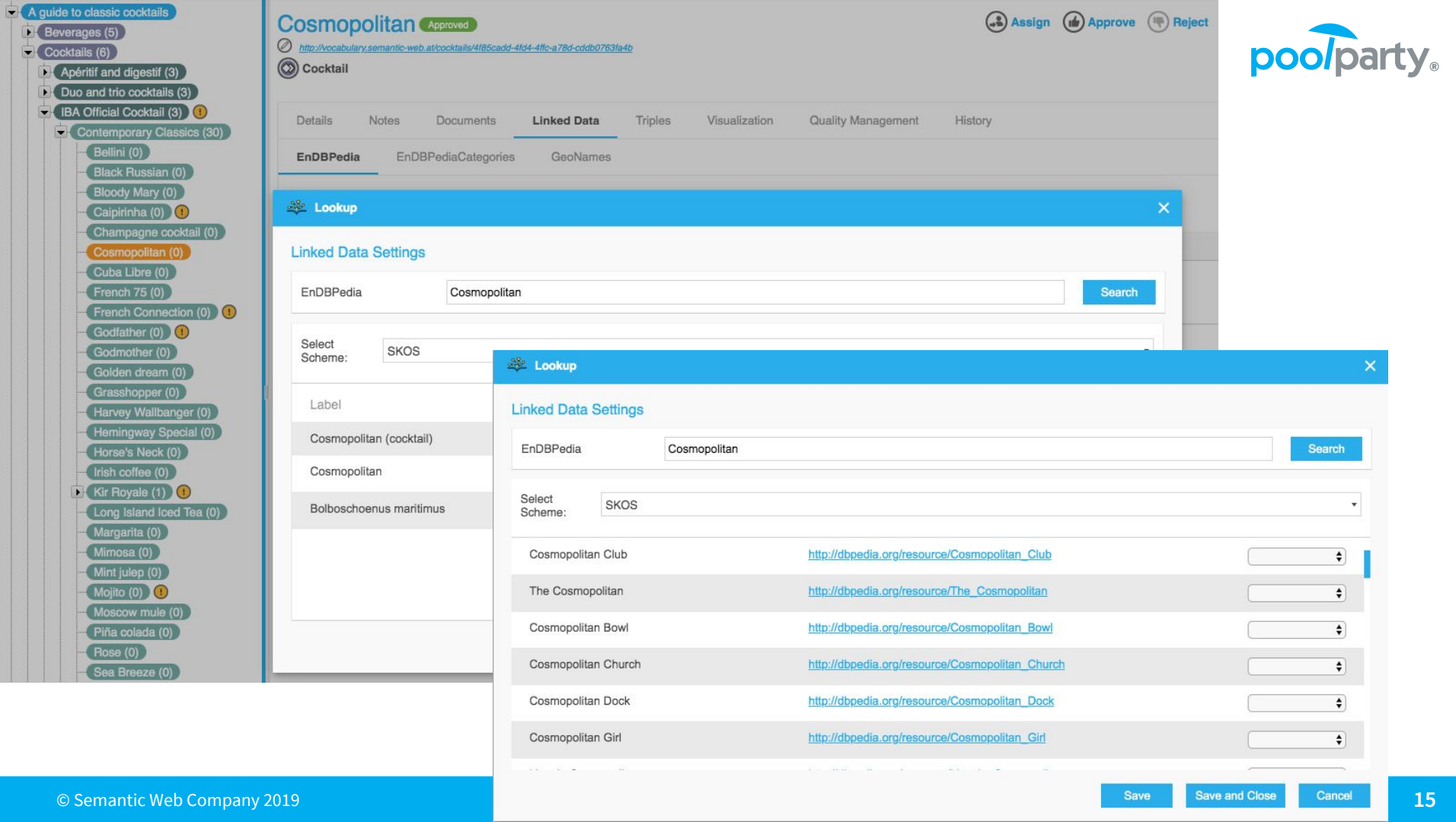
Entity Extraction & Text Mining

Bain Capital is a venture capital company based in Boston, MA. Since inception it has invested in hundreds of companies including AMC Entertainment, Brookstone, and Burger King. The company was co-founded by Mitt Romney.

