Taxonomies and Text Analytics for Recommendation Systems

KM World Connect: Text Analytics Forum
November 18, 2021

Heather Hedden
Data & Knowledge Engineer
Semantic Web Company

About the Speaker

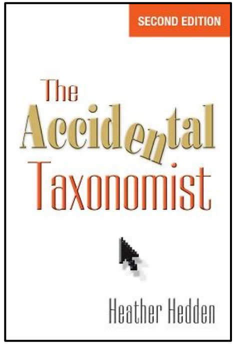
Heather Hedden
Data and Knowledge Engineer
Semantic Web Company

Over 25 years of experience in developing and managing taxonomies, metadata, and other knowledge organization systems for various organizations and applications.

Prior taxonomy consultant and staff taxonomist.

Instructor of self-paced online taxonomy courses.

Author of the book *The Accidental Taxonomist*.



Semantic Web Company (SWC) and PoolParty

poolparty®

SWC is developer / vendor of **PoolParty Semantic Suite**

Most complete and secure **Semantic Middleware / Semantic AI platform** on the Global Market

W3C standards compliant

ISO

ISO 27001:2013 certified

First release in 2009

Current version **8.0**

On-premises or **cloud-based**

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

Semantic AI: Fusion of Graphs, NLP, and Machine Learning

Named as Visionary in **Gartner's Magic Quadrant** for Metadata Management Systems 2019, 2020

KMWorld listed PoolParty as one of the **Trend-Setting Products** 2015 - 2020 and listed SWC in the **AI 50** list of companies in 2020

Outline

- Why Recommendation Systems and Types
- HR Recommender Example Demo
- How a Semantic Recommendation System is Built: HR Recommender Example
  - Taxonomy and Ontology Development
  - Text Mining
  - Knowledge Graph and Search Application

Why Recommendation Systems

Getting the right information to the right people

- There is a lot of information and content people can benefit from;
  they don’t know how best to look for information that would benefit them.
- They don't know that the information is there or how to find it.

Making matches of what goes together

- Standard search does not support complex matching queries.

A system that provides **suggestions** or **recommendations** to users can be very helpful.

Why Recommendation Systems

A recommender system (engine) can recommend to its users:

- content of interest
- products to purchase
- people to connect with
- job opportunities
- training to improve skills
- knowledge assets to reuse

A match-making kind of recommender system can recommend:

- matches of applicants to job openings
- matches of consultants to projects
- matches of buyers and sellers

Recommender System Types

Recommender Technologies

1. **Content-based filtering** - Similar content recommended based on a single user's interactions
   - Can only make recommendations on previous interactions or feedback of the user
2. **Collaborative filtering** - Recommendations based on interactions from multiple similar users
   - Requires a large number of users
3. **Support Vector Machines (SVM)** - Machine learning classification method, using algorithms, training examples, statistical learning, which calculates distances between categories.
   - Often used in combination with **collaborative filtering**
4. **Knowledge-based systems** - Based on explicit knowledge of the content, stored in a graph database, making use of a knowledge graph

Recommender System Types

Disadvantages to both content-based and collaborative filtering

- New users or items, which had not been trained upon, don't get recommendations initially: "cold start" problem due to insufficient data.
- By recommending more of the same, new ideas are lacking; it becomes an echo chamber
- By recommending more of the same, system does not "understand" what makes a good recommendation.
- The choice made by the algorithms are not apparent.
- Can only recommend to the user and not do other matchmaking.

