LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

Horizon 2020 Grant Agreement No 809965
Contract start date: July 1st 2018, Duration: 30 months

LAMBDA Deliverable 3.2
Enterprise Knowledge Graphs

Due date of deliverable: 30/06/2019
Actual submission date: 30/06/2019

Revision: Version 1.0

Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme, H2020-WIDESPREAD-2016-2017 Spreading Excellence and Widening Participation under grant agreement No 809965.

Author(s)	Sahar Vahdati (UBO),Emanuel Sallinger (UOXF)
Contributor(s)	Tim Furche, Luigi Bellomarini (UOXF)
Internal Reviewer(s)	Valentina Janev (PUPIN)
Approval Date	
Remarks	

Workpackage	WP 3 Cooperation for Teacher and Student Training
Responsible for WP	Institute for Computer Science - University of Bonn
Deliverable Lead	University of Oxford (Emanuel Sallinger)
Related Tasks	Task 3.2 'Train the Trainer' Lectures (UBO)

Document History and Contributions

Version	Date	Author(s)	Description
0.1	10.06.2019	Sahar Vahdati	First Draft
0.2	17.06.2019	Emanuel Sallinger	Section 2 &Section 3
0.3	18.06.2019	Tim Furche, Luigi Bellomarini	Section 2 &Section 3
0.4	28.06.2019	Valentina Janev	Internal Review
1.0	30.06.2019	Sahar Vahdati	Final version

Institute Mihajlo Pupin (**PUPIN**)	Co-ordinator	Serbia
Fraunhofer Institute for Intelligent Analysis and Information Systems (**Fraunhofer/IAIS**)	Contractor	Germany
Institute for Computer Science - University of Bonn (**UBO**)	Contractor	Germany
Department of Computer Science - University of Oxford (**UOXF**)	Contractor	UK

Executive Summary

This document contains details about learning material developed as one of the specific objectives of LAMBDA (Learning, Applying, Multiplying Big Data Analytics) project. The reporting period is (M1-M12). The project partners developed a series of learning material during this period which is made available for BDA school participants. The learning material will be presented as OpenCourseWare (OCW) on a shared platform. In this deliverable, we report about one of the fundamental topics of the BIG Data hype namely Enterprise Knowledge Graphs. This deliverable will have two further follow up deliverables D3.3 Semantic BD Architectures and D3.4 Smart Data Analytics.

Table of Contents

Abbreviations and Acronyms

BDA	Big Data Analytics
NoE	Network of experts
OCW	OpenCourseWare
PPT	Power Point File format/extension (Microsoft)
WP	Work Package

List of Figures

1. Introduction

Learning is considered as one of the main building blocks for the LAMBDA (**L**earning, **A**pplying, **M**ultiplying **B**ig **D**ata Analytics, http://www.project-lambda.org/) project. Aligned with the overall objectives of the project to stimulate scientific excellence and innovation capacity, initial consideration of the consortium goes in providing high quality and up to date material related to Big Data. Therefore, a set of lectures and online material has been produced for the Belgrade Big Data Analytics Summer School 2019 and 2020. The 1st Summer school was held in the PUPIN premises (part of the University of Belgrade) from the 17th to the 20th of June 2019. The website of the summer school, https://project-lambda.org/Summer-School-2019 , provides more details about the organization and topics discussed at the school, while Deliverable 3.5 briefly documents the event.

1.1 Relation to Other Deliverables

Deliverable 3.2 **Enterprise Knowledge Graphs** is the first of three reports associated with Task 3.2 'Train the Trainer' Lectures. The whole set of LAMBDA lectures contain:

1. **Deliverable 3.2 Enterprise Knowledge Graphs**: lectures that include formal conceptual frameworks for designing and maintaining knowledge graphs; such as strategies for the semi-automatic construction of such graphs from the combination of proprietary enterprise data and relevant public domain knowledge; opportunities and implications in terms of performance and access control.
2. **Deliverable 3.3 Semantic BD Architecture:** lectures that include approaches for better supporting the variety dimension of Big Data comprising RDF, RDF-Schema and OWL knowledge representation formalisms, mapping standards such as R2RML, JSON-LD and CSVW, the SPARQL query language, etc. Integrating semantic and Big Data technologies can help to make Big Data architectures and applications more flexible, adaptive and their implementation more efficient.
3. **Deliverable 3.4 Smart Data Analytics**: lectures that include different algorithms and tools related to Distributed Semantic Analytics, Semantic Question Answering, Structured Machine Learning, Deep Learning, Software Engineering for Data Science, Semantic Data Management, Knowledge Extraction and Validation.