© Semantic Web Company 2019

Linking Concepts to DBpedia





© Semantic Web Company 2019

About: Cosmopolitan (cocktail)

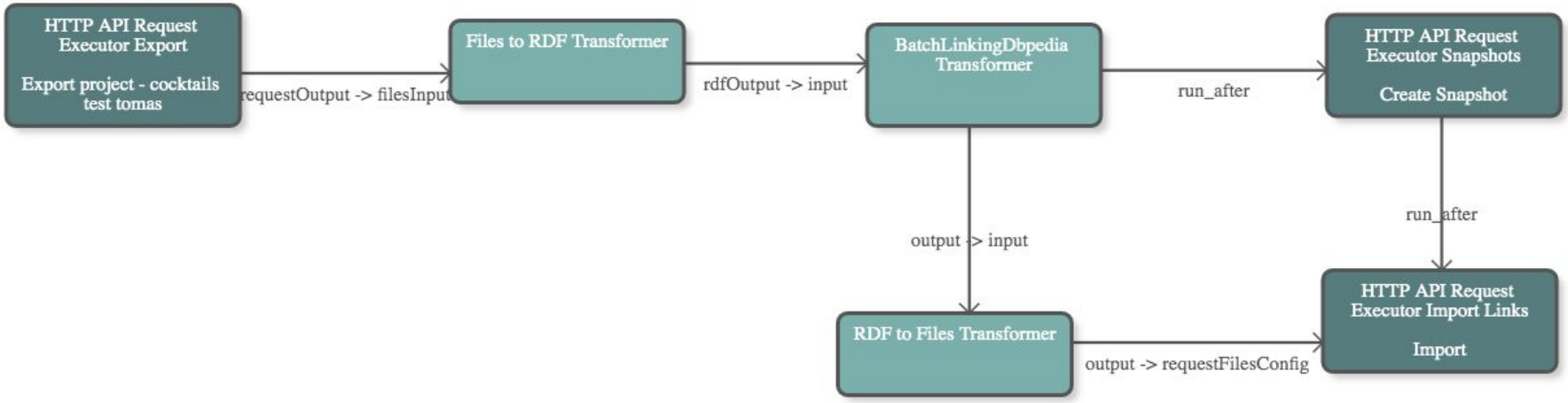
An Entity of Type : Food, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

A cosmopolitan, or informally a cosmo, is a cocktail made with vodka, triple sec, cranberry juice, and freshly squeezed or sweetened lime juice.

Property	Value
dbo:abstract	• A cosmopolitan, or informally a cosmo, is a cocktail made with vodka, triple sec, cranberry juice, and freshly squeezed or sweetened lime juice. (en)
dbo:thumbnail	• wiki-commons:Special:FilePath/Cosmopolitan_(5076906532).jpg?width=300
dbo:wikiPageID	• 409113 (xsd:integer)
dbo:wikiPageRevisionID	• 732587202 (xsd:integer)
dbp:caption	• A cosmopolitan garnished with a lemon twist
dbp:drinkware	• cocktail
dbp:garnish	• Lemon slice, lime wedge
dbp:iba	• yes
dbp:ingredients	• * 4 cl Vodka Citron * 1.5 cl Cointreau * 1.5 cl Fresh lime juice * 3 cl Cranberry juice

Linking Concepts to DBpedia - Batchlinking

- UnifiedViews pipeline for batchlinking the thesaurus to DBpedia
  - Solr to provide efficient search for potential candidates
  - Spreading activation to take into account context of the concept



Linking Concepts - Batchlinking Evaluation

- STW Thesaurus for Economics
  - http://zbw.eu/stw/version/latest/about
  - ~3500 links discovered
  - Compared against gold standard provided by STW (~2600 links) and manually extended (~900 links)
- Results:
  - Precision **96%**
    - The suggested link was correct
  - Recall almost 40%

