University of Dublin
Trinity College

CS7IS1
Knowledge and Data Engineering

Prof. Declan O'Sullivan
declan.osullivan@tcd.ie

PRELIMINARY REMARKS

College has asked us to remind you

1. Recommendation is to wear masks in crowded settings like lectures, tutorials etc.

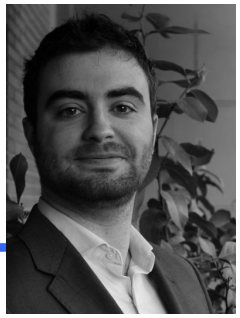
2. If you are unwell, do not attend college but take a Covid test, and seek materials that you miss
   - For this module, access can be arranged for access to recordings of lecture sessions

University of Dublin
Trinity College

CS7IS1
Knowledge and Data Engineering

Prof. Declan O'Sullivan
declan.osullivan@tcd.ie

Module Demonstrator: Albert Navarro
(https://www.adaptcentre.ie/experts/albert-navarro-gallinad/)
He is contactable through "**The Clinic"** Link on Blackboard



Reflection Task

You have all done some form of application development with database or knowledge base technologies

What properties of today's data do we need to deal with?

Properties of today's data

- Increasing Scale of data to be integrated
- Increasing uncertainty
  - *Co-ordinated vs autonomous publication*
  - *Manual data curation vs semi-automated curation (at best!)*
  - *Schemas created by DBAs/Experts vs end-user created/emergent schemas like tagging*
- Increasing dynamism
  - *Stable set of sources vs network churn*
  - *Sources known a priori vs runtime discovery of sources*
- Changing expressivity
  - *Relational data vs semi-structured or human-readable content*
  - *Integrity constraints vs no guarantees*
  - *Structured queries vs simple subject-predicate or free text queries*

Slide source: Rob Brennan

Huge Interoperability Problems

Diversity of data formats

Diversity of data platforms

Diversity of standards bodies

Can we have a machine readable network of data?

But if we do…

=> Which data can we trust, where did it come from (provenance)

=> What does the data really mean? (semantics: common understanding)

⇒ Quality of data – incomplete, wrong, malicious, …

⇒ Value of the data? (economic, social, etc.)

Representing knowledge

Many options

- As *objects*, using the well-accepted techniques of object-oriented analysis and design to capture a model
- As *clauses*, going back to the early days of AI and Lisp
- As *XML*, using the industry-standard structured mark-up language
- As *sets* of *entities* that support logical operations to define the sets and classify the entities
- As *graphs*, making use of the things we know about graph theory and entity-attribute-value models
- As some combination of these

Advantages of a Graph based approach

- Graphs provide a **concise and intuitive abstraction** for a variety of domains, where edges capture the (potentially cyclical) relations between the entities
- Graphs allow maintainers to **postpone the definition of a schema**, allowing the data – and its scope – to evolve in a more flexible manner
- Specialised graph query languages support not only standard **relational operators** (joins, unions, projections, etc.), but also **navigational operators** for recursively finding entities connected through arbitrary-length paths

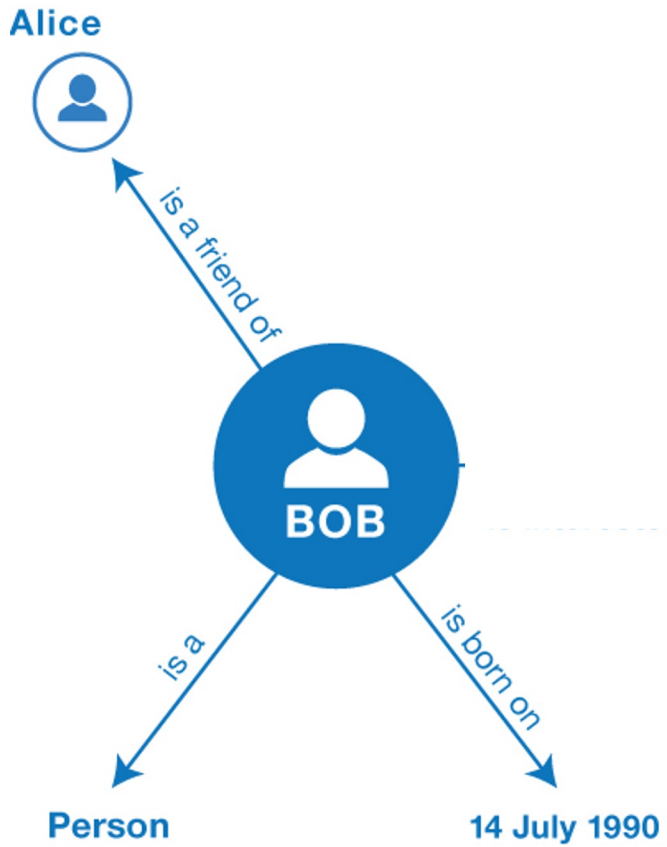
Advantages of a Graph based approach

- Standard knowledge representation formalisms – such as **ontologies and rules** – can be employed to define and reason about the semantics of the terms used to label and describe the nodes and edges in the graph
- Scalable frameworks for **graph analytics** can be leveraged for computing centrality, clustering, summarisation, etc., in order to gain insights about the domain being described.
- Various representations have also been developed that **support applying machine learning techniques** both directly and indirectly over graphs (e.g. Knowledge Graph Embeddings)

In summary, the decision to build and use a knowledge graph opens up a range of techniques that can be brought to bear for integrating and extracting value from diverse sources of data at large scale.

