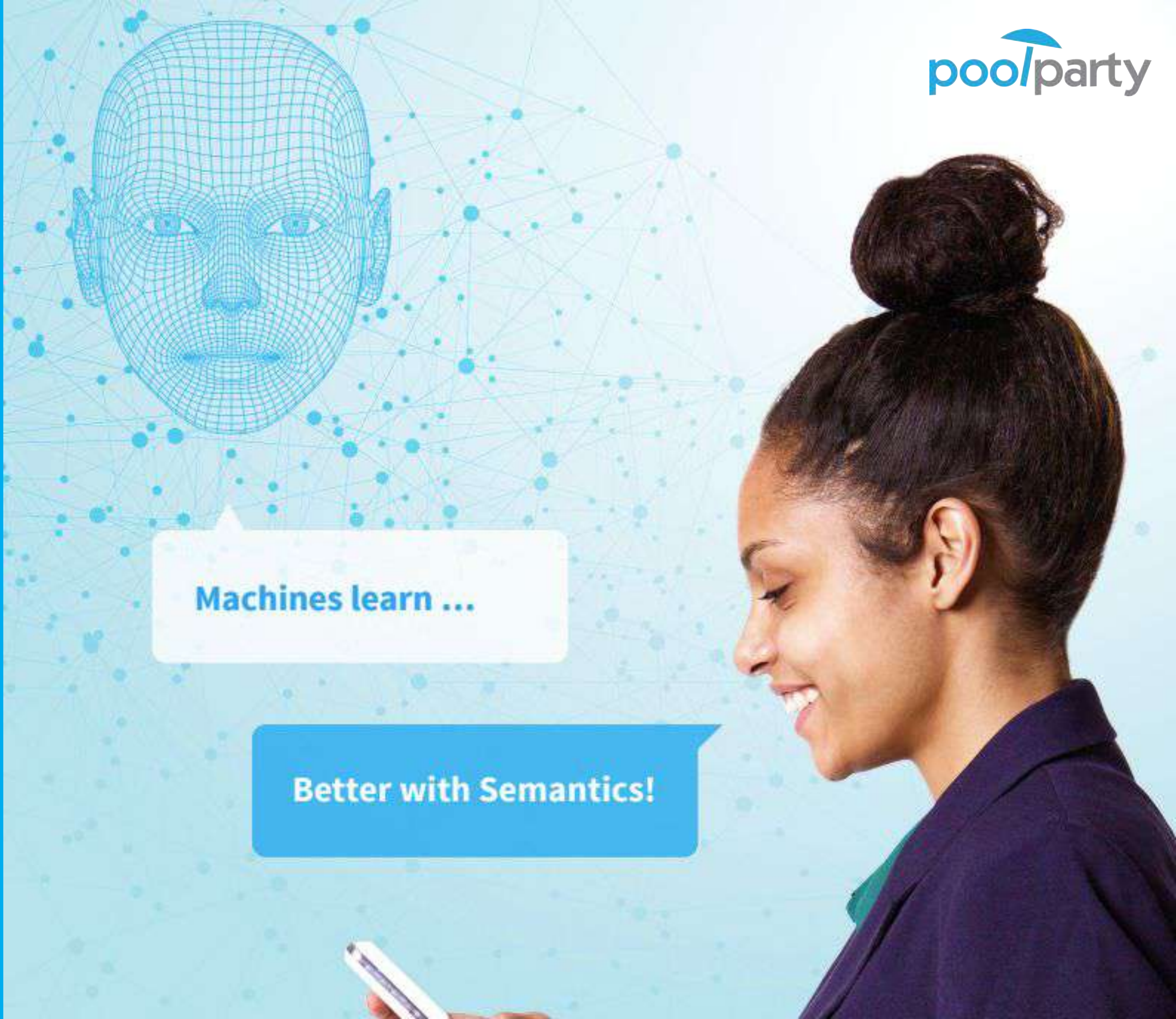
SHACL-based data life cycle management



Robert David
CTO
Albin Ahmeti
Data Engineer

Semantic Web Company /
PoolParty Semantic Suite

2

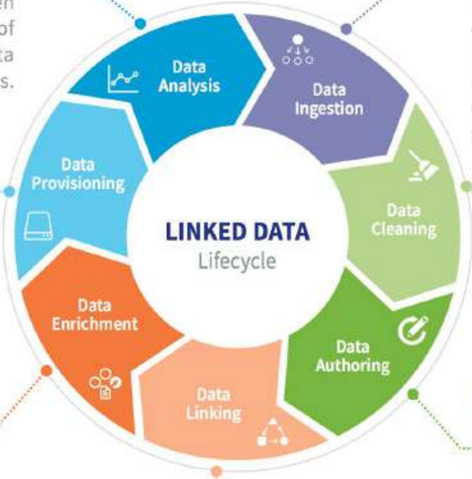
About

- Linked Data Lifecycle
- Software Components
- Data consistency requirements
- Data validation standards
- Validation use cases
- Live demo

3

Knowledge Graph Management

Along the Linked Data Life Cycle



ANALYSIS

Insights from linked data

Discover and analyze complex - and often hidden - relations in your data by means of customized data portals and state-of- the-art data visualization based on Semantic Data Lakes.

PROVISIONING

Data-driven personalization

Use semantic technologies to make your content and data management workflows smarter and more dynamic. Create personalized digital experiences. Feed your bots and machine learning algorithms with higher-quality data.

ENRICHMENT

Actionable data

Provide automatic tagging functions based on controlled vocabularies to ensure precise content classification. Semantic models help to enrich all content and data consistently and based on standards.

INGESTION

Avoid data migration