© Semantic Web Company 2019

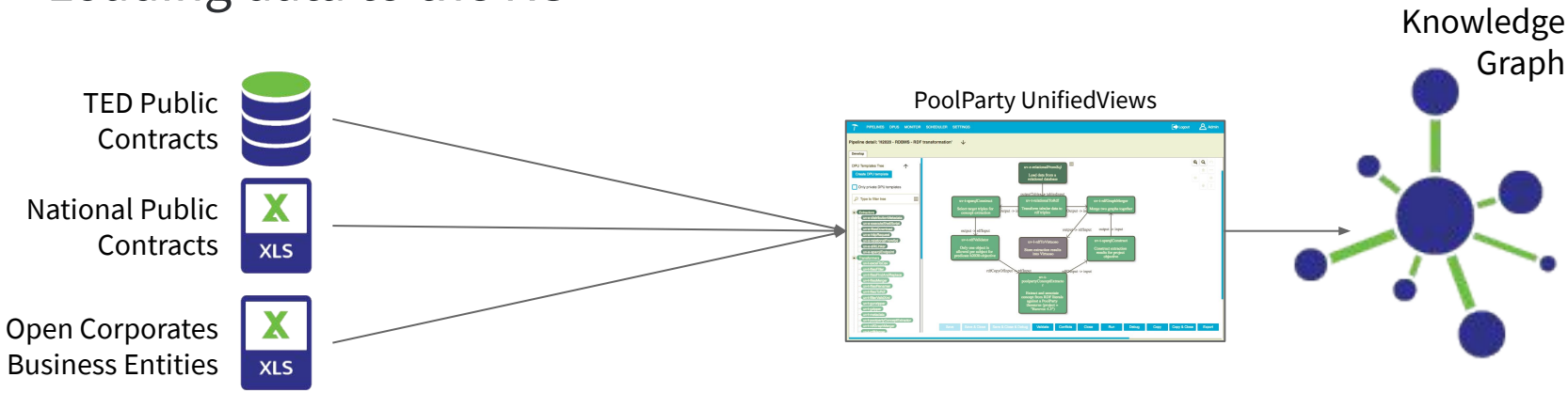
KB Enrichment Scenarios

Structured Data

© Semantic Web Company 2019

KG Enrichment Scenarios for Structured Data

- Data collection/pre-processing
- **Schema mapping, Entity linking, Data fusion**
- Quality assessment
- Loading data to the KG



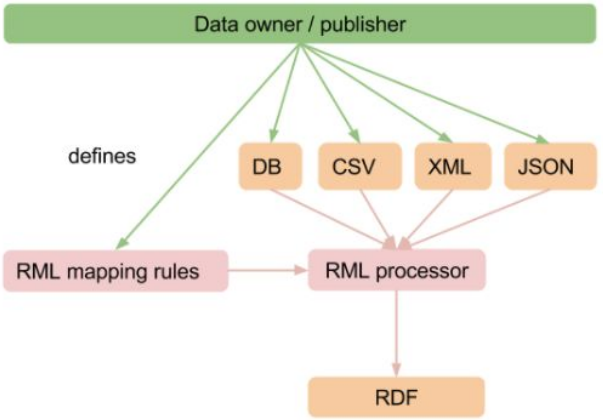
© Semantic Web Company 2019

Schema Mapping - Idea

- Map schemas of the input data sources to the target schemas maintained by PoolParty
  - Support for arbitrary sources (CSV, JSON, XML, relational data sources)
  - Suggestion of mapping rules
  - Support for data transformations

Contract Number	Vendor	Vendor City	Vendor State	Vendor Zip Code
1069374	Tom Woll and Associates	Fairfax	VA	22030
1053701	Setty & Associates International PLLC	Washington	DC	20016
1065909	Keller Brothers Inc	Mount Airy	MD	21771
9508000263AA	Ariana's Transportation Services, LLC	Washington	DC	20022
1069194	Community Ministries of Rockville, Inc.	Rockville	MD	20851
1020522	Family Services, Inc.	Landover Hills	MD	20784
1080123	Pallavi Belur	Potomac	MD	20854

Schema Mapping - UnifiedViews



- Schema mapping DPUs
  - R2RML (W3C Rec)
  - RML - generalization of R2RML
    - http://rml.io