Slide source: Hogan et al. : KG Book

Simple Example



Simple Example

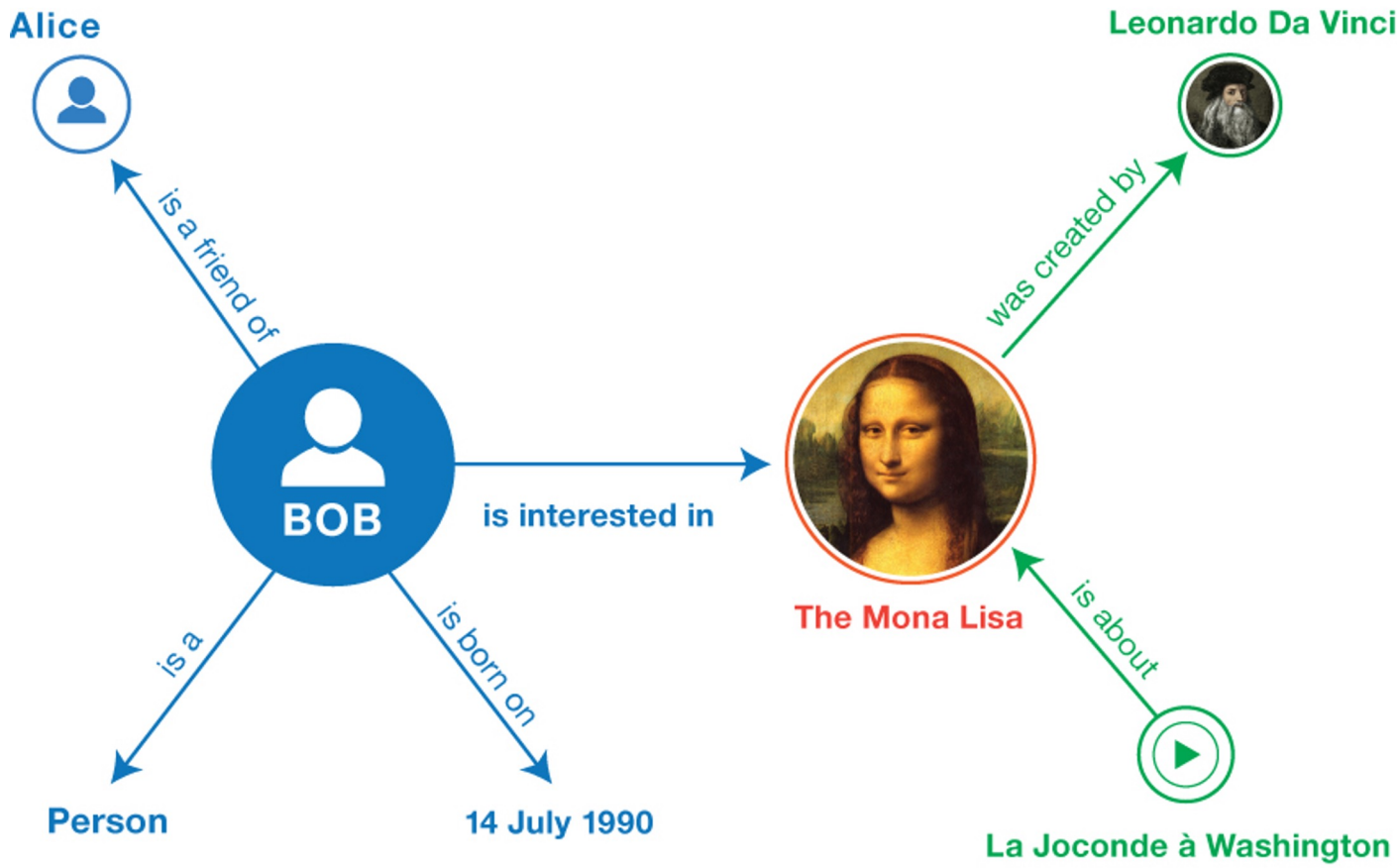
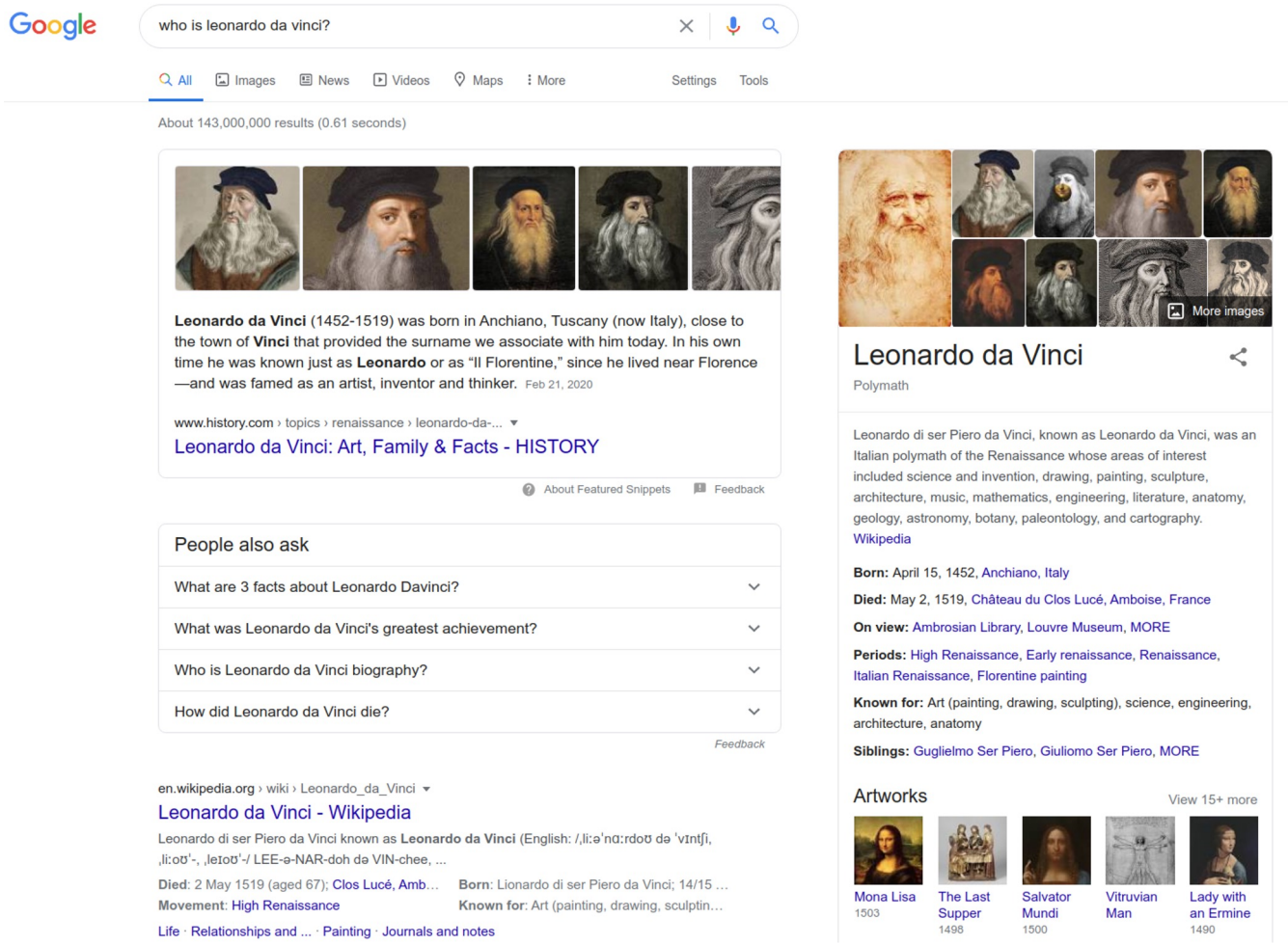


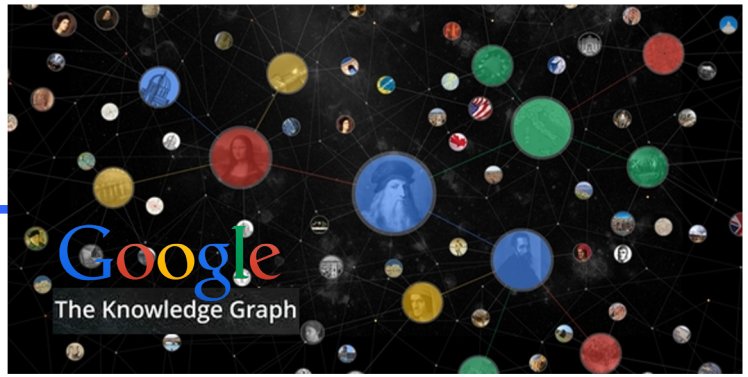
Image source: https://www.w3.org/TR/rdf11-primer/

Where do you come across Knowledge Graphs in action?



Slide source: Fabrizio Orlandi

Example: Google Knowledge Graph



Google who is leonardo da vinci?

All Images News Videos Maps More Settings Tools

About 143,000,000 results (0.61 seconds)

Leonardo da Vinci (1452-1519) was born in Anchiano, Tuscany (now Italy), close to the town of **Vinci** that provided the surname we associate with him today. In his own time he was known just as **Leonardo** or as "Il Florentine," since he lived near Florence —and was famed as an artist, inventor and thinker. Feb 21, 2020

www.history.com › topics › renaissance › leonardo-da-...

Leonardo da Vinci: Art, Family & Facts - HISTORY

About Featured Snippets Feedback

People also ask

- What are 3 facts about Leonardo Davinci?
- What was Leonardo da Vinci's greatest achievement?
- Who is Leonardo da Vinci biography?
- How did Leonardo da Vinci die?

Feedback

en.wikipedia.org › wiki › Leonardo_da_Vinci

Leonardo da Vinci - Wikipedia

Leonardo di ser Piero da Vinci known as **Leonardo da Vinci** (English: /ˌliːəˈnɑːrdoʊ də ˈvɪntʃi, ˌliːoʊˈ-, ˌleɪoʊˈ-/ LEE-ə-NAR-doh də VIN-chee, ...

Died: 2 May 1519 (aged 67); Clos Lucé, Amb... **Born**: Lionardo di ser Piero da Vinci; 14/15 ...
Movement: High Renaissance **Known for**: Art (painting, drawing, sculptin...

Life · Relationships and ... · Painting · Journals and notes

Leonardo da Vinci

Polymath

Leonardo di ser Piero da Vinci, known as Leonardo da Vinci, was an Italian polymath of the Renaissance whose areas of interest included science and invention, drawing, painting, sculpture, architecture, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, paleontology, and cartography. Wikipedia

Born: April 15, 1452, Anchiano, Italy

Died: May 2, 1519, Château du Clos Lucé, Amboise, France

On view: Ambrosian Library, Louvre Museum, MORE

Periods: High Renaissance, Early renaissance, Renaissance, Italian Renaissance, Florentine painting