Disadvantages of Support Vector Machines (SVM)

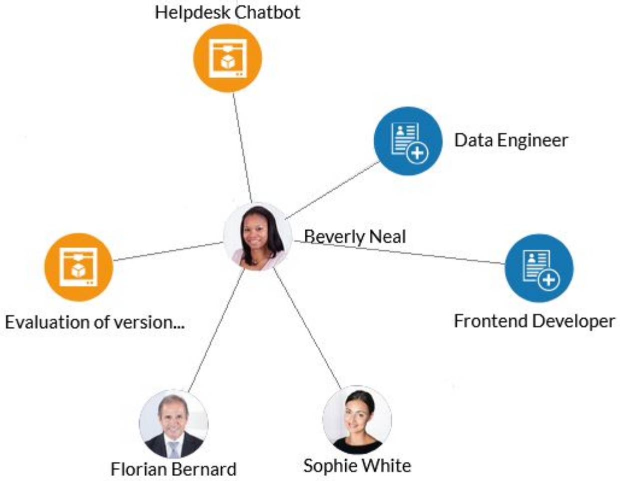
- Requires time to train data, and performance varies based on the data.
- Designed for limited, distinct content and categories;
  doesn't have the benefit a taxonomy with synonyms and semantic relationships

HR Recommender Example

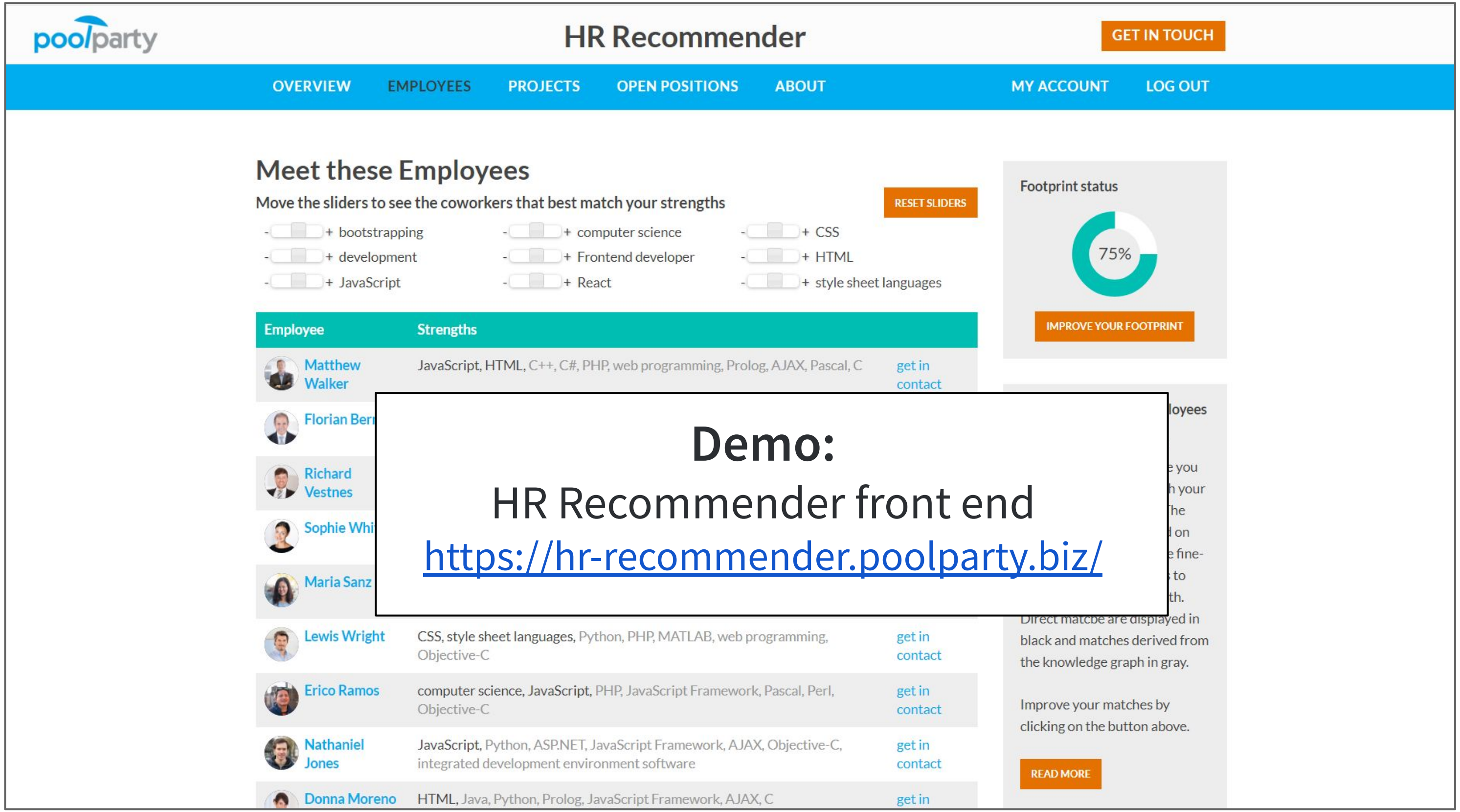
A semantic recommendation/matchmaking tool based on a knowledge graph