This document corresponds to deliverable 3.2, while the other two deliverables will be developed accordingly in the future. Lectures that were presented at the Big Data Analytics Summer School 2019 refer to the three selected topics **Enterprise Knowledge Graphs (D3.2), Semantic BD Architecture (D3.3) and Smart Data Analytics (D3.4).** The slides have been uploaded to the LAMBDA portal (see Figure 1 or https://project-lambda.org/Knowledge-repository/Lectures) and to the SlideWiki.org OpenCourseWare platform, https://slidewiki.org/deck/128440-1/big-data-analytics-lectures (total 468 slides).

LEARNING, APPLYING, MULTIPLYING BIG DATA ANALYTICS

SUMMER SCHOOL Lectures from the LAMBDA Consortium

Enterprise Knowledge Graphs

University of Oxford	Introduction to Knowledge Graphs
	Extraction for Knowledge Graphs
	Reasoning in Knowledge Graphs

Semantic Big Data Architectures

Fraunhofer Institute / University of Bonn	Introduction to Big Data Architecture
	Distributed Big Data Frameworks

Smart Data Analytics

Fraunhofer Institute / University of Bonn	Distributed Big Data Libraries
	Distributed Semantic Analytics I
	Distributed Semantic Analytics II

Examples

Fraunhofer Institute / University of Bonn	Big Data Solutions in Practical Use-cases
Institute Mihajlo Pupin	Data Analytics for Energy Sector

The LAMBDA Consortium

institute MIHAILO PUPIN

Fraunhofer IAIS

UNIVERSITÄT BONN

UNIVERSITY OF OXFORD

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme, H2020-WIDESPREAD-2016-2017 Spreading Excellence and Widening Participation under grant agreement No 809965.

Figure 1. Lectures from the LAMBDA consortium presented at the Summer school 2019

1.2 Structure of the Deliverable

This deliverable contains the following sections:

1. Definitions of Knowledge Graphs - definitions are given based on keynote and lecture content of BDA. (section 2)
2. Enterprise Knowledge Graphs - an overview of lectures related to the VADALOG system[1,2], developed by the University of Oxford. (Section 3)

2. Definitions of Knowledge Graphs

Although different communities tried to define KGs from different perspectives, there is no unique definition for that. The first definition of knowledge graphs to the BDA school participants was presented by the first keynote speaker, Sören Auer, the director of the German National Library for Science and Technology. The following figure shows the corresponding slide about KG definition which is taken from his presentation[3].