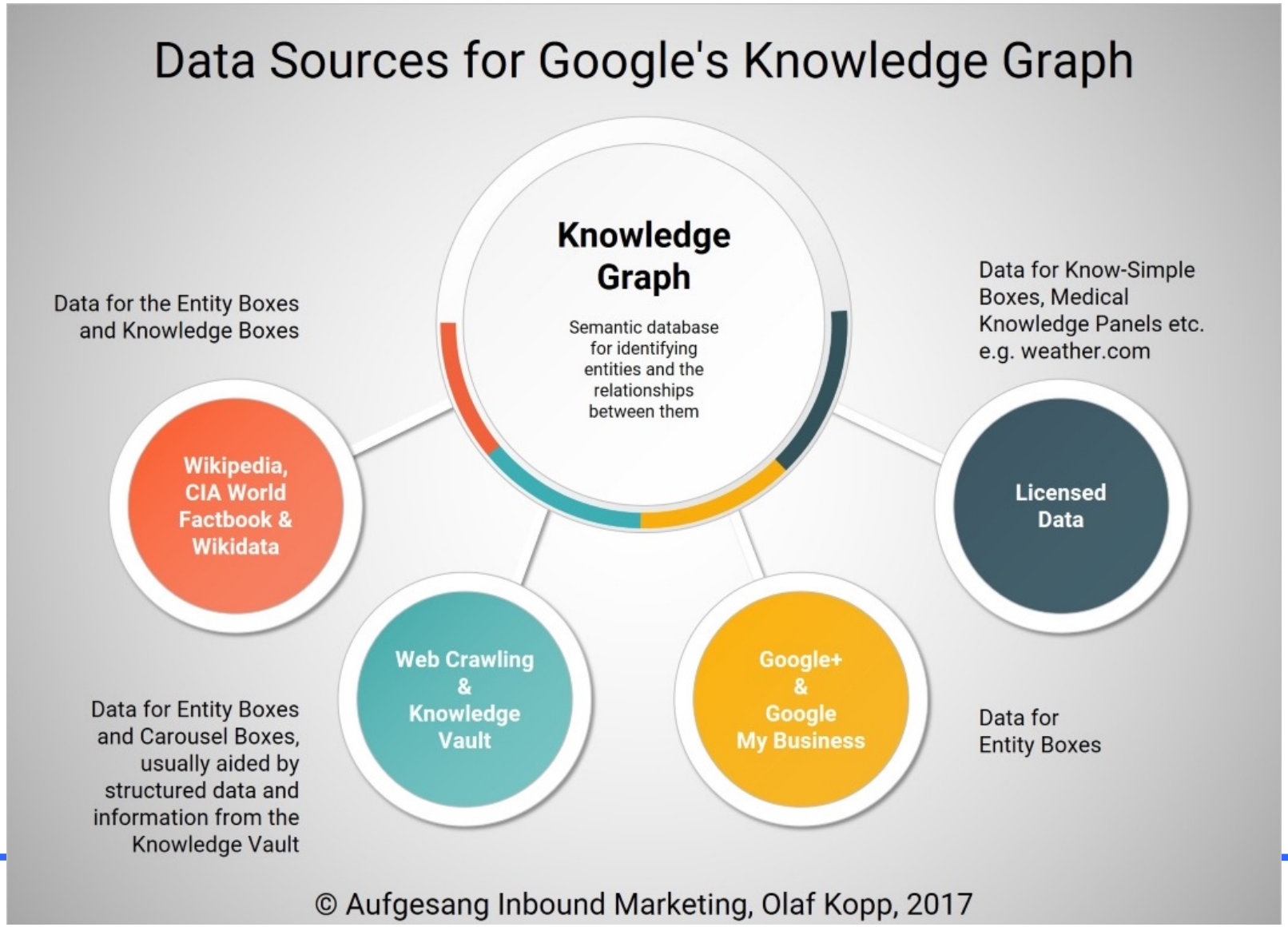
Known for: Art (painting, drawing, sculpting), science, engineering, architecture, anatomy

Siblings: Guglielmo Ser Piero, Giuliomo Ser Piero, MORE

Artworks View 15+ more

Mona Lisa 1503 · The Last Supper 1498 · Salvator Mundi 1500 · Vitruvian Man · Lady with an Ermine 1490

Where does Google draw its KG data from?



A definition of Knowledge Graph (KG)

"*a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities.*"

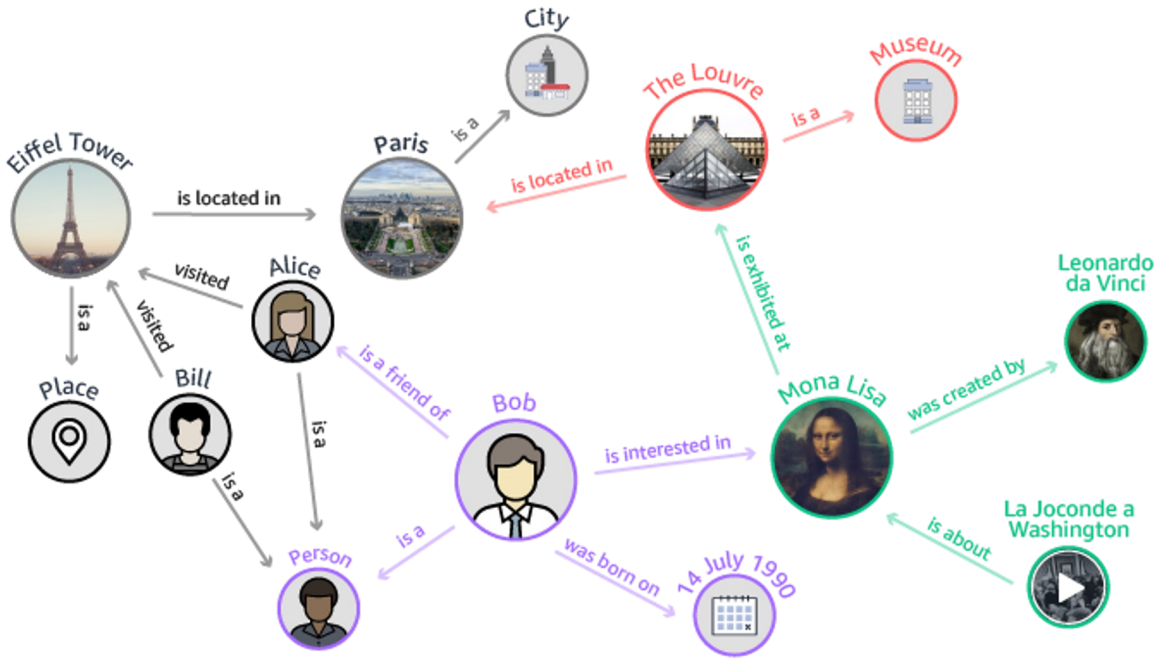


Image source: https://aws.amazon.com/neptune/

Slide source: Hogan et al.: KG Book

Representing Knowledge in a Knowledge Graph

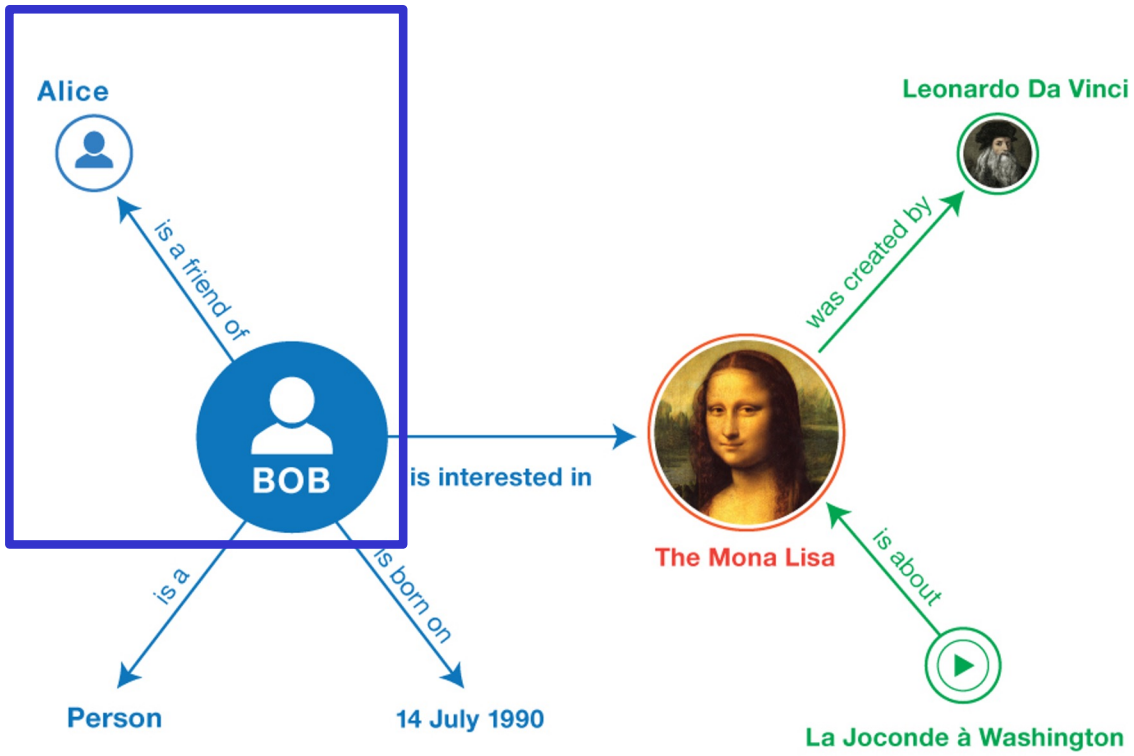
Knowledge may be composed of simple statements, such as “Dublin is the capital of Ireland”, or quantified statements, such as “all capitals are cities”

If the knowledge graph intends to accumulate quantified statements, a more expressive way to represent knowledge – such as ontologies or rules – is required. Deductive methods can then be used to entail and accumulate further knowledge (e.g., “Dublin is a city”).