Use case

- An organization wants to make the best use of the strengths and skills of its employees.
- Employees, as self-service users, should be able to:
  - Connect with interesting coworkers
  - Browse relevant projects
  - Find career opportunities within the organization
- Matchmaking HR staff should be able to:
  - Find candidates for open positions
  - Staff projects
  - Identify professional development needs



HR Recommender Example



Demo:

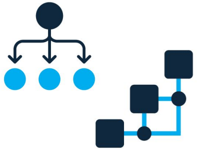
HR Recommender front end

https://hr-recommender.poolparty.biz/

How the Recommendation System was Built

HR Recommender Components

1. **Semantic model**
   - Taxonomies containing concepts and labels
   - Ontology of semantic relations



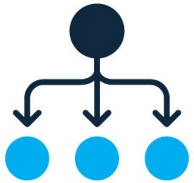
2. **Content that is text-mined**
   - CVs, personal profiles, job descriptions, project descriptions

3. **Stored data**
   - Knowledge graph and a Solr search index

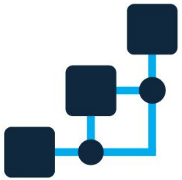
4. **Recommender application**
   - Algorithms for calculating similarities and recommendations to *enrich* the semantic footprint (using a SPARQL endpoint)
   - Web application user interface on top of an API

Semantic Model

Taxonomy & Ontology for the HR Recommender



- **Taxonomy created from multiple sources**
  - Fully developed taxonomies
    - ESCO (https://ec.europa.eu/esco)
    - SEMWEB custom created taxonomy