PoolParty is a semantic middleware that enriches and contextualizes your data with relevant information. Link and integrate different datasets by applying semantic knowledge models. Enterprise data gets smarter while remaining in the original storage system.

CLEANING

Data quality management

PoolParty allows organizations to ensure that their data is consistent, reliable and usable for analysis. Benefit from highly automated 'Data Repair' features.

AUTHORING

Knowledge graph capabilities

Enable collaboration and develop synergies between your departments by establishing defined knowledge graphs and controlled vocabularies

LINKING

The power of connected data

Link internal and external data resources and provide your organization with a customized and relevant knowledge platform.

4

Fact sheet: PoolParty

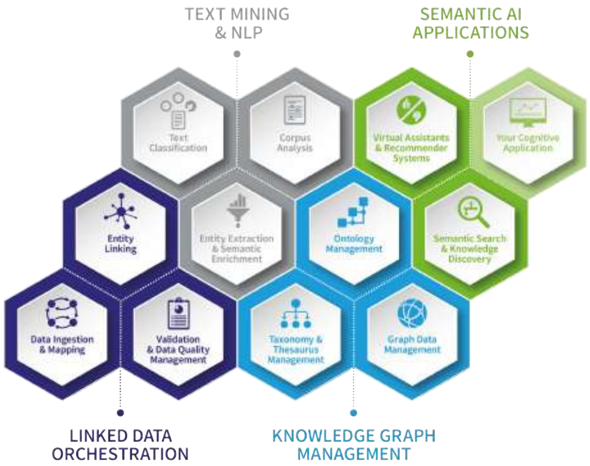
poolparty

PoolParty Semantic Suite

- Most complete Semantic Middleware on the Global Market
- *Semantic AI*: Fusion of Knowledge Graphs, NLP, and Machine Learning
- Linked Data Management along the whole Data Life Cycle
- W3C standards compliant

- First release in 2009
- Current version 7.0
- Over 200 installations world-wide

- On-premises or cloud-based
- KMWorld listed PoolParty as Trend-Setting Product 2015, 2016 and 2017
- www.poolparty.biz