Additional knowledge – based on simple or quantified statements – can also be extracted from and accumulated by the knowledge graph using inductive methods.

Slide source: Hogan et al.: KG Book

How to make a graph “machine-processable”?

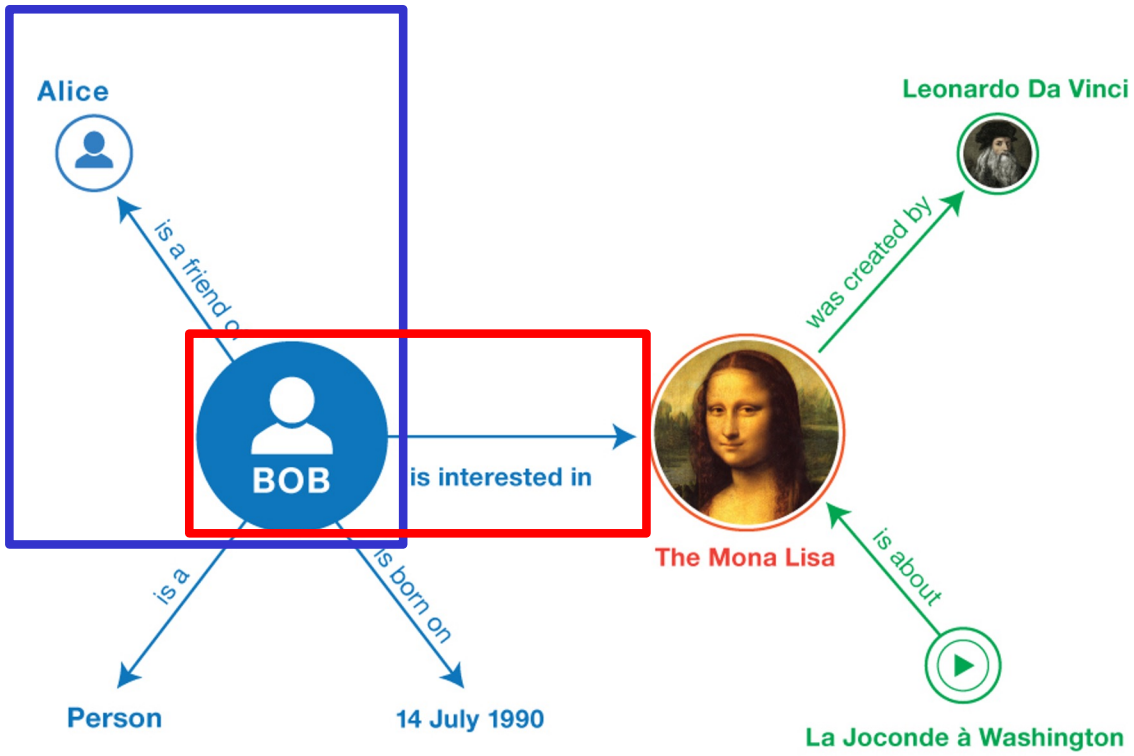


Subject (node)	Predicate (edge)	Object (node)
BOB	is a friend of	Alice
…	…	…

Image source: https://www.w3.org/TR/rdf11-primer/

Slide source: Fabrizio Orlandi

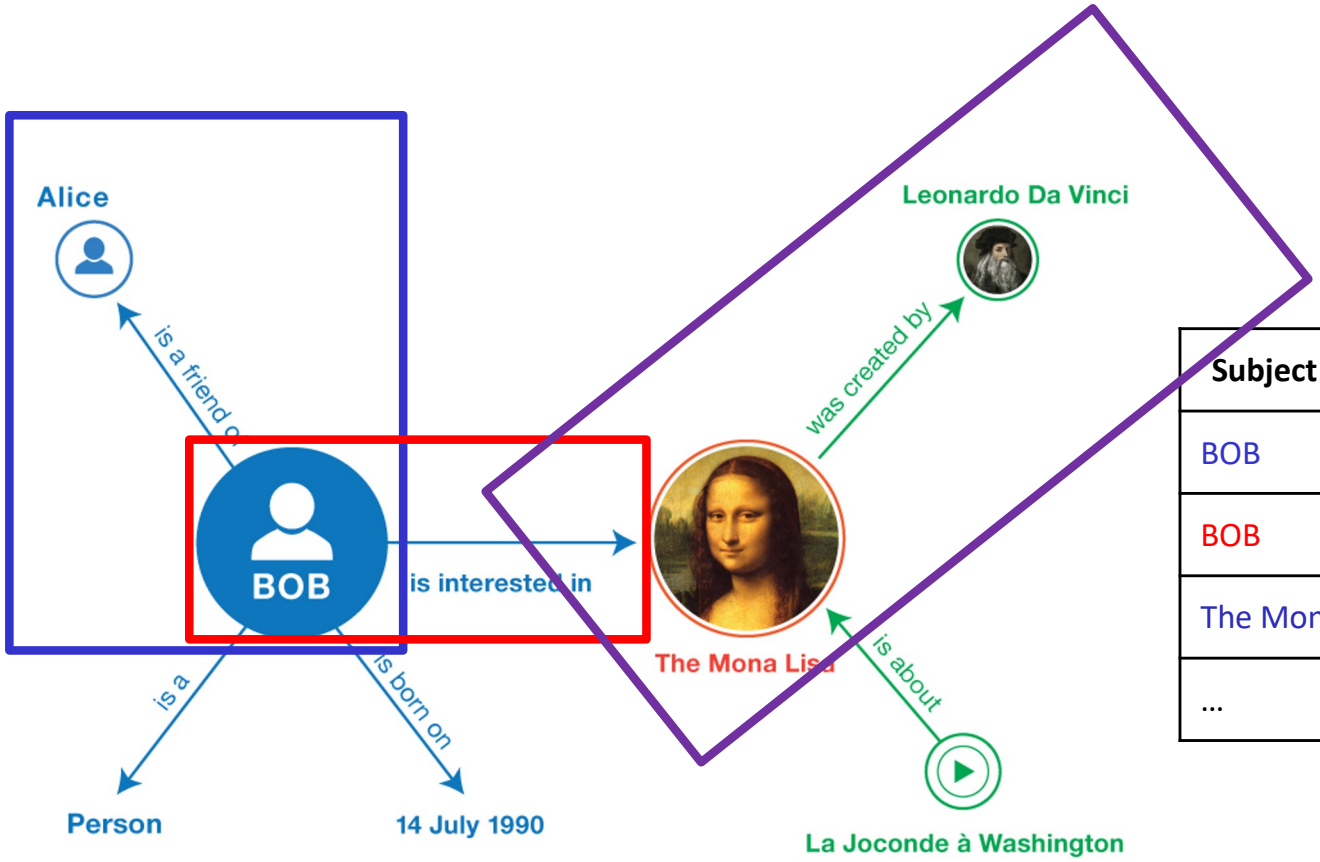
How to make a graph “machine-processable”?



Subject (node)	Predicate (edge)	Object (node)
BOB	is a friend of	Alice
BOB	is interested in	The Mona Lisa
…	…	…

Image source: https://www.w3.org/TR/rdf11-primer/

How to make a graph “machine-processable”?



Subject (node)	Predicate (edge)	Object (node)
BOB	is a friend of	Alice
BOB	is interested in	The Mona Lisa
The Mona Lisa	was created by	Leonardo Da Vinci
…	…	…

Image source: https://www.w3.org/TR/rdf11-primer/

Slide source: Fabrizio Orlandi

How to make data graph accessible on the web

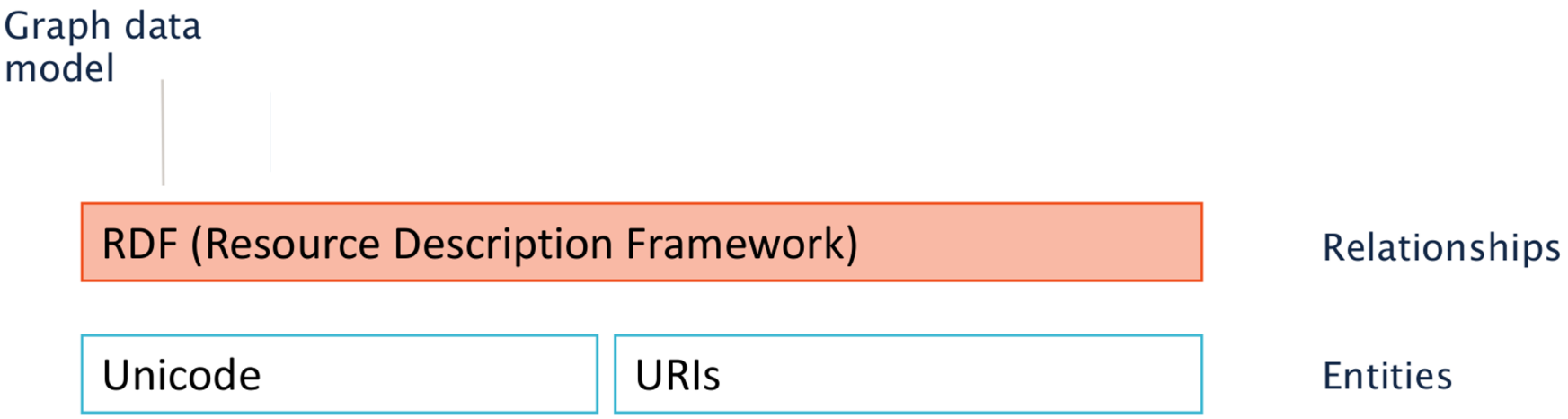
The **Semantic Web (SW)** is an extension of the current Web where data is given well defined meaning and where the relationships between data, and not just documents, are defined in a common machine-readable format - creating a Web of Data.