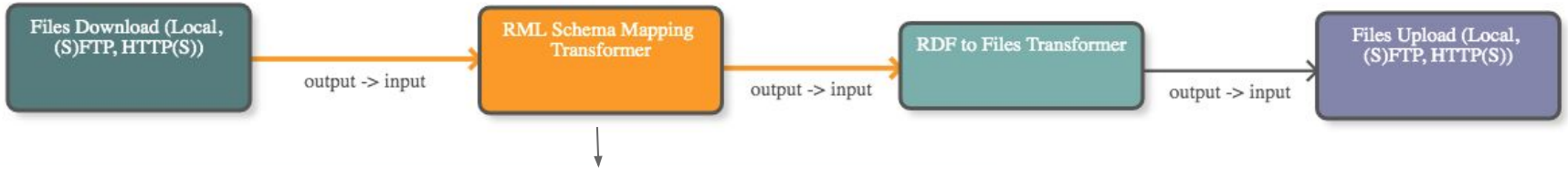
Step 2: Define mappings for subject columns

Subject CSV column	Object CSV column	Property
Contract Number	Vendor	Supplier
Vendor	Vendor Address Line 2	Address
Vendor Address Line 2	Vendor Address Line 1	Street
Vendor Address Line 2	Vendor Zip Code	Zip code
Vendor Address Line 2	Vendor City	City
Vendor Address Line 2	Vendor State	State

```
@prefix rml: <http://semweb.mmlab.be/ns/rml#>.
@prefix ql: <http://semweb.mmlab.be/ns/ql#>.
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@base <http://unifiedviews.eu/rml/csv/base> .
<#Contract+Number>
 rml:logicalSource [
   rml:source "file.csv" ;
   rml:referenceFormulation ql:CSV
 ];
 rr:subjectMap [
   rr:template "http://demo.semantic-web.com/data/contract/{Contract Number}";
   rr:class <http://purl.org/procurement/public-contracts#Contract>
 ];
 rr:predicateObjectMap [
   rr:predicate <http://purl.org/procurement/public-contracts#supplier>;
   rr:objectMap [
     rr:parentTriplesMap <#Vendor>
   ]
 ].
```

© Semantic Web Company 2019

Schema Mapping - UnifiedViews Pipeline



Step 2: Define mappings for subject columns

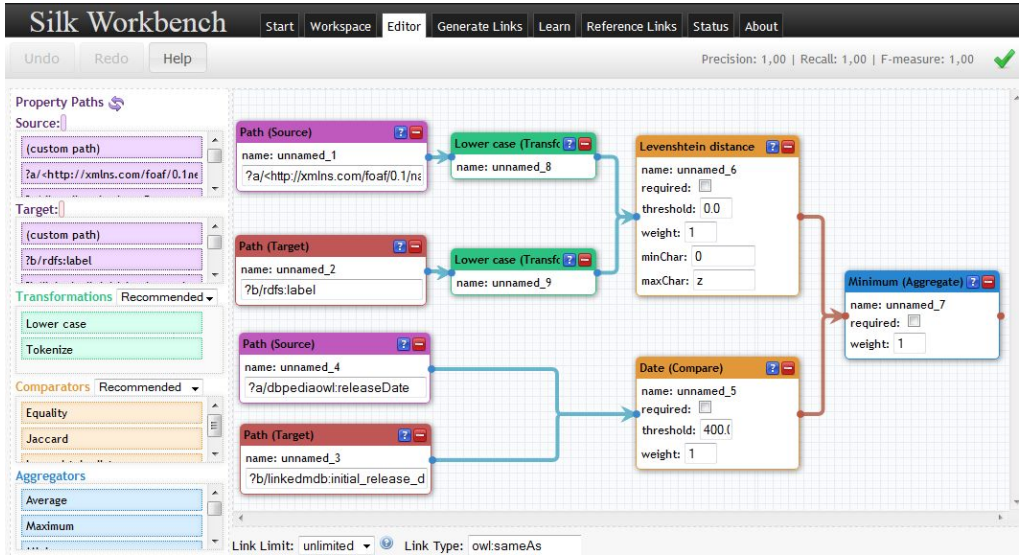
Subject CSV column	Object CSV column	Property
Contract Number	Vendor	Supplier
Vendor	Vendor Address Line 2	Address
Vendor Address Line 2	Vendor Address Line 1	Street
Vendor Address Line 2	Vendor Zip Code	Zip code
Vendor Address Line 2	Vendor City	City
Vendor Address Line 2	Vendor State	State

```
@prefix rml: <http://semweb.mmlab.be/ns/rml#>.
@prefix ql: <http://semweb.mmlab.be/ns/ql#>.
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@base <http://unifiedviews.eu/rml/csv/base> .
<#Contract+Number>
 rml:logicalSource [
  rml:source "file.csv" ;
  rml:referenceFormulation ql:CSV
 ];
 rr:subjectMap [
  rr:template "http://demo.semantic-web.com/data/contract/{Contract Number}";
  rr:class <http://purl.org/procurement/public-contracts#Contract>
 ];
 rr:predicateObjectMap [
  rr:predicate <http://purl.org/procurement/public-contracts#supplier>;
  rr:objectMap [
   rr:parentTriplesMap <#Vendor>
  ]
 ].
```