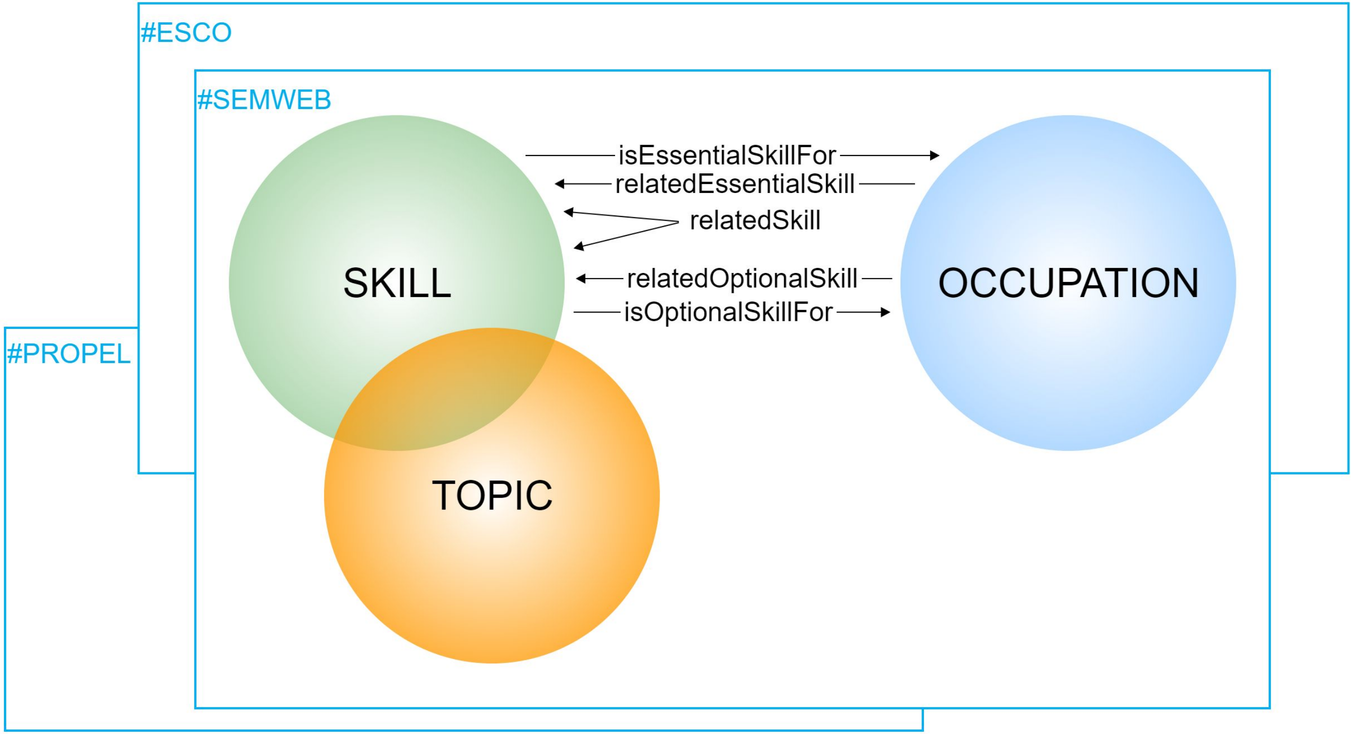
  - Enrich the taxonomy with text mining (entity extraction)
    - Propel
      - Industry conference content: submitted papers, speakers
      - Fictitious CVs
- **Ontology model to add semantic relationships**