Linked Data (LD) describes a set of principles and best practices for publishing, interlinking and engaging with graph data using **standard web technologies**.

These principles include the use of HTTP Uniform Resource Identifiers (URIs) for naming resources and for retrieving data using the existing HTTP stack.

Standards-based Eco-system for processing graph data (W3C)

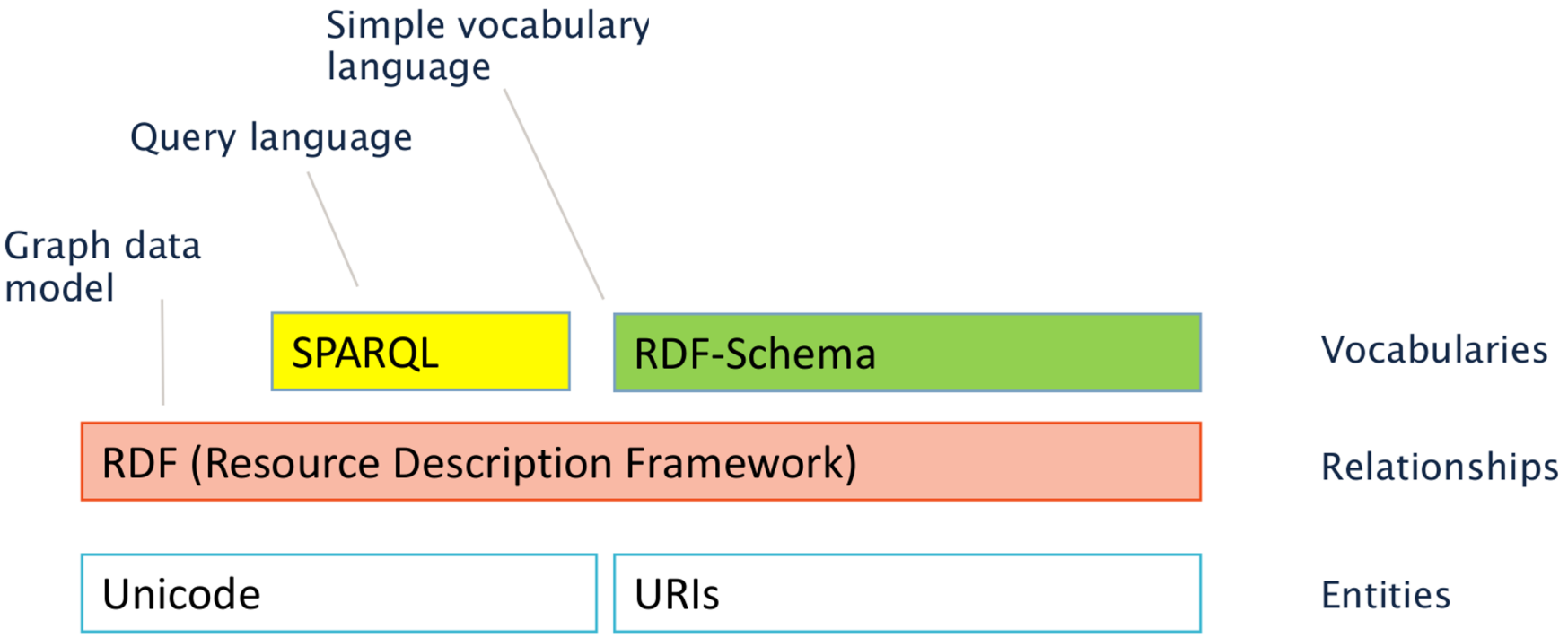
We will focus in this module on W3C standards for Knowledge Graphs whenever possible



Slide source: Fabrizio Orlandi

Standards-based Eco-system for processing graph data (W3C)

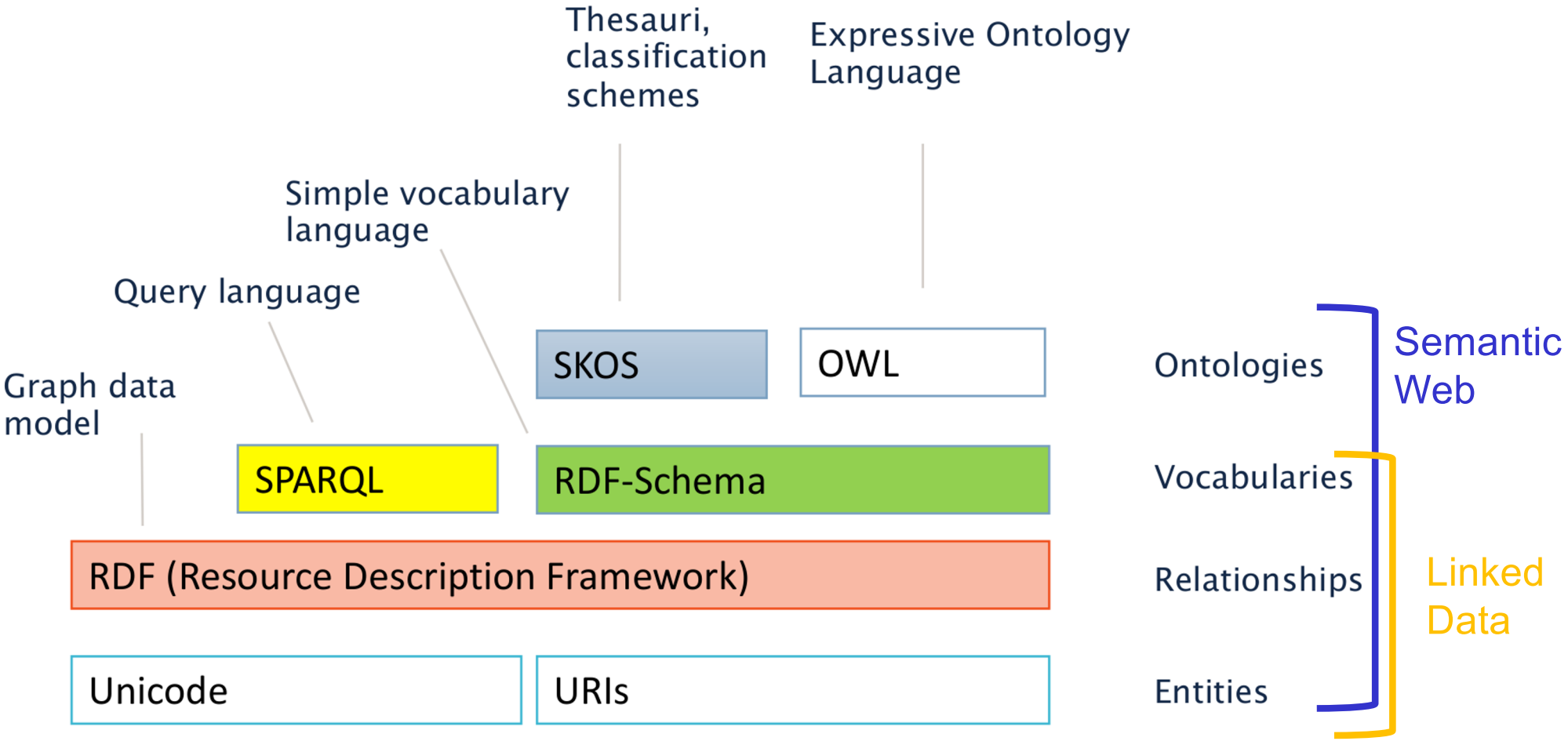
We will focus in this module on W3C standards for Knowledge Graphs whenever possible



Slide source: Fabrizio Orlandi

Standards-based Eco-system for processing graph data (W3C)

We will focus in this module on W3C standards for Knowledge Graphs whenever possible



CS7IS1 Module Aim

To explore the knowledge and data management challenges for 21st Century:

- Knowledge representation and management
- Querying knowledge bases
- Knowledge interoperability
- Publishing and linking data on the web

=> Enabling knowledge-driven intelligent systems that can deal with the current data deluge