© Semantic Web Company 2019

Entity Linking - Idea

- Link attributes to the entities in the existing KG
- Discover duplicates between the acquired data and the KG

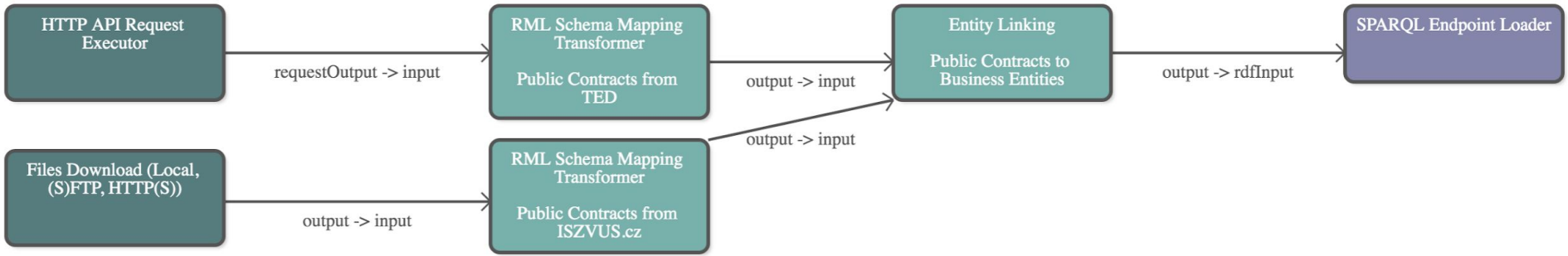



```
<Interlinks>
  <Interlink id="cities">
    <LinkType>owl:sameAs</LinkType>

    <SourceDataset dataSource="dbpedia" var="a">
      <RestrictTo>
        ?a rdf:type dbpedia:City
      </RestrictTo>
    </SourceDataset>

    <TargetDataset dataSource="geonames" var="b">
      <RestrictTo>
        ?b rdf:type gn:P
      </RestrictTo>
    </TargetDataset>

    <LinkageRule>
      <Aggregate type="average">
        <Compare metric="levenshteinDistance" threshold="1">
          <Input path="?a/rdfs:label" />
          <Input path="?b/gn:name" />
        </Compare>
        <Compare metric="num" threshold="1000" >
          <Input path="?a/dbpedia:populationTotal" />
          <Input path="?b/gn:population" />
        </Compare>
      </Aggregate>
    </LinkageRule>
```

© Semantic Web Company 2019

Entity Linking - UnifiedViews

- UnifiedViews DPUs
  - DPUs executing SPARQL (Update) queries
  - DPUs using Silk, http://silkframework.org/