TEXT MINING & NLP

SEMANTIC AI APPLICATIONS

Text Classification

Corpus Analysis

Virtual Assistants & Recommender Systems

Your Cognitive Application

Entity Linking

Entity Extraction & Semantic Enrichment

Ontology Management

Semantic Search & Knowledge Discovery

Data Ingestion & Mapping

Validation & Data Quality Management

Taxonomy & Thesaurus Management

Graph Data Management

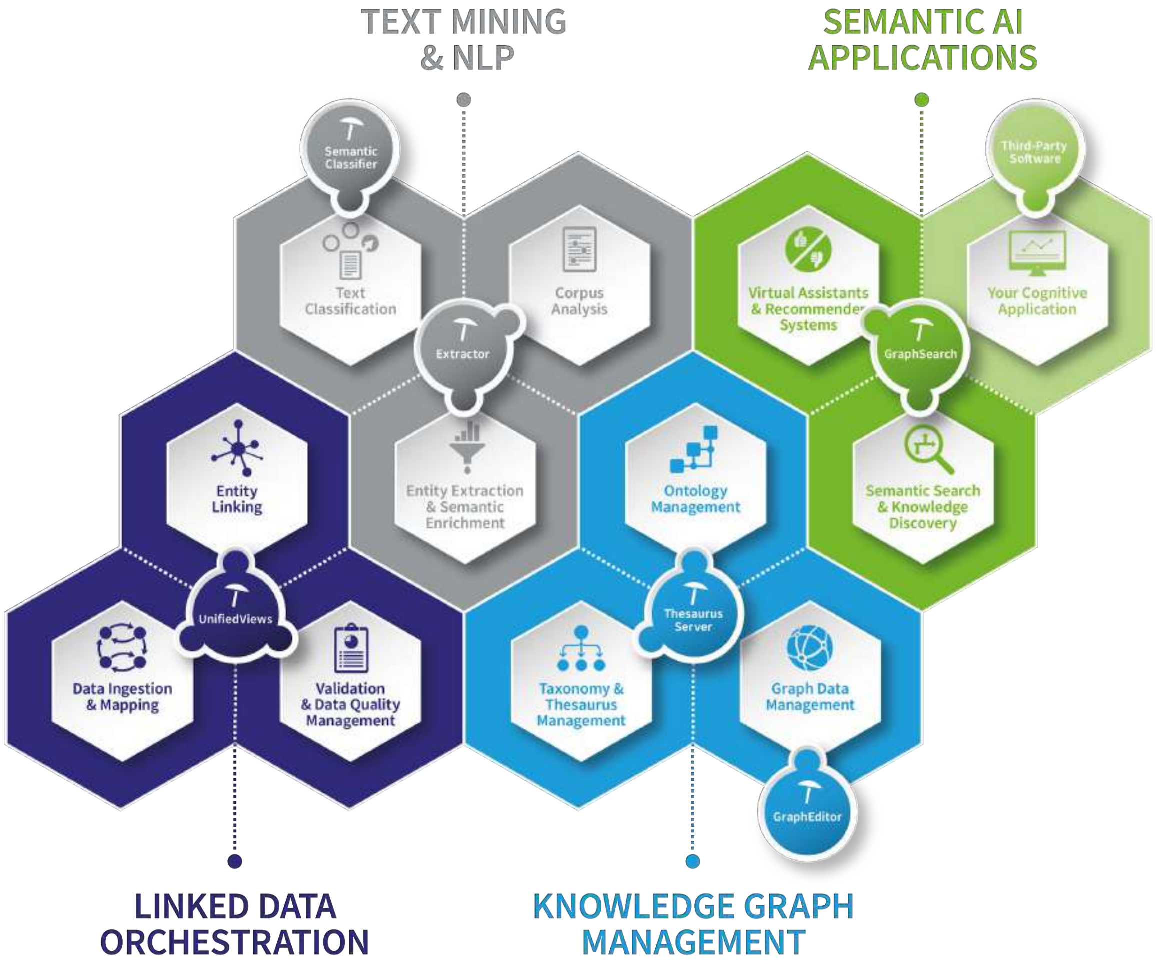
LINKED DATA ORCHESTRATION

KNOWLEDGE GRAPH MANAGEMENT

KMWorld Trend-Setting PRODUCT OF 2017

5

Functions and Components


TEXT MINING & NLP
SEMANTIC AI APPLICATIONS
Semantic Classifier
Text Classification
Corpus Analysis
Extractor
Entity Extraction & Semantic Enrichment
Virtual Assistants & Recommender Systems
Third-Party Software
Your Cognitive Application
GraphSearch
Semantic Search & Knowledge Discovery
Entity Linking
UnifiedViews
Data Ingestion & Mapping
Validation & Data Quality Management
Ontology Management
Thesaurus Server
Taxonomy & Thesaurus Management
Graph Data Management
GraphEditor
LINKED DATA ORCHESTRATION
KNOWLEDGE GRAPH MANAGEMENT