Topic Coverage

Building, maintaining and using knowledge-based systems:

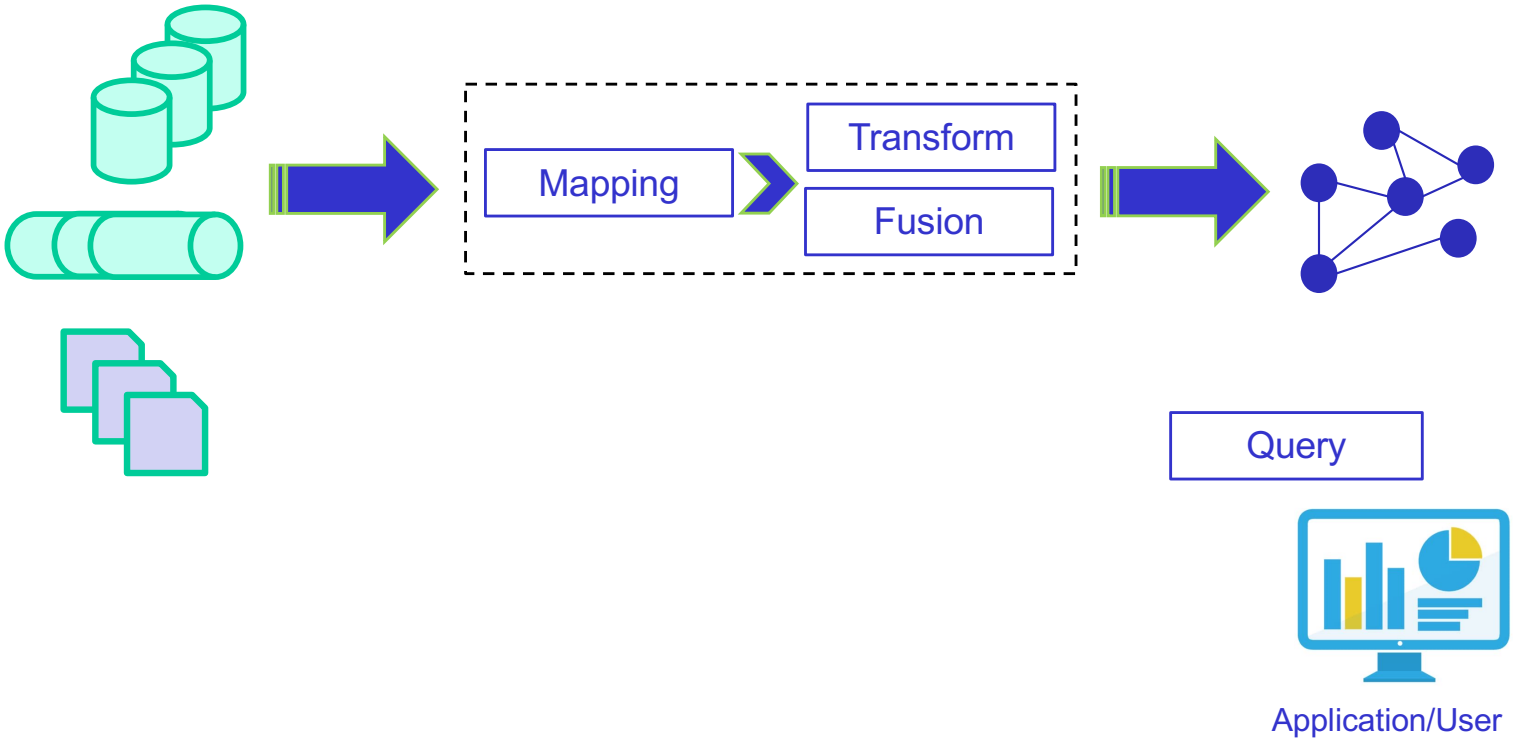
- Representing and manipulating knowledge in computer systems
  - Knowledge Graph/ Linked Data/Semantic Web
- Integrating knowledge and data from diverse sources
  - Mappings, Uplift, Provenance, Data Governance

Learning Outcomes

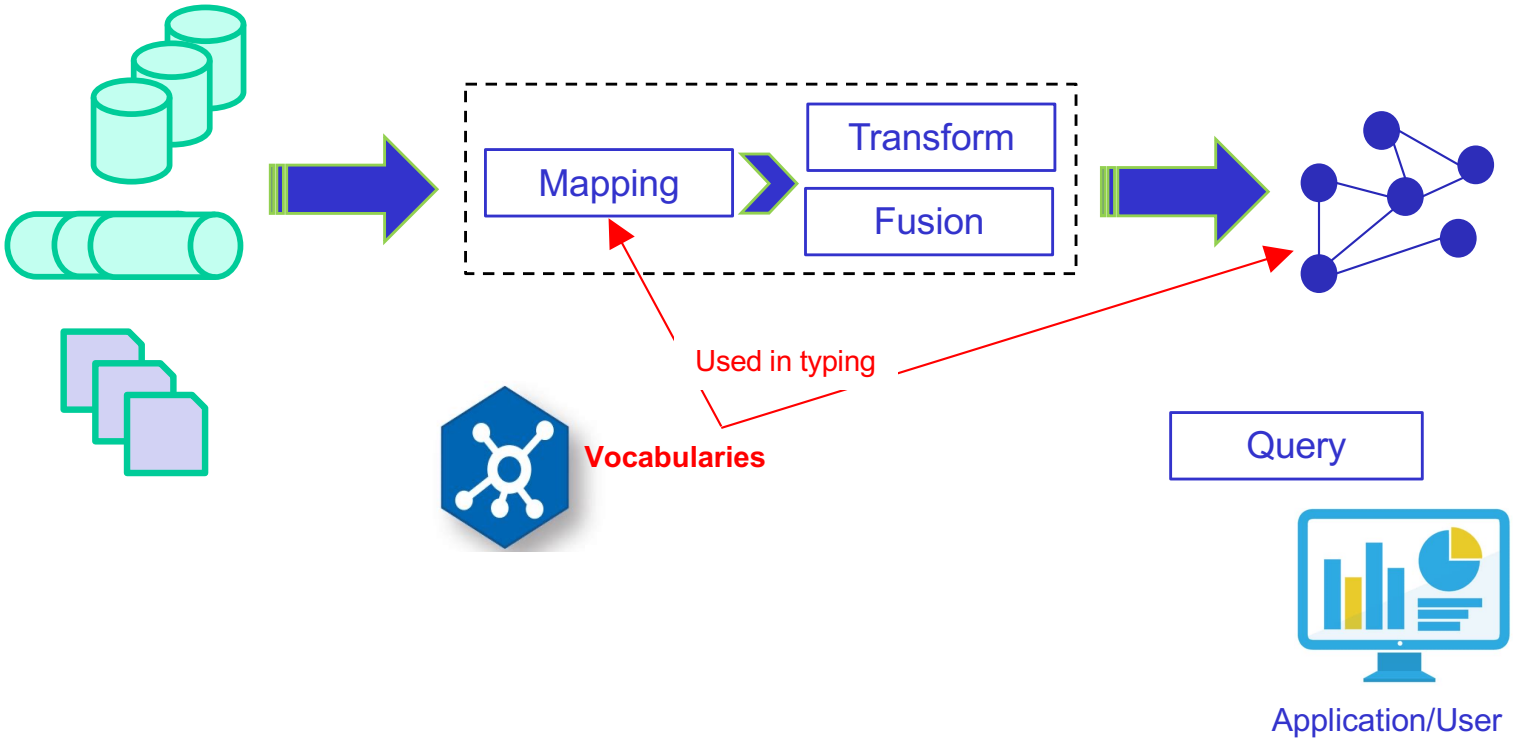
By the end of this module you should be able to:

- Compare and contrast different approaches to **modelling** information and knowledge
- Model information and produce rich **semantic models and ontologies**
- Survey the **state of the art** in semantic technologies and applications
- Demonstrate a clear understanding of the principles underlying **information interoperability** and **transformation**
- **Apply** semantic modelling and transformation techniques to a range of applied problems
- Use sophisticated **querying approaches** to facilitate distributed information retrieval and aggregation

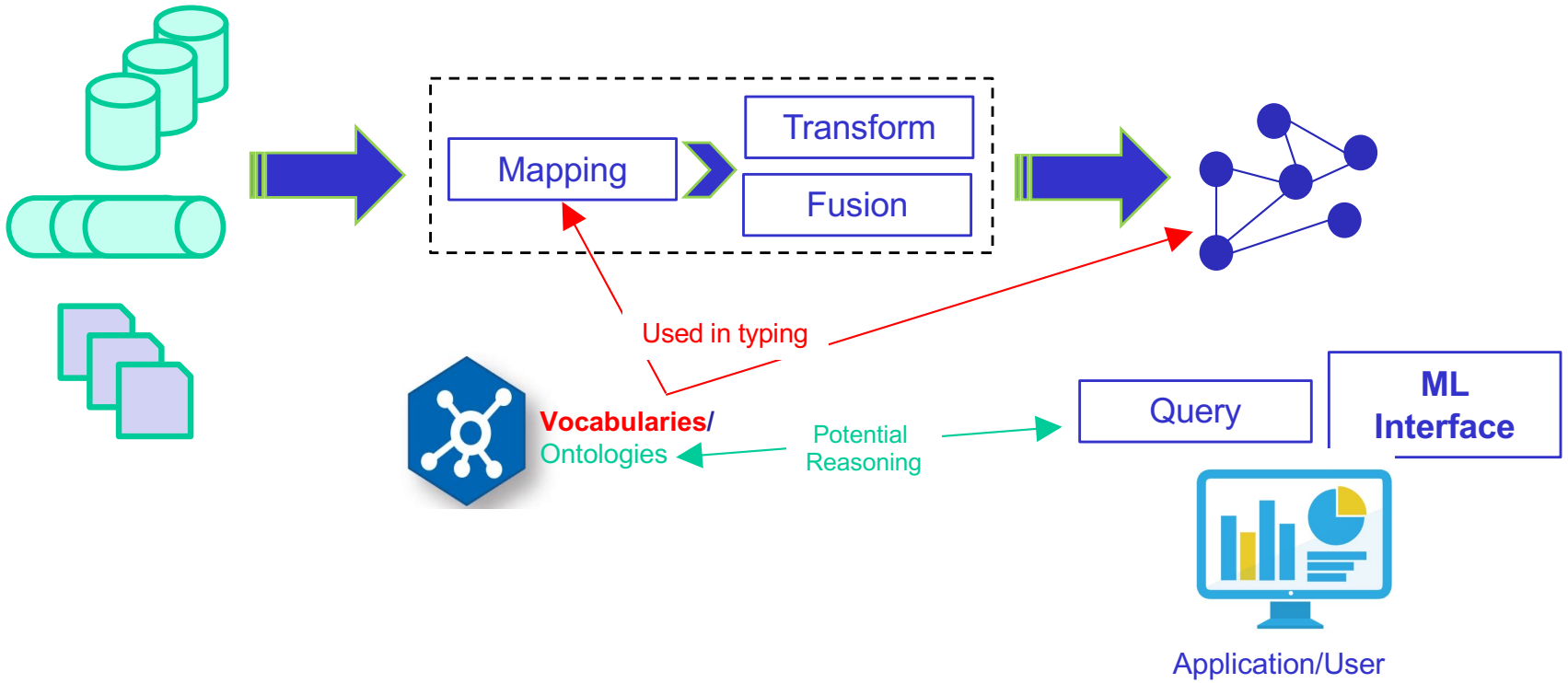
Rough Abstraction technologies in pipeline addressed in Module



Rough Abstraction technologies in pipeline addressed in Module



Rough Abstraction technologies in pipeline addressed in Module



Rough Roadmap for module

KG representation - RDF and RDFS

querying KGs - SPARQL

Uplift mappings – transform non-RDF to RDF

Ontology engineering – Semantics, OWL

Linked Data – publishing KGs

Interlinking – linking RDF to RDF

Advanced Topics: ML/KG, Context and KGs

Housekeeping

Module slides, reading material and lab info will be made available via the TCD Blackboard system

- Session Materials available **around 10pm evening before** each session

  …. But note that tasks and various solutions posted after session with updated Session Materials

- ”The Clinic” discussion board – monitored by Module Demonstrator

Assessment: 100% Coursework

- 35% for Individual Research task
- 50% for Group Project
- 15% for Individual Portfolio

Individual Research task: Recorded presentation critique of identified research paper (details given **TODAY**)

Group Project: Development of Knowledge Graph (details given during week Oct 3rd)

Individual portfolio: self directed activities/practical tasks to complete, that contribute towards your ability to execute the group project e.g. small tutorial exercises (ongoing)

ASSIGNMENT:
YOUR PORTFOLIO

(15% OF FINAL MARKS)

Your Portfolio

During the module there will be a number of **small** Self Directed Tasks that you are asked to do to help reinforce a concept we cover or practice a skill

The outputs of these individual tasks are **not marked**, but your engagement with the module as demonstrated in engaging with these tasks will be marked

The idea of the Portfolio is to gather your outputs in one place

- Under “Assignments” link you will see “Your Portfolio” discussion forum
  - Available after 2pm today

- Simply create **ONCE** a Thread with your **NAME**, and then in that thread put the output for each self directed task as a separate post in the thread with your name

To avoid a backlog for yourself (and to aid in group project activities), I normally **suggest a target date to aim for submission**

For Portfolio

Self Directed Task 1 (SDT1) – Semantic Web Paper

Read the following paper and briefly (say one paragraph for each question) add to your portfolio your own personal thoughts on the answers to the following questions about this paper (suggested submission by Monday 19th September):

1. What did you find interesting about the paper?

2. What do you believe are blockers to the widespread adoption and implementation of this approach?

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). **The semantic web**. Scientific American, 284(5), 34-43. https://www.lassila.org/publications/2001/SciAm.pdf

For Portfolio

Self Directed Task 2 (SDT2) – KG Review Paper

Review the following paper and briefly (say one paragraph for each question) add to your portfolio your own personal thoughts on the answers to the following questions about this paper (suggested submission by Monday 19th September)

1. What did you find interesting about the paper?

2. What do you believe are blockers to the widespread adoption and implementation of this approach?

Hogan et al. ***Knowledge Graphs****. ACM Comput. Surv. 54, 4, Article 71 (May 2022), 37 pages.*

https://doi.org/10.1145/3447772

ASSIGNMENT:
INDIVIDUAL RESEARCH TASK

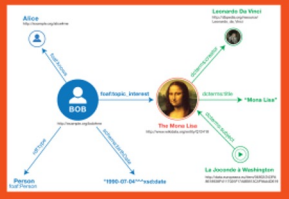
(35% OF FINAL MARKS)

Individual Research Task Steps

1. Pick either:
   - **Good** technical paper on one of the underlying technologies
   - **Good** application paper of where semantic web/linked data/knowledge graph is being applied
2. Propose it (see later slide)
3. Record your presentation on the selected paper
   - **Minimum** 4 minutes
   - **Maximum** 6 minutes
4. Upload recorded presentation **in MP4 format** to Blackboard Assignment area **before 5pm on Friday September 30th**
5. Be prepared to answer questions on the presentation at live sharing session of selected videos

Papers related to W3C based KG techniques

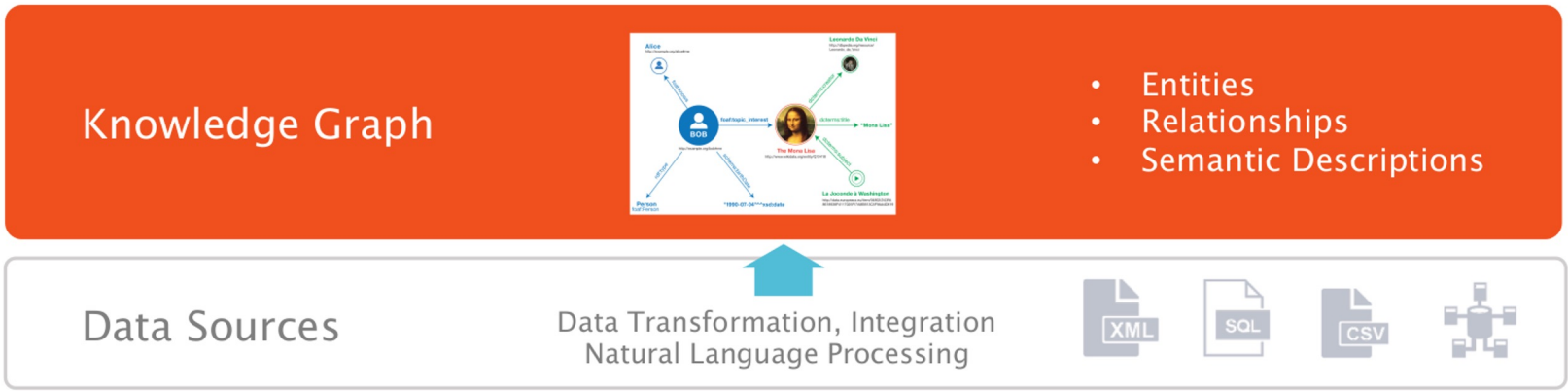
Knowledge Graph



- Entities
- Relationships
- Semantic Descriptions

P. Haase et al., **Enterprise KGs tutorial**, ISWC 2018

Papers related to constructing W3C based KGs



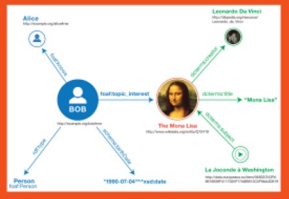
P. Haase et al., **Enterprise KGs tutorial**, ISWC 2018