© Semantic Web Company 2019

Data Fusion - Idea

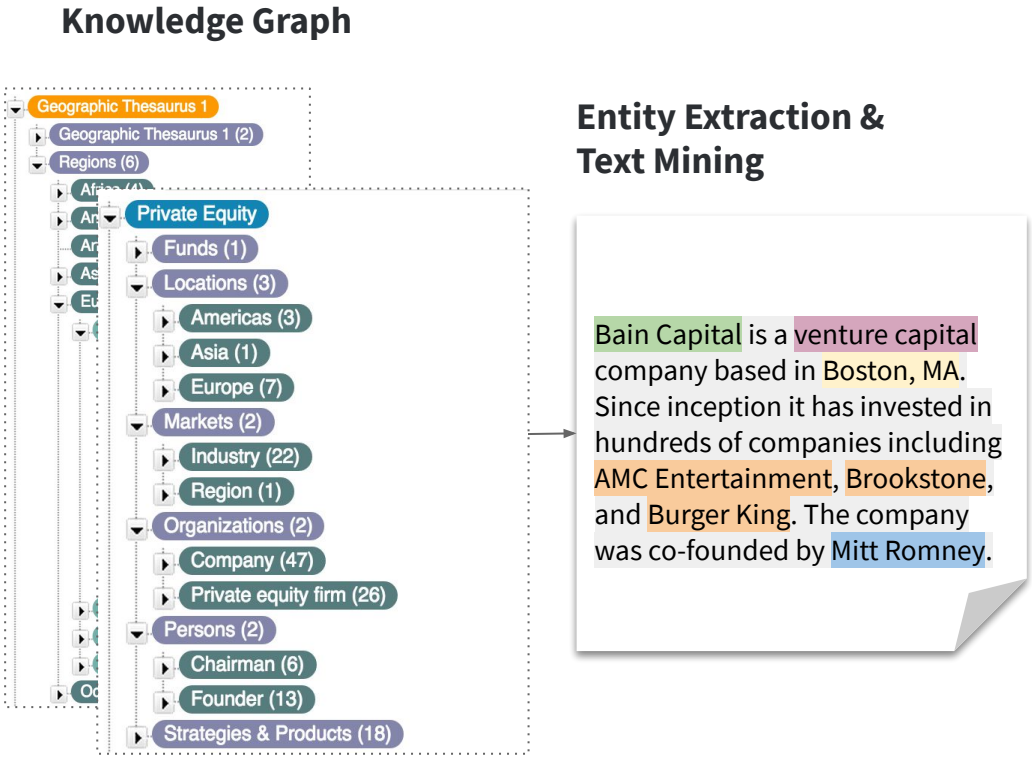
- Fuse different representations of the same resources
  - Define what happens when there are conflicting attributes/relations
  - Apply a conflict resolution policy

Company Name	Founded	Number of employees	Address
Food Council, Inc.	12/20/2004	30	...
Montgomery county Food Council, Inc.	12/21/2004	35	...

© Semantic Web Company 2019

Extracted Structured Data

- Unstructured data with:
  - Recognized (named) entities
  - Recognized attributes and relations of (named) entities
- Follow the recipe for structured data
  - Schema mapping
  - Entity Linking
  - Data Fusion