Semantic Model

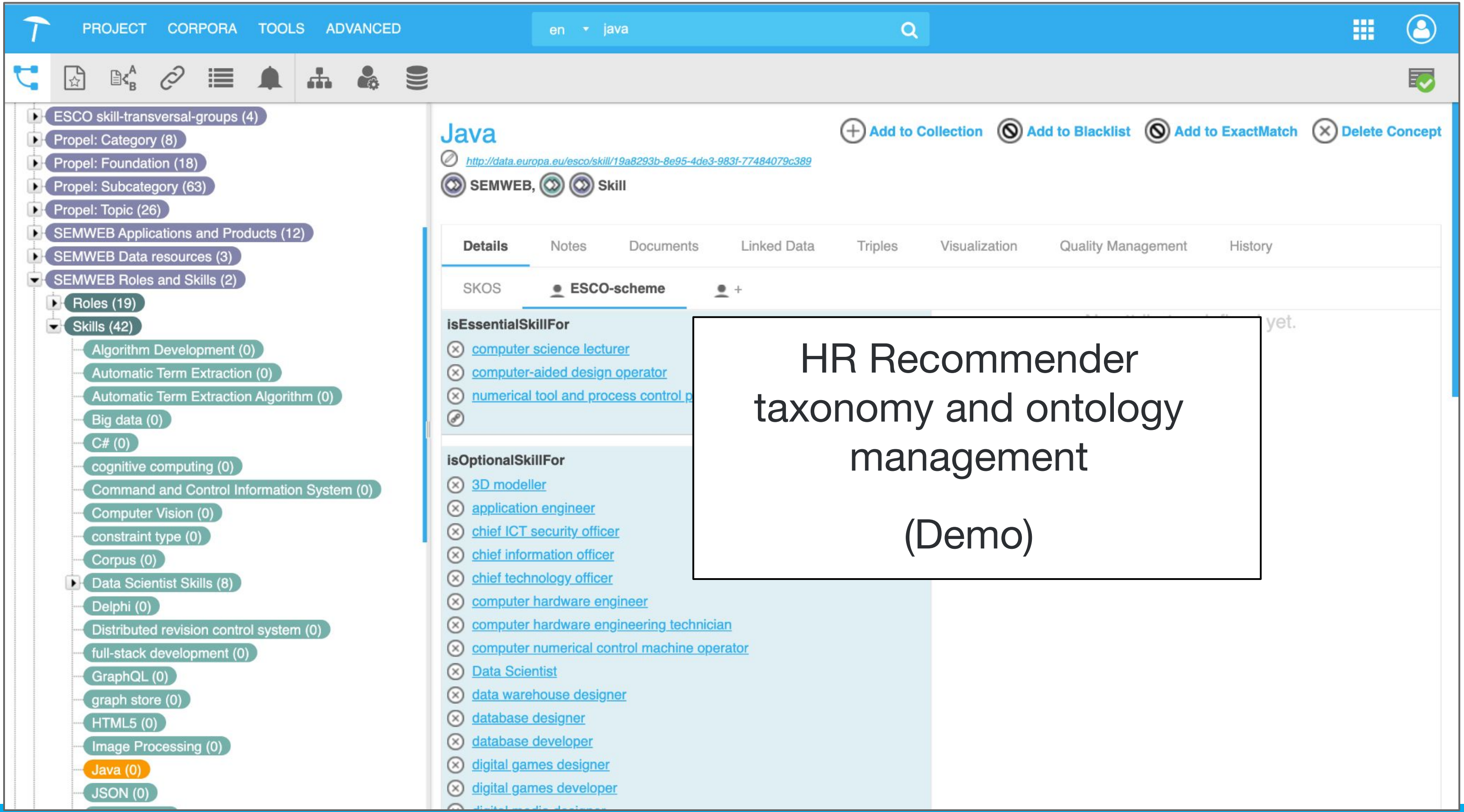
Taxonomy sources:

- Skills & Occupations Topics:
  SEMWEB custom taxonomy
- Skills & Occupations:
  ESCO Classification
- Taxonomy enriched with text mining (term extraction) of Topics:
  Propel corpus of industry conference content: submitted papers, speakers

Ontology model (as a layer):
Adds semantic relationships



Semantic Model



HR Recommender taxonomy and ontology management

(Demo)

Text Mining

What is text mining?

- An application of text analytics, utilizing AI technologies of Natural Language Processing (NLP).
- Extracting passages from text that are relevant in a particular business context.
- Automatically deriving information, and not merely strings of words.
- Transforming unstructured text into meaningful information.