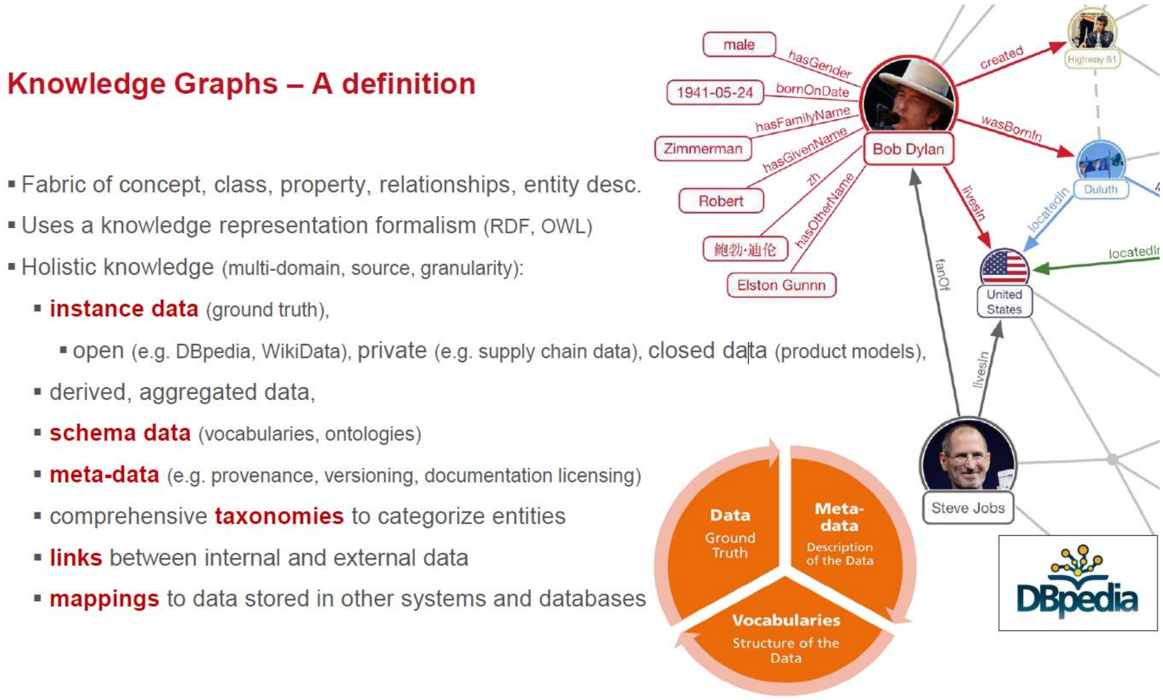


Figure 2. KG Definition in TIB Hannover Presentation

[1] Georg Gottlob, Andreas Pieris, Emanuel Sallinger (2019) Vadalog: Recent Advances and Applications, European Conference on Logics in Artificial Intelligence, Springer, 21-37.
[2] Luigi Bellomarini, Georg Gottlob, Andreas Pieris, Emanuel Sallinger (2018) Vadalog: A Language and System for Knowledge Graphs, International Joint Conference on Rules and Reasoning, Springer, 3-8.
[3] Sören Auer (2019) Towards Knowledge Graph based Representation, Augmentation and Exploration of Scholarly Communications, Keynote at the 1st LAMBDA Big Data Analytics Summer School (18-20.06.2019, Belgrade, Serbia)

Atanas Kiryakov, CEO of OntoText was the second keynote and presented another definition of KG from their perspective[4].

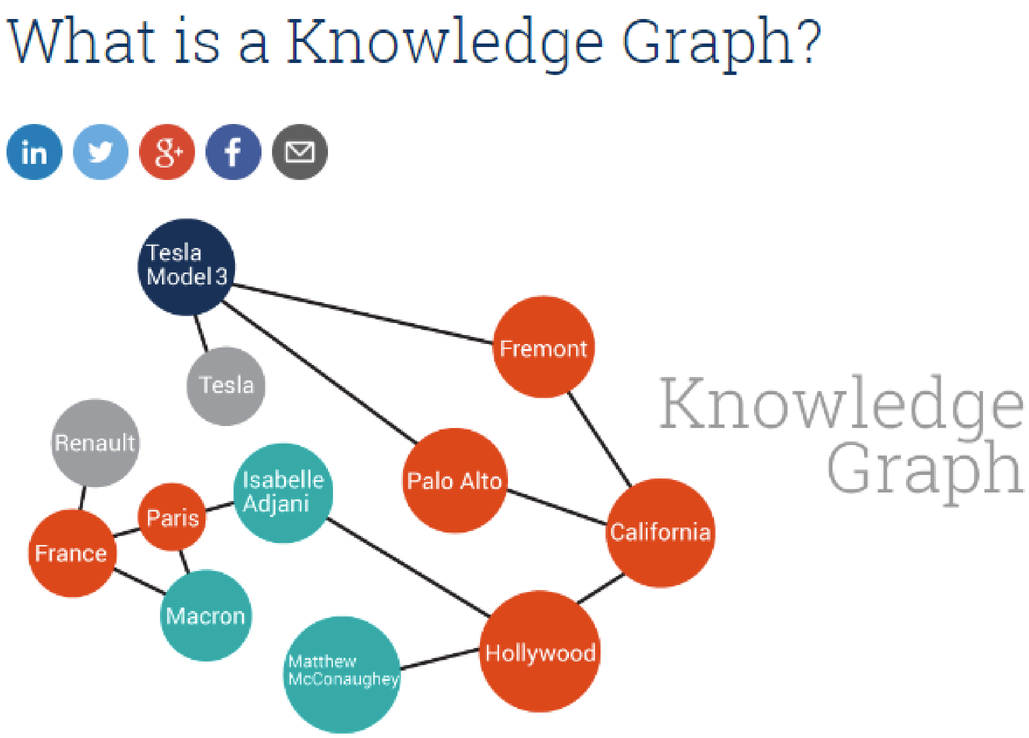


Figure 3. KG Definition from OntoTex Website (2)

