Attendees:

- Georg Ferdinand Schneider (Individual CLA but affiliated with Schaeffler)
- John Maiden (Cherre)
- Jyrki Oraskari (RWTH-Aachen, BIM4Ren)
- Joel Bender (Cornell University)
- Serge Chávez Feria (Ontology Engineering Group - Universidad Politécnica de Madrid)
- María Poveda-Villalón (Ontology Engineering Group - Universidad Politécnica de Madrid)
- Sjoerd Rongen (Taxonic)
- Mads Holten Rasmussen (NIRAS)
- Erik Wallin (Idun Real Estate Solutions / RealEstateCore)
- Bart Hanssens
- Manuel Lopez-Enriquez (Schaeffler)
- Katja Breitenfelder (Fraunhofer IBP / Technical University of Munich)
- Jereon Werboruck
- Gonçal Costa (LaSalle University)
- Calin Boje (LIST)

Date and time

- 20/10/2020
- 15:00-16:30@UTC, 17:00-18:30@CEST, 16:00-17:30@BST, 08:00-09:30@PDT, 23:00-00:30@CST
- Connection details: https://lists.w3.org/Archives/Member/internal-lbd/2020Oct/0000.html

Agenda (tentative)

1. Introduction (5min)
2. 'Building A Knowledge Graph Of Commercial Real Estate At Cherre' by John Maiden (45 min)
3. Questions and Discussion (30 min)
4. Open discussion and Agenda Items

Minutes

1. Introduction (5min)
    - https://www.w3.org/community/lbd/
    - https://github.com/w3c-lbd-cg/lbd/blob/gh-pages/presentations/out/TPAC2020.pdf
    - w3id.org/bot
    - https://www.w3.org/community/lbd/2020/09/04/w3c-lbd-cg-meets-selected-topics-on-linked-data-semantic-web-and-graph-technologies-in-the-built-environment/

2. 'Building A Knowledge Graph Of Commercial Real Estate At Cherre' by John Maiden (45 min)

Building-up a CRE Knowledge Graph

- Short presentation of the US/ New York based company “Cherre”
- Background/ goal: using commercial real estate (CRE) data to transfer questions like:
    - Property address, owner (tax assessor..), actual owner, couple of large providers;
    - The property’s true owner: Contact info, portfolio holdings etc.;
    - Follow-up questions, eg. “Which properties has this owner bought and sold in the past five years?”;
    - Decision Support: Property Comps, Valuation Models;
    - Making public data “transparent”.
- Once the data is collected it is possible to answer certain questions on the properties, eg. on environmental questions, on stakeholders involved, on data collected over a certain time period..
- Difference between data collection on residential real estate and on commercial real estate, the latter being more open for interpretation.
- Building-up the CRE Knowledge Graph is followed by multiple opportunities for data usage eg. deep learning models, questioning on assessed taxes, stakeholder contact information, owner information etc.
- Data sources:
    - Huge efforts are made into building-up the “NYC public real estate data” (open data initiative) sourced national data and supplemented from: “NYC Department of Finance”, “NYC Buildings” (collection of building data).
- General data requirements
    - (1) on Buildings: tax payer, transaction history, permits..;
    - (2) on Corporate Data: corporate structure (locations..), contact info, LLC registration.
- Building-up the Knowledge Graph - first step defining the taxonomy:
    - (1) Nodes: Property, Address, Person, Corporation (registered vs. unknown..);
    - (2) Edges: Sources / Data Related Attributes, Recency / Frequency (collected data can overspan decades).
- Knowledge Graph Mining & Data Analysis:
    - Scanning for certain “Motifs”: Identify relevant data patterns, assigning confidence levels to sources, edges and patterns;
    - Everything is done on “Scale”: US has around 150m registered tax lots ..posing challenges in Engineering and Analysis;
    - Further tasks: identifying missing information and new data sources, securing consistency of collected data for customers;
    - Implementing SQL graph technology.
- Securing clean and consistent data:

    - Huge time effort necessary for cleaning-up the data before building-up the knowledge graph;
    - Workflow: Extract the data sources > cleaning names/addresses > creating “standardized names/ addresses” > Building-up the Knowledge Graph;
    - Most data is collected by data partnerships, only little use of online sources.
- People / Corporation Standardization:
    - Other standardization work is necessary to create valid data (sources)
    - Cleaning the names and categorization are important
    - Solutions to solve the problem: Regex, NLP-based classification models, Graph & Fuzzy Matching, good reference data.
- Address Standardization:
    - Abbreviations/ alternate names, spelling variations, obvious typos / sticky components, embedded addresses;
    - Implemented technology: Parsing (Regex, hidden Makov Models, Conditional Random Fields, Neural network); Good Address Data (mailing-/property-/ landmark-/ vanity addresses).
- Lessons learned from Standardization:
    - Business knowledge / context is critical;
    - Dealing with scale is important to standardization;
    - Focus on continuous improvement.

“Placekey” - the company’s new initiative

- “Placekey” is an universal Identifier for physical places: https://www.placekey.io/
- Technical aspects: making use of an H3 hierarchical hex-grid system, currently representing 9 different places
- Encoding addresses > understanding where the building (unit) is located is considered a very helpful feature

- Contact details John Maiden: john@cherre.com

3. Questions and Discussion (30 min)
    - Q [Manuel López]: How do you make usage of “Placekey”? -> A: “Placekey” is a compliment national/ local data source providing encoded data.
    - Q [Vladimir Alexiev]: What is your motivation for today’s presentation? Are you looking for cooperation opportunities? -> A[John]: Focus was on information about the technical challenges of machine learning etc. the company is confronted with by building-up the CRE knowledge graph, but there is also interest in cooperation from the company’s perspective. A[Georg]: The presentation showed a industrial large-scale application of knowledge graph technology in the real estate domain, which is fully inscope to the W3C LBD CG

topics and interest. W3C LBD CG chairs and members are happy for the opportunity to listen and gain insights from John´s presentation.
    - Q [Manuel López]: Do you use the CRE knowledge graph representation for analysis on an urban scale? -> A: Yes, for the purpose of urban analysis multiple sources are used for collecting data. General goal: Providing as much data as possible to the customer.
    - Q [Manuel López]: Do you loop back Feedback from customers to your knowledge graph -> A: Yes
    - Q [Vladimir]: Did you also integrate information on crime or similar? Yes
    - from Vladimir Alexiev to everyone: 5:53 PM GS1 GLN, Google's place id (was introduced... maybe 1y ago?)
    - Q [Erik]: Do you provide the ontologies you use in your use case? No
    - from Serge Chavez to everyone: 5:56 PM Q [Serge] How much time of the whole pipeline do you spend cleaning and standardizing the data? A: More than wanted ;-). It is an effort.
    - Q [Georg]: Do you use any open schemas or ontologies? A: Not so much. The output of the extraction pipeline is a graph. The semantic expressivity used is rather limited.
    - Q [Georg]: When refreshing the graph how much effort is related to this and is the graph completely redone or only in parts. A: Extraction and update pipeline is fully automated. Versioning applied. Redoing is not too much of a problem and large parts are created.

4. Open discussion and Agenda Items
    - None

Next Call

- 03/11/2020, Tue, 15:00-16:30@UTC, 17:00-18:30@CEST, 16:00-17:30@BST, 08:00-09:30@PDT, 23:00-00:30@CST

We are interested in getting suggestions from the community about potential agenda items for the following calls. Please send your suggestions to public-lbd@w3.org, whether you have a short presentation to bootstrap the discussion, and an approximate duration you think the discussion will last.

Previous minutes

https://docs.google.com/document/d/1PaEGwLSeyPOT7lijGkIt7s6gN2Mfi1t4P5kcA9sFCBM/edit?usp=sharing