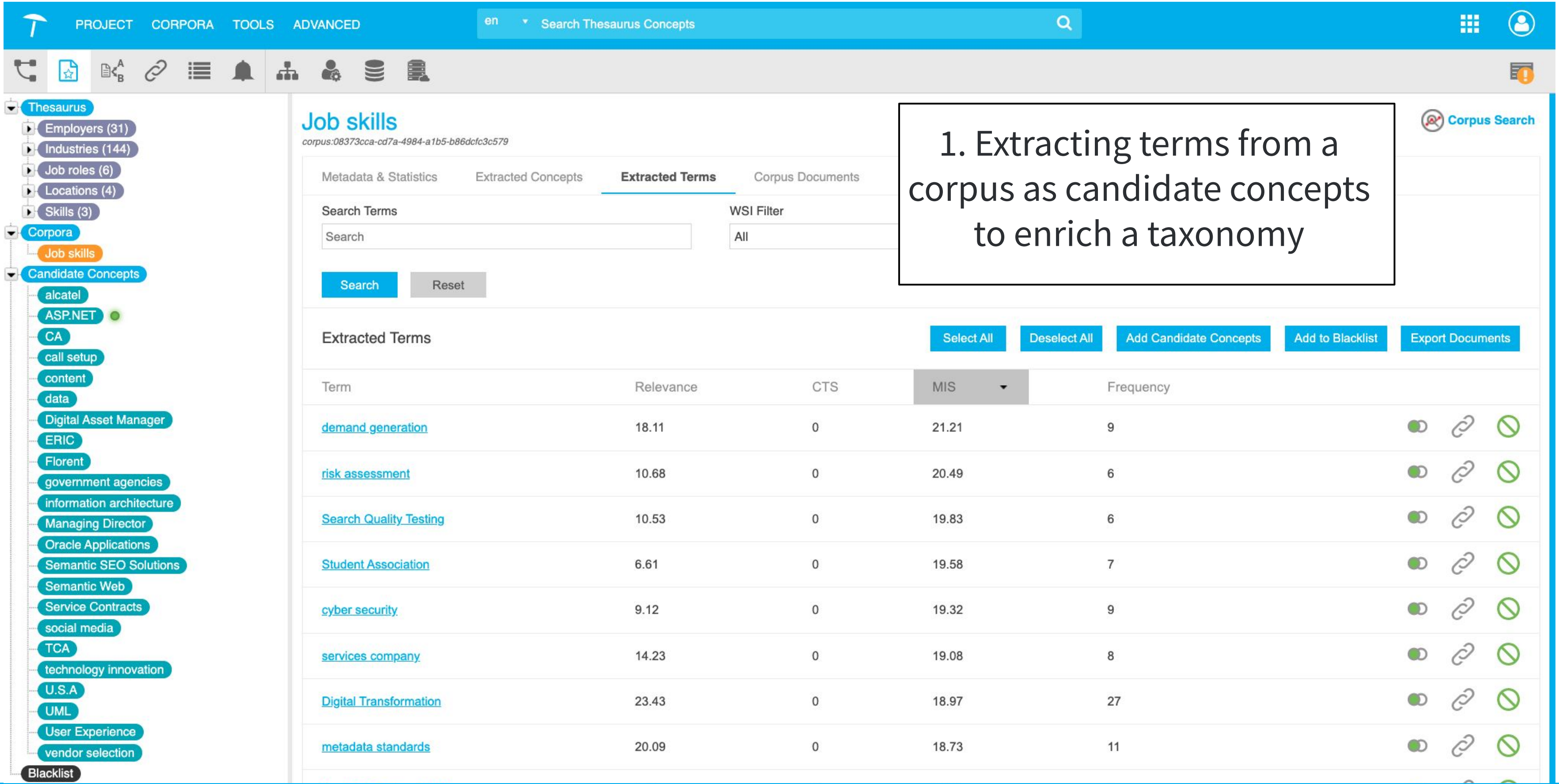
Text mining functions:

1. Extracting terms from a corpus as candidate concepts to enrich a taxonomy
2. Extracting taxonomy concepts from content for auto-tagging it

For the HR Recommender:

1. Extracted terms from the Propel corpus of conference content to enrich the taxonomy
2. Auto-tagged documents of profiles, CVs, projects, and job openings with the taxonomy

Text Mining



1. Extracting terms from a corpus as candidate concepts to enrich a taxonomy

Text Mining

PoolParty Extractor - 7.0.5

Display Debug Information

Concepts

Concept Preferred Label	Normalized Score	Corpora Score	Transitive Broaders	Transitive Broader Top Concepts	Related Concepts
Data Scientist	100.0				
requirements	69.0				
pipelining					
modeling					
dataset					
AWS					
computer science					
infrastructure					
design					
candidate					

2. Auto-tagging with Extractor API
https://hr-recommender-poolparty.poolparty.biz/extractor/test/extraction
(Demo)

Shadow Concepts

Concept Preferred Label	Normalized Score	Corpora Score	Transitive Broaders	Transitive Broader Top Concepts	Related Concepts
interaction	100.0	11073.25			?
Graph	92.0	10287.55			?
data extraction	85.0	9430.44			?
communication	83.0	9257.79			?
JSON	47.0	5236.87			?
remote	42.0	4730.15			?
troubleshoot	35.0	3927.06			?