[4] Atanas Kiryakov (2019) How Analytics on Big Knowledge Graphs Help Data Linking: Company Importance and Similarity Demo, Keynote at the 1st LAMBDA Big Data Analytics Summer School (18-20.06.2019, Belgrade, Serbia), see also https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

The Knowledge Graph represents a collection of interlinked descriptions of entities – real-world objects, events, situations or abstract concepts – where:

- Descriptions have a formal structure that allows both people and computers to process them in an efficient and unambiguous manner;
- Entity descriptions contribute to one another, forming a network, where each entity represents part of the description of the entities, related to it.

Knowledge Graphs combine characteristics of several data management paradigms. The Knowledge Graph can be seen as a specific type of:

- **Database**, because it can be queried via structured queries;
- **Graph**, because it can be analyzed as any other network data structure;
- **Knowledge base**, because the data in it bears formal semantics, which can be used to interpret the data and infer new facts.

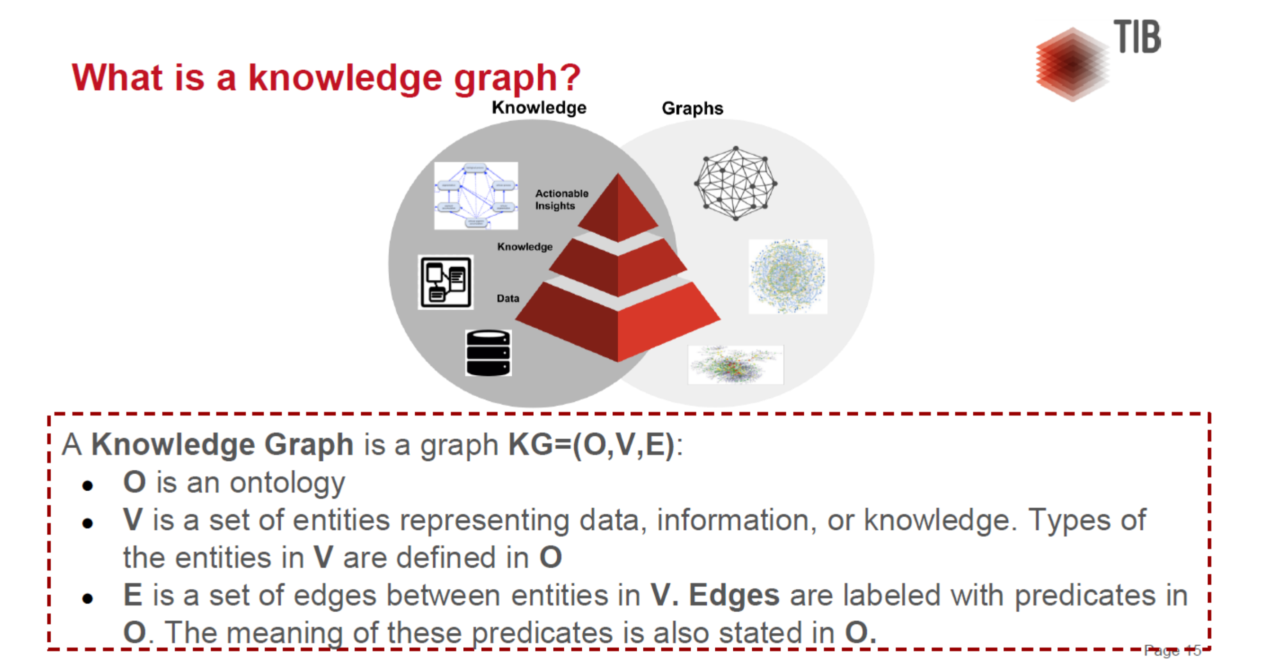


Figure 4. KG Definition in TIB Hannover Presentation

3. Enterprise Knowledge Graphs

The LAMBDA consortium members and representatives from the University of Oxford were:

- Emanuel Sallinger, University of Oxford, TU Wien and DeepReason.ai
- Luigi Bellomarini, Banca d'Italia and University of Oxford
- Tim Furche, Meltwater and University of Oxford

The UOXF lectures about

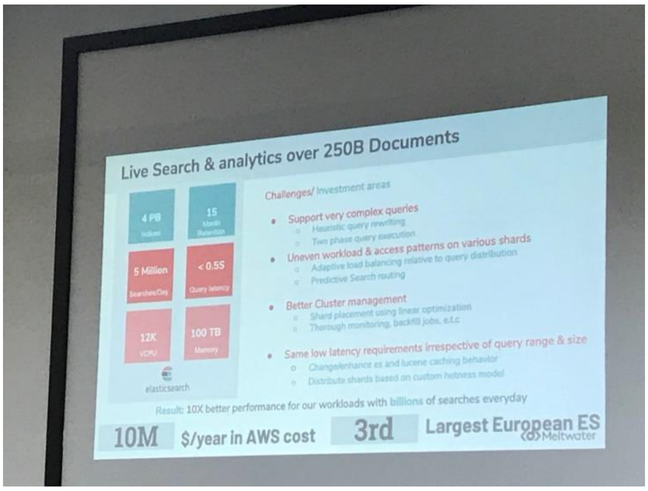
- Knowledge Graphs
- Reasoning in Knowledge Graphs
- Extraction for Knowledge Graphs

are based on the experience gained in the project VADA[5], see research objectives[6] and publications[7], as well as commercial projects conducted by UOXF spin-off companies Wrapidity (acquired by Meltwater) and DeepReason.ai. In addition, the work done at the Banca d'Italia was presented by Luigi Bellomarini, see Figure 5.

University of Oxford - Knowledge Graphs Definition

Reasoning in Knowledge Graphs at Banca d'Italia

[5] http://www.cs.ox.ac.uk/projects/vada/
[6] http://vada.org.uk/research-objectives/
[7] http://vada.org.uk/publications-by-topic/



Live Search and Analytics on Big Data at Meltwater

Meltwater represented by Dr. Tim Furche

Figure 5. Lectures by the University of Oxford

3.1 Knowledge Graphs:

The first lecture, see https://project-lambda.org/EKGs-Lecture-1, includes a comprehensive introduction to Knowledge Graphs (KGs). The start of this lecture was designed to clarify these perspectives and leave the participants with a broad view of KGs and their application in Big Data challenges. The aim was to harness various natures of KGs, which emerge from literature and practice.