© Semantic Web Company 2019

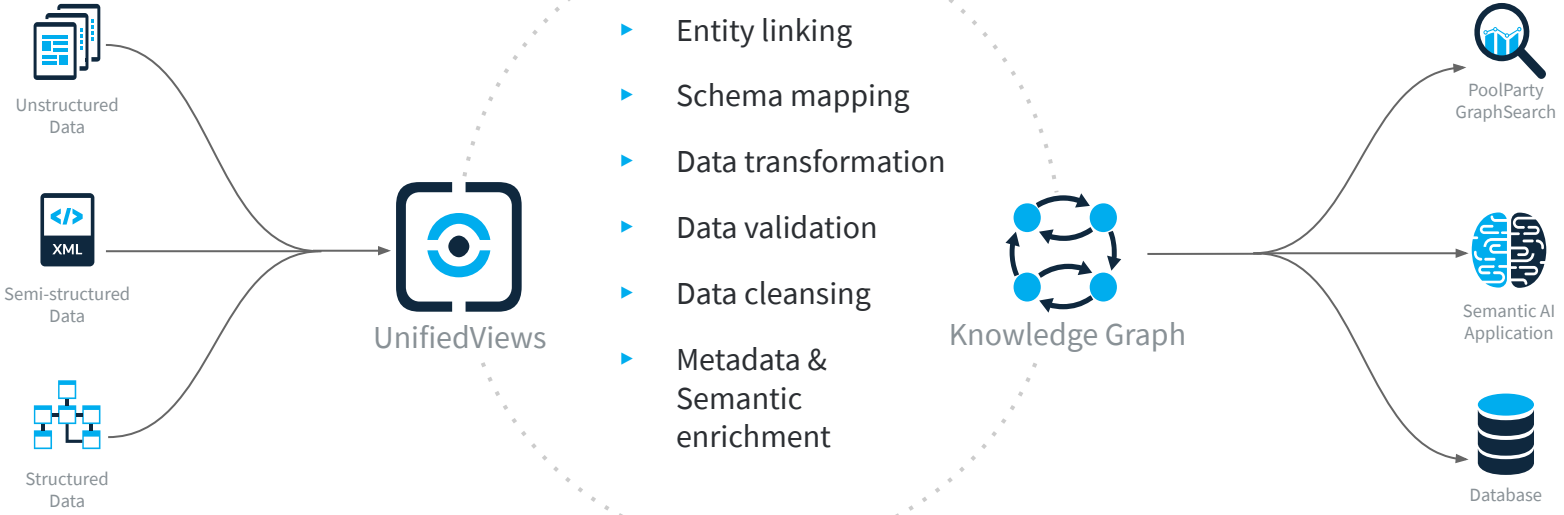
Summary

© Semantic Web Company 2019

Knowledge Graphs

- “PoolParty brings the power of Google into the enterprise environment through Knowledge Graphs.”

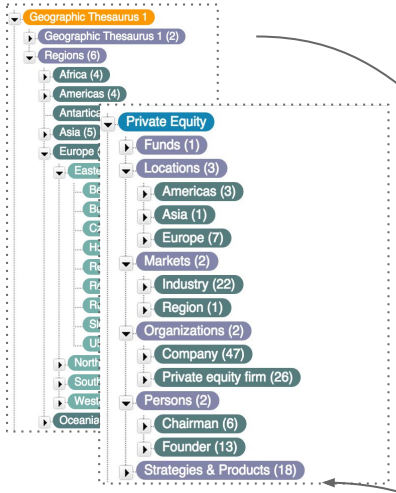
Knowledge Architect,
Top 3 Oil & Gas Company



© Semantic Web Company 2019

Knowledge Graph Enrichment Scenarios

Taxonomy & Ontology Server



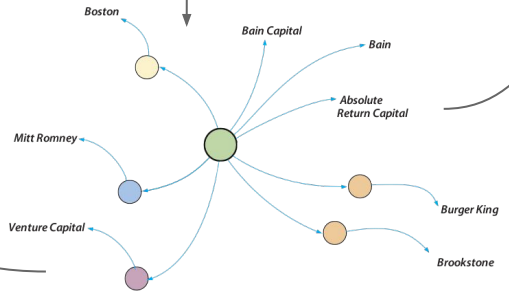
Identify new candidate concepts to be included in a controlled vocabulary

Entity Extractor & Semantic Classifier

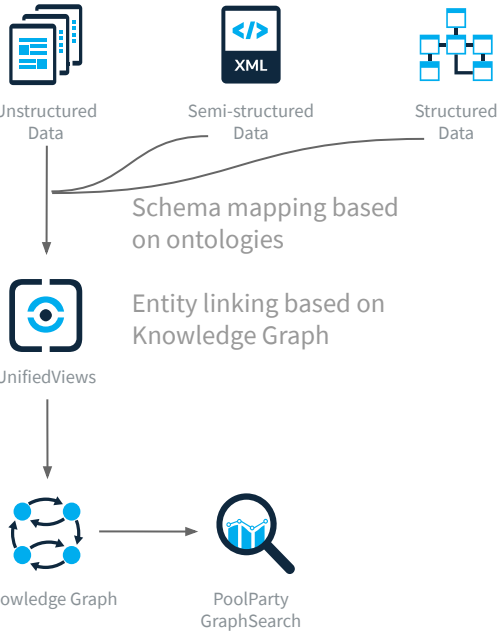
Controlled vocabularies as a basis for highly precise knowledge extraction and text classification

Bain Capital is a venture capital company based in Boston, MA. Since inception it has invested in hundreds of companies including AMC Entertainment, Brookstone, and Burger King. The company was co-founded by Mitt Romney.

Factsheet



Data Integration & Data Linking



© Semantic Web Company 2019

Thank You!

Tomas Knap, PhD

Software Engineer and Researcher,
Semantic Web Company

tomas.knap@semantic-web.com