Stored Data in a Knowledge Graph

poolparty®

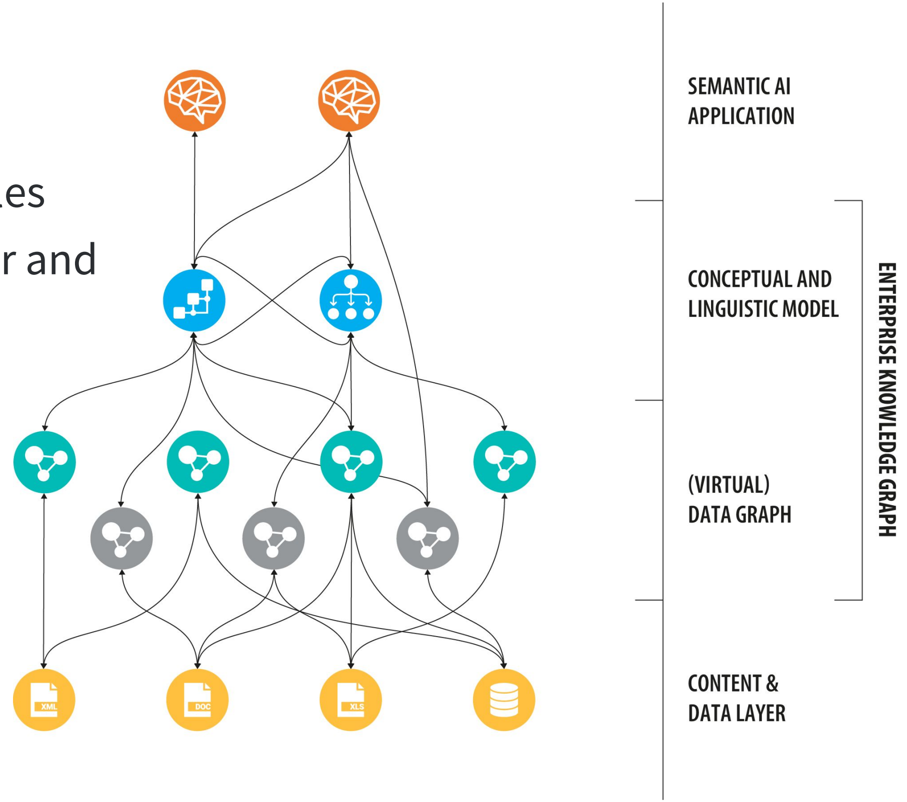
What is a knowledge graph?

- Taxonomy + Ontology + Instance Data stored in a graph database, often as triples
- Connects the content/external data layer and the semantic application layer

In the HR Recommender:

The semantic application is based on the Solr search Index.

Instance data are text snippets about each employee.



Application Build: Enrich the Footprint

SPARQL query endpoint

Algorithms for calculating similarities and recommendations to *enrich* the semantic footprint