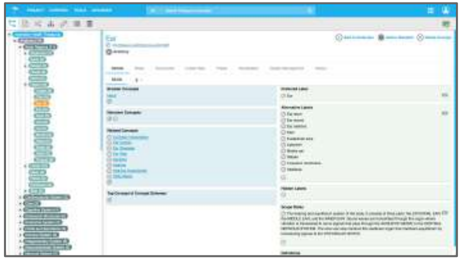
6

PoolParty Components

Screenshots

PoolParty GraphEditor

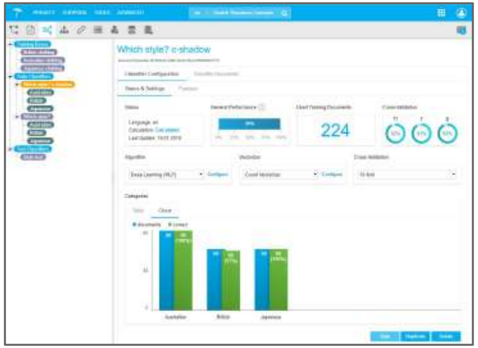
PoolParty Thesaurus Server



PoolParty UnifiedViews

PoolParty Extractor

PoolParty Semantic Classifier



PoolParty GraphSearch

7

Data Consistency

Motivation

Why do we need data consistency?

Software components:

- Stability for application logic
- Correctness of processed results

Users:

- Correctness of analysis results
- Quality of data

8

Data Consistency

Software Components and the Linked Data Life Cycle

- Software components support the Linked Data Lifecycle
- Managed data has to conform to requirements of software components
- Components need input / output validation for data
- Ensure stability for software components
- Correctness of processed results

9

Data Validation Standards

Validation for the Linked Data Lifecycle

RDF based validation approaches:

- SPARQL
- Closed World OWL
- ShEx
- SHACL

10

SHACL

Shapes Constraint Language

“a language for validating RDF graphs against a set of conditions”

- RDF based
- Easy to understand by humans
- Can be processed by machines
- Well defined semantics
- Extendible
- W3C Recommendation

11

SHACL

Shapes Constraint Language

How does it work?

- Define shapes using RDF
- Shapes describe how the data should look like
- Shapes define constraints
- A SHACL processor validates the data against the shapes
- The validation result is a report listing constraint violations

12

Validation Use Cases

Validating data for PoolParty components

13

Use Case 1

SKOS Thesaurus Management

Component: PoolParty Thesaurus Server

- SKOS based data model
- Users can import RDF into project
- The components has requirements:
  - SKOS
  - Additional component-specific constraints
- Data has to be validated on import
- Data can be repaired for conformance

14

Use Case 2

Graph Data Validation

Component: PoolParty GraphEditor

- Ontology based data model
- Ontology driven UI
- Users can connect to graphs
- Users can work freely with RDF data
- Not restricted to SKOS
- But also less stability for data
- Flexible data validation is needed
- Define checks for different use cases

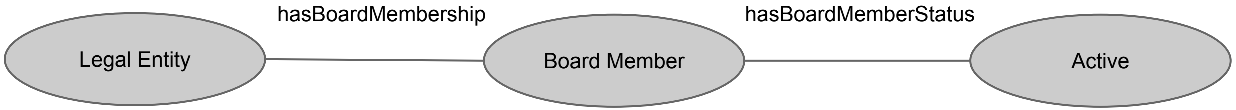
15

Use Case 3

GraphEditor
Validation

Legal Data
Legal Definitions

Component: PoolParty GraphEditor



Constraint:

Not more than two active board members for each Legal Entity

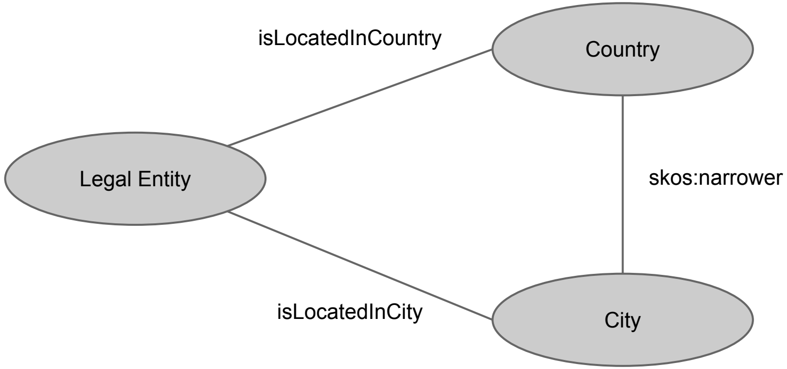
16

Use Case 4

GraphEditor
Validation

Legal Data
Geo Consistency

Component: PoolParty GraphEditor



Constraint:

If a Legal Entity has a country and a city assigned,
then both must be related with a skos:narrower path

17

Use Case 5

GraphEditor Validation

Component: PoolParty UnifiedViews

- Linked data orchestration tool for RDF data processing
- Users define pipelines based on the plugins (DPUs)
- Users transform, merge different formats XLS, CSV, XML creating “free-form” RDF
- Data validation using SPARQL and ASK queries
- More structured and focused data validation is needed

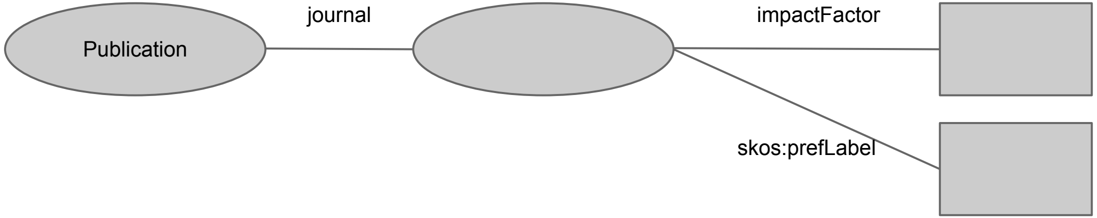
18

Use Case 5

PoolParty UV Validation

Publication dataset

Component: PoolParty UnifiedViews/GraphEditor



journal ⇒ impactFactor ⇔ ¬journal ∨ impactFactor

Constraint:

If a publication has a relation to a journal, that journal should have an impactFactor and a skos:prefLabel

Use Case 5

Shape with logical operators vs SPARQL

```
:PublicationShape a sh:NodeShape ;
  sh:targetClass :Publication ;
   sh:property [
     sh:path  sweb:journal ;
     sh:sparql [
            a sh:SPARQLConstraint ;
            sh:select
            """SELECT $this    WHERE {
                        $this $PATH ?journal;
                       FILTER NOT EXISTS {
                       ?journal :impactFactor ?impactFactor .
                       ?journal skos:prefLabel  ?label . }
                     }""" ;
                  ] ;
            ] .
```