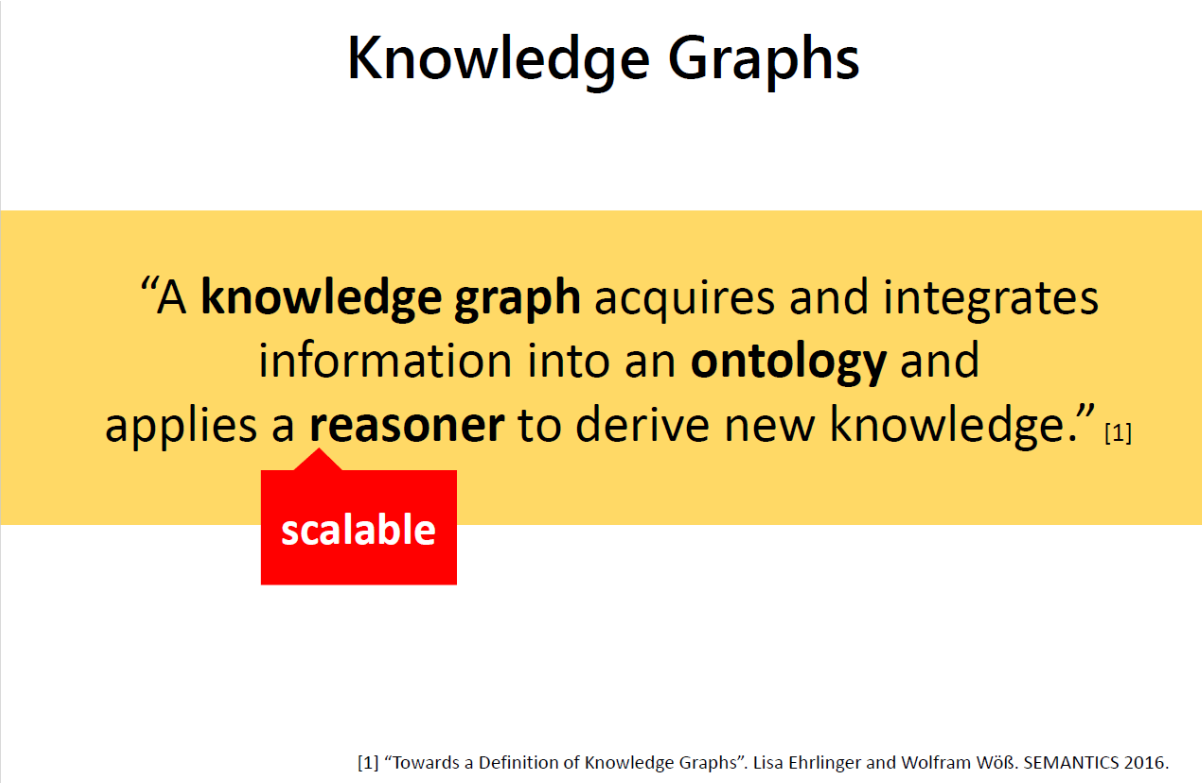


Figure 5. KG Definition in Oxford presentation

The lecture was followed by motivating applications on knowledge representation, interpretation, inference, and reasoning. Considering Big Data applications in modern IT scenarios, a KG-based IT architecture with knowledge graph management systems (KGMS) was introduced.

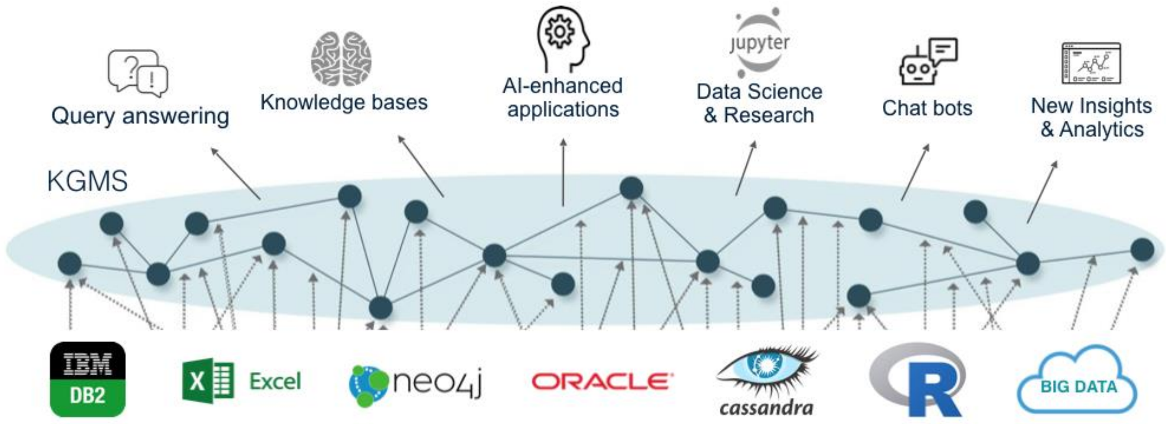


Figure 6. IT architecture based on Enterprise Knowledge Graphs

Requirements for such a management system would be to provide reasoning services via a rich set of APIs. This depends on the knowledge representation language which should be simple and modular. It should also be of high expressive power, so to be suitable to represent problems of real interest applications. Another requirement would be to handle Big data sources properly, in the sense that processing different data formats e.g., relational, RDF. A good yardstick is Datalog, a formalism or query language based on logic programming initiated from databases and logic programming communities from the 80s. Therefore the lecture was followed by an introduction to syntax and semantics of Datalog[8].