SPARQL Endpoint

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX propel: <https://pp-semantics-dev.semantic-web.at/PROPELontology#>
PREFIX esco:<http://data.europa.eu/esco/model#>
SELECT *
WHERE {
 ?uri skos:prefLabel ?label .
 {
  SELECT ?uri (MAX(?distScore) AS ?maxDistScore)
  WHERE {
   VALUES ?x { <http://data.europa.eu/esco/skill/19a8293b-8e95-4de3-983f-77484079c389> }
   {
    BIND(?x AS ?uri)
    BIND(STRDT('1.00',xsd:float) AS ?distScore)
   } UNION {
    ?x esco:isEssentialSkillFor ?uri.
    BIND(STRDT('0.5',xsd:float) AS ?distScore)
   } UNION {
```

Add Namespace

- [ ] SKOS
- [ ] DC
- [ ] DCterms
- [ ] OWL
- [ ] RDF
- [ ] RDFS
- [ ] SWC

Run Query

uri	label	maxDistScore
http://data.europa.eu/esco/skill/19a8293b-8e95-4de3-983f-77484079c389	"Java"@en	"1.00"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/eb0e5615-1575-4a86-a1a2-7d39595033c5	"ABAP"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/b4dc6e4f-dc7d-445f-8ce2-d7b9d225e282	"AJAX"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/58d7a289-dafd-4363-833f-d1dc4140885e	"APL"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/56a7f561-1d55-43c9-9cd7-36a0a9bc6c50	"ASP.NET"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
https://pp-semantics-dev.semantic-web.at/PeopleandContentMatchmaker/820a0683-d2ae-4a88-a648-3feb2f104e44	"Angular"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
https://pp-semantics-dev.semantic-web.at/PeopleandContentMatchmaker/3bfcfbe2-4f52-4b55-99d3-cb76a7e8131e	"AngularJS"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
https://pp-semantics-dev.semantic-web.at/PeopleandContentMatchmaker/f31ae51f-aff9-42b1-81e9-fc9a55302090	"Apex"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/47b9bbcf-356c-4782-83a4-7f5a1b2b51a3	"Assembly"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
https://pp-semantics-dev.semantic-web.at/PeopleandContentMatchmaker/3014fd67-33b2-4992-bffc-042338cdb026	"C"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/4c016b68-4116-468c-9dc6-42710c239e4a	"C#"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/b633eb55-8f1f-4ae6-ab4c-2022ffe2cb7f	"C++"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/def007fa-5fed-4a5f-91a2-b0d7e3db1be1	"COBOL"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/993b1e23-f2de-4bd8-b33f-f86dde1c8e9d	"CoffeeScript"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
http://data.europa.eu/esco/skill/0cd6dcf1-5778-42a5-b685-4d01ae4a4871	"Common Lisp"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
https://pp-semantics-dev.semantic-web.at/PeopleandContentMatchmaker/661ac55d-3e7f-4cd8-bad3-4dc4af6efab0	"Crystal"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>
https://pp-semantics-dev.semantic-web.at/PeopleandContentMatchmaker/8e7dabd3-bcd8-4309-adb0-dac16ee331db	"Delphi"@en	"0.7"^^<http://www.w3.org/2001/XMLSchema#float>

Conclusions

poolparty

Semantic recommender systems are based on:

- A knowledge graph comprising:
  1. A taxonomy, whose concepts are tagged to and/or extracted from the content to be recommended *and* to either matchable content or a user profile
  2. An ontology that links concepts with additional semantic relationships
  3. Instance data linked to the taxonomy/ontology stored in a search index or graph DB
- A large body of content tagged with the taxonomy

Optionally enhanced with:

- Algorithms for weighting/scoring relations

And:

- A front-end (user interface) application

Resources

- “From Taxonomies to Recommendation Systems” webinar recording
  www.poolparty.biz/events/from-taxonomies-to-recommendation-systems
- Recommendation/matchmaking demos
  - HR Recommender
    https://hr-recommender.poolparty.biz
  - Wine & Cheese Harmonizer
    http://vocabulary.semantic-web.at/GraphSearch
  - Semantic Matchmaker (Matching consultants to projects)
    https://semantic-matchmaker.poolparty.biz
- “Natural Language Processing with PoolParty” white paper
  www.poolparty.biz/resources/natural-language-processing-with-poolparty

Questions/Contact

Heather Hedden
Data and Knowledge Engineer
Semantic Web Company Inc.
One Boston Place, Suite 2600
Boston, MA 02108

857-400-0183
heather.hedden@semantic-web.com
www.linkedin.com/in/hedden

Semantic Web Company www.semantic-web.com

PoolParty Semantic Suite www.poolparty.biz