Papers related to Applications directly using W3C based KGs

Applications

- Semantic Search
- Analytics
- Knowledge Sharing
- Question Answering
- Dashboards
- Knowledge Management

Knowledge Graph



- Entities
- Relationships
- Semantic Descriptions

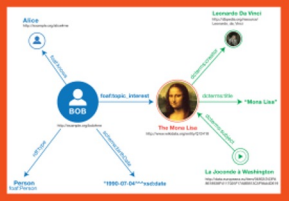
P. Haase et al., **Enterprise KGs tutorial**, ISWC 2018

Papers related to Algorithms interacting with W3C based KGs

Algorithms

- Inferencing
- Machine Learning
- Entity Recognition
- Disambiguation
- Text Understanding
- Recommendations

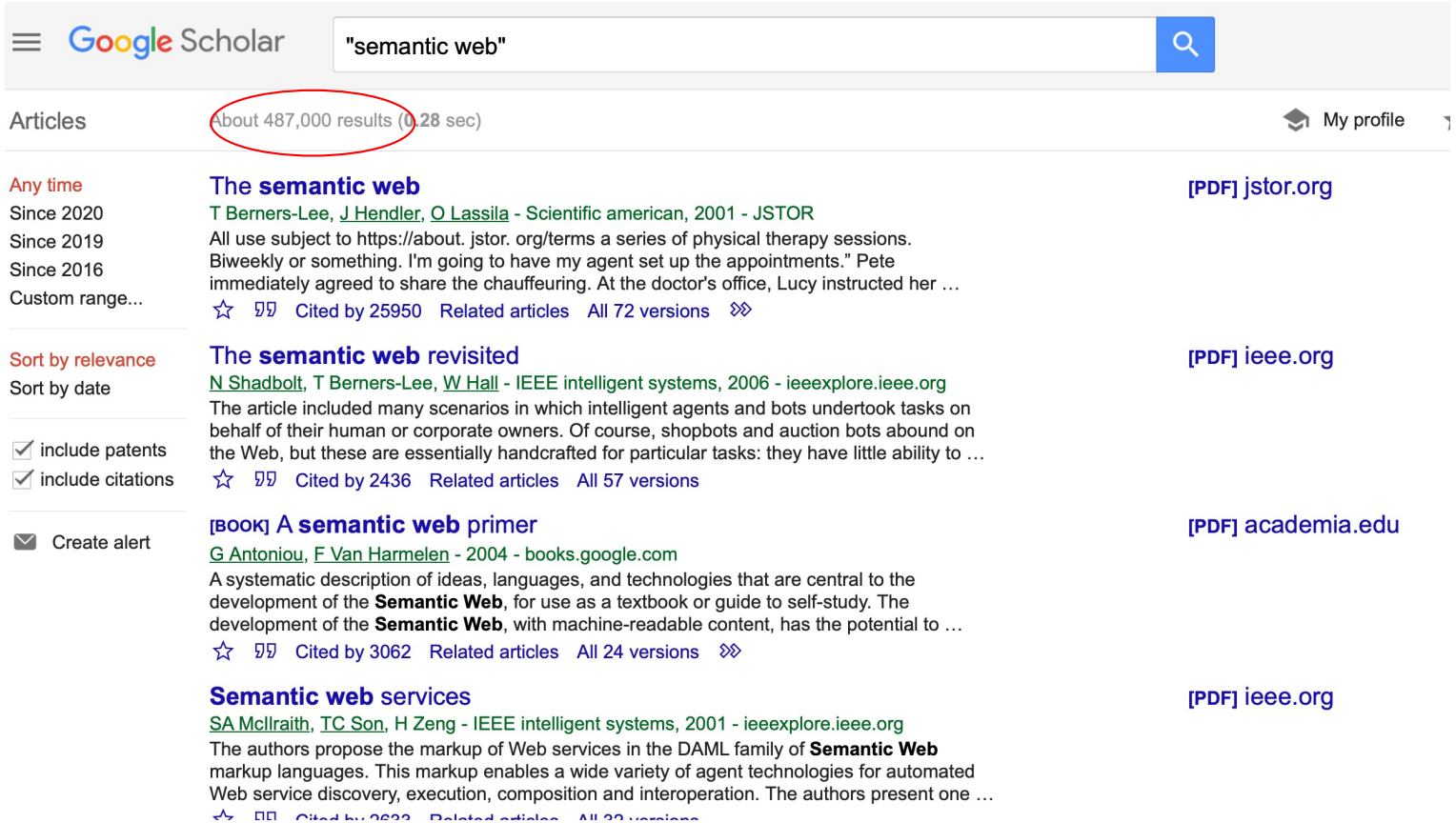
Knowledge Graph



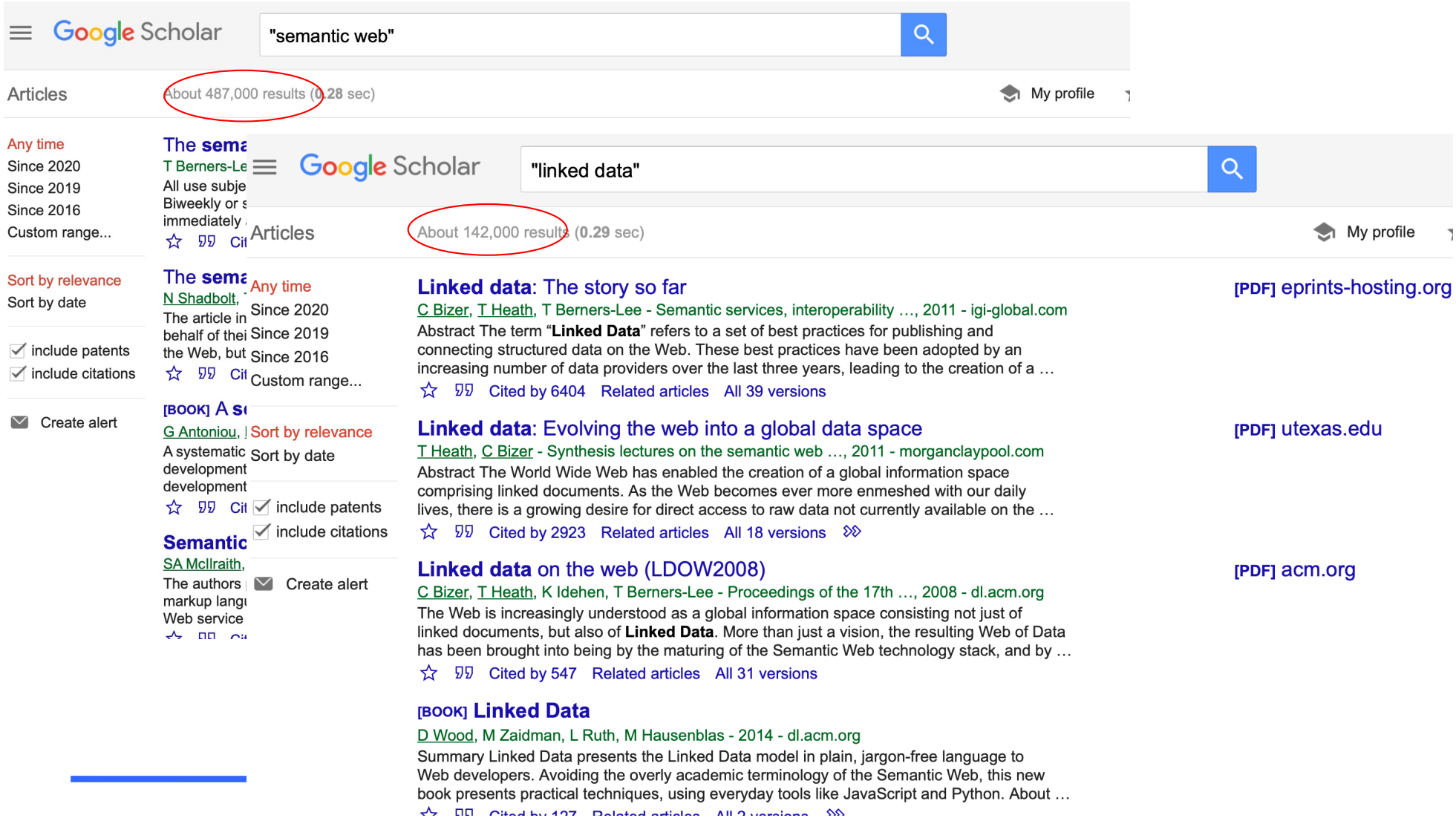
- Entities
- Relationships
- Semantic Descriptions

P. Haase et al., **Enterprise KGs tutorial**, ISWC 2018