3.2 Reasoning in Knowledge Graphs:

Beyond the specific requirements which were motivated in lecture one, in lecture two https://project-lambda.org/EKGs-Lecture-2 the participants have been introduced to the core challenge for an effective KGMS which lies in providing efficient reasoning solutions. The second lecture focuses on reasoning using VADALOG[9] - A Core System which is a KG management system developed by Oxford University is introduced. Motivated by deductive database systems, Datalog has been successfully used in academia and industry for several reasoning tasks. It fulfills the requirements for KGMS handling Big Data. It turned out to be a key factor and many companies and research projects directed their attention to logic-based KGMS.

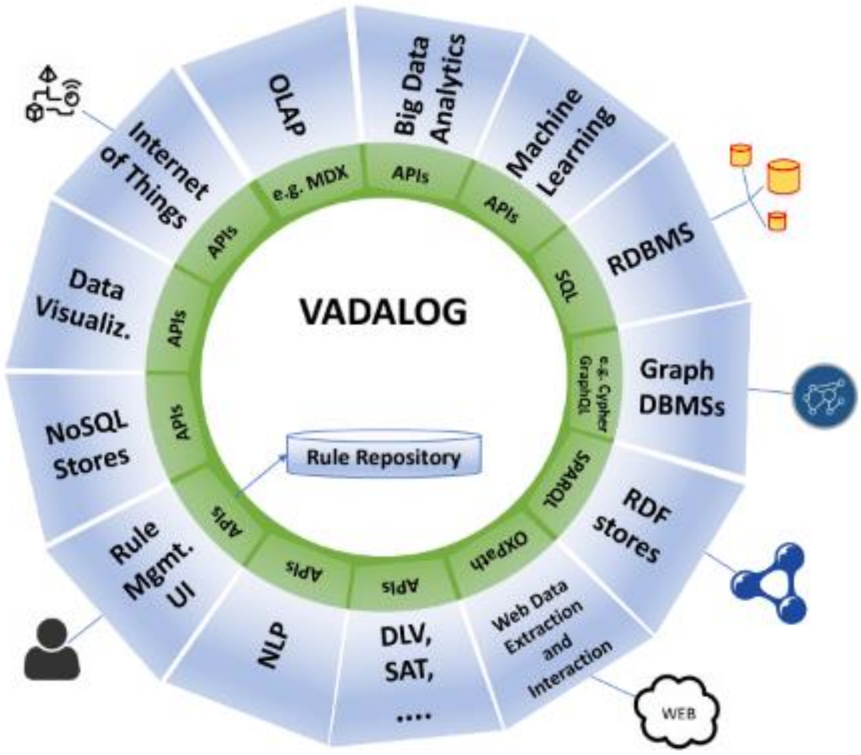


Figure 7. VADALOG System

Reasoning is essential to gain value from Knowledge Graphs by deriving insights and making available new implicit data from existing data. The VADALOG system is using an extended version of Datalog which was introduced with different examples mainly from the commercial domain.

Figue 9 shows an example of a simple database, a query, and results in Datalog. The lecture had a number of easily accessible examples based on Oxford's Vadalog system using the Datalog language was shown. In this lecture, the theory and practice of reasoning in Knowledge Graphs were shown.

[8] https://www.cs.ox.ac.uk/publications/publication6640-abstract.html

[9] Luigi Bellomarini, Emanuel Sallinger, Georg Gottlob (2018) The Vadalog system: datalog-based reasoning for knowledge graphs, Proceedings of the VLDB Endowment - Proceedings of the 44th International Conference on Very Large Data Bases (VLDB), Rio de Janeiro, Brazil, Volume 11 Issue 9, May 2018, Pages 975-987.

Database: $\{A(1), C(1,2)\}$

Rules: $\{\forall x, y\ A(x) \wedge C(x, y) \rightarrow B(y, x)\}$

Query: $B(y, x)$

```
a(1).
c(1,2).

b(Y,X) :- a(X), c(X,Y).

@output("b").
```

Figure 8. Datalog Example 1

$$\forall t, i, n, e, z\ (\textbf{Report}(t, i, n, e, z) \wedge \textbf{Report}(t, i, n', e', z') \rightarrow n = n')$$

Figure 9. Datalog Example 2

3.3 Extraction for Knowledge Graphs:

The third lecture, see https://project-lambda.org/EKGs-Lecture-3 focuses on data extraction problems from the Web and in-use approaches with regard to Big Data. Data-driven websites are mostly accessed through search interfaces. Such sites follow a common publishing pattern that, surprisingly, has not been fully exploited for unsupervised data extraction yet: the result of a search is presented as a paginated list of result records. Each result record contains the main attributes about one single object, and links to a page dedicated to the details of that object. Several examples and challenges in data extraction are introduced in this lecture. Figure 11 illustrates an example of a visual wrapper used for extracting details of car dealerships. The numbers indicate the steps for user interaction.

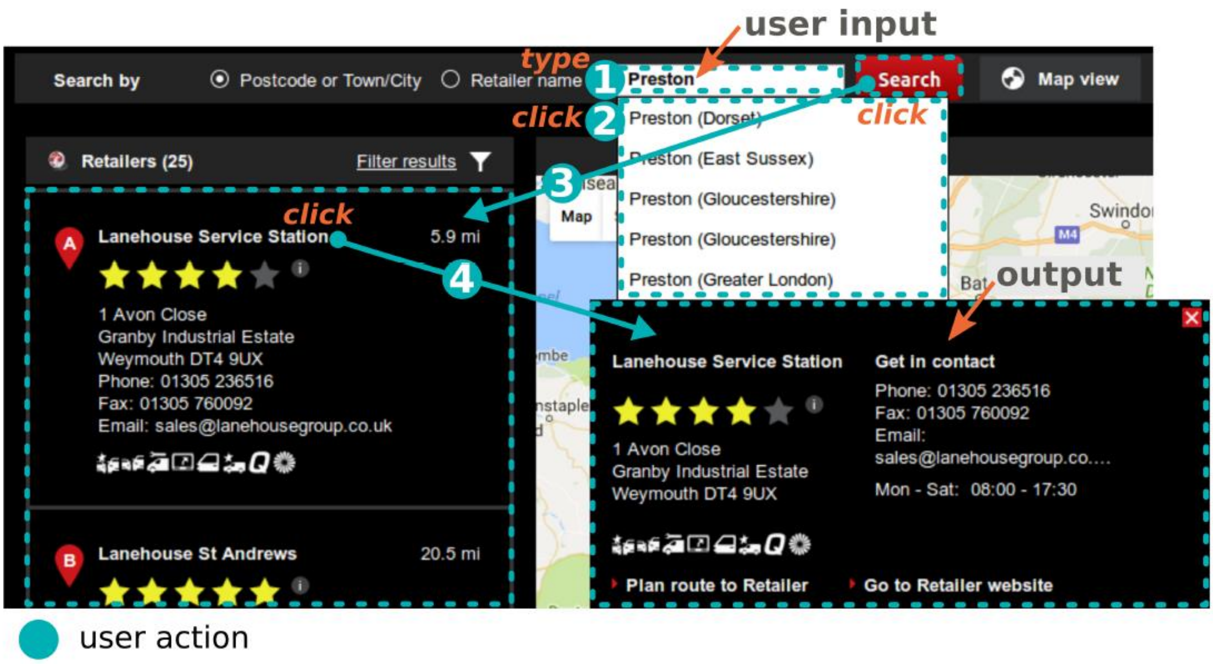


Figure 10. Web Data Extraction

One of the main challenges in the Big Data domain and KGs has been addressed in this lecture. Despite the large volume of data in such KGs, one of the main issues is their incompleteness. This limits the inference of knowledge and influences the performance of systems utilizing such KGs. An elegant solution to solve incompleteness of KGs is Knowledge Graph Embeddings (KGE).

4. Conclusion

This deliverable provides a summary of the work carried out in the WP3 Task 3. We provided a comprehensive discussion on KGs as the enabling technology for enterprise AI applications. Considering the flow of Big Data (Variety, Velocity, Variety) from different sources, recently, knowledge graphs (KGs) have been announced as one of the most important technologies for the next wave of artificial intelligence and knowledge management solutions across research and industries. Experts of the future require a deep understanding of the principles together with the advanced in-used technologies as well as the challenges. The lectures by Oxford has been developed to prepare the participants of BDA with a strong introduction to Big Data challenges and definition of Knowledge Graphs and how different phenomena and key phrases (e.g., Semantic Web, Link Data) appearing different in technology can, in fact, come together under the umbrella of Big Data. This provides necessary base camps as well as advanced technologies.