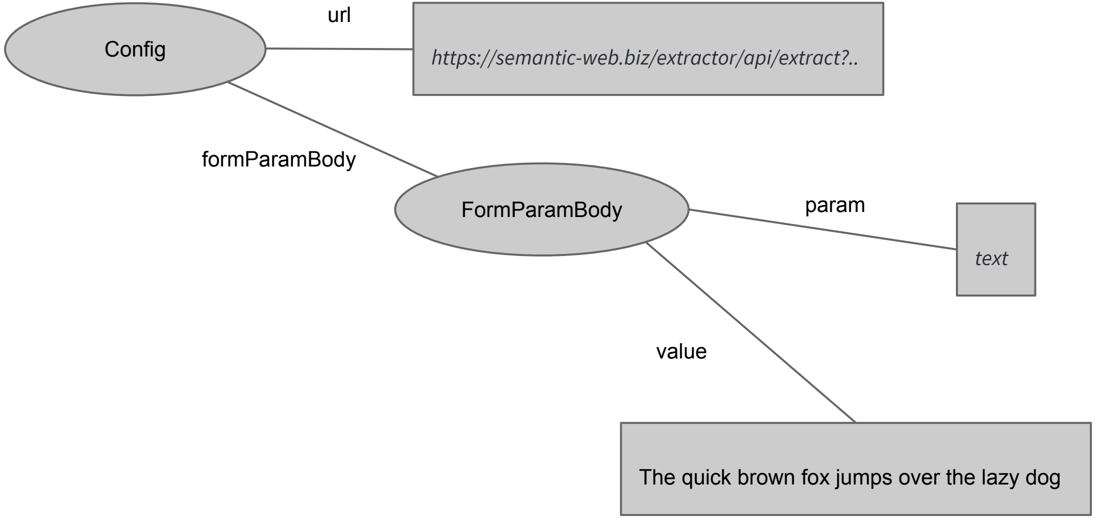
20

Use Case 6

PoolParty UV Validation

Dynamic DPU configuration

Component: PoolParty UnifiedViews/GraphEditor


Config
url
https://semantic-web.biz/extractor/api/extract?..
formParamBody
FormParamBody
param
text
value
The quick brown fox jumps over the lazy dog

Constraint:

HttpRequest DPU configuration should have an url and a body that has a parameter and a value

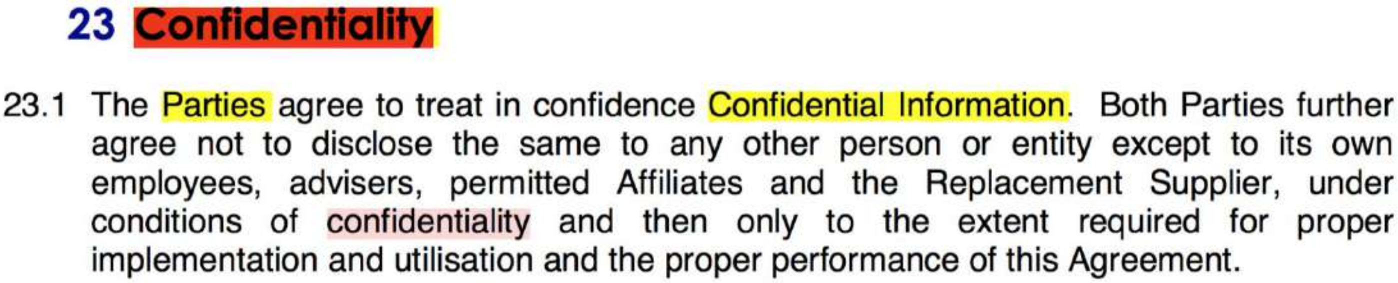
Use Case 7

UnifiedViews/GraphEditor Validation

Rule satisfiability checking

Component: PoolParty UnifiedViews/GraphEditor

- Various rules represented by an *exact set of concepts*
- Concepts should be in a *certain window* of characters subject to a rule



23 Confidentiality

23.1 The Parties agree to treat in confidence Confidential Information. Both Parties further agree not to disclose the same to any other person or entity except to its own employees, advisers, permitted Affiliates and the Replacement Supplier, under conditions of confidentiality and then only to the extent required for proper implementation and utilisation and the proper performance of this Agreement.

Use Case 7

SPARQL plus
SHACL combined

```
sweb:ConfidentialityShape a sh:NodeShape ;
  sh:targetClass :Tag;
  sh:sparql [
    a sh:SPARQLConstraint ;
    sh:select
        """SELECT $this WHERE {
        $this :hasIndex/schema:beginningIndex> ?begIndex1 .
        $this :uri <https://semantic-web.com/Rules/9> .
        BIND (STRDT(?begIndex1, xsd:integer) as ?intBegIndex1)
        FILTER EXISTS
        {   ?s2 :hasIndex/schema:beginningIndex ?index2;
          :uri <https://semantic-web.com/Rules/8>.
          BIND (STRDT(?index2, xsd:integer) as ?intBegIndex2) .
          filter (abs(?intBegIndex1 - ?intBegIndex2) < 100)  }
...
```

23

CONNECT

Robert David

CTO, Semantic Web Company

- robert.david@semantic-web.com
- https://www.linkedin.com/in/robert-david-39b47692/
- https://twitter.com/semwebcompany
- https://blog.semantic-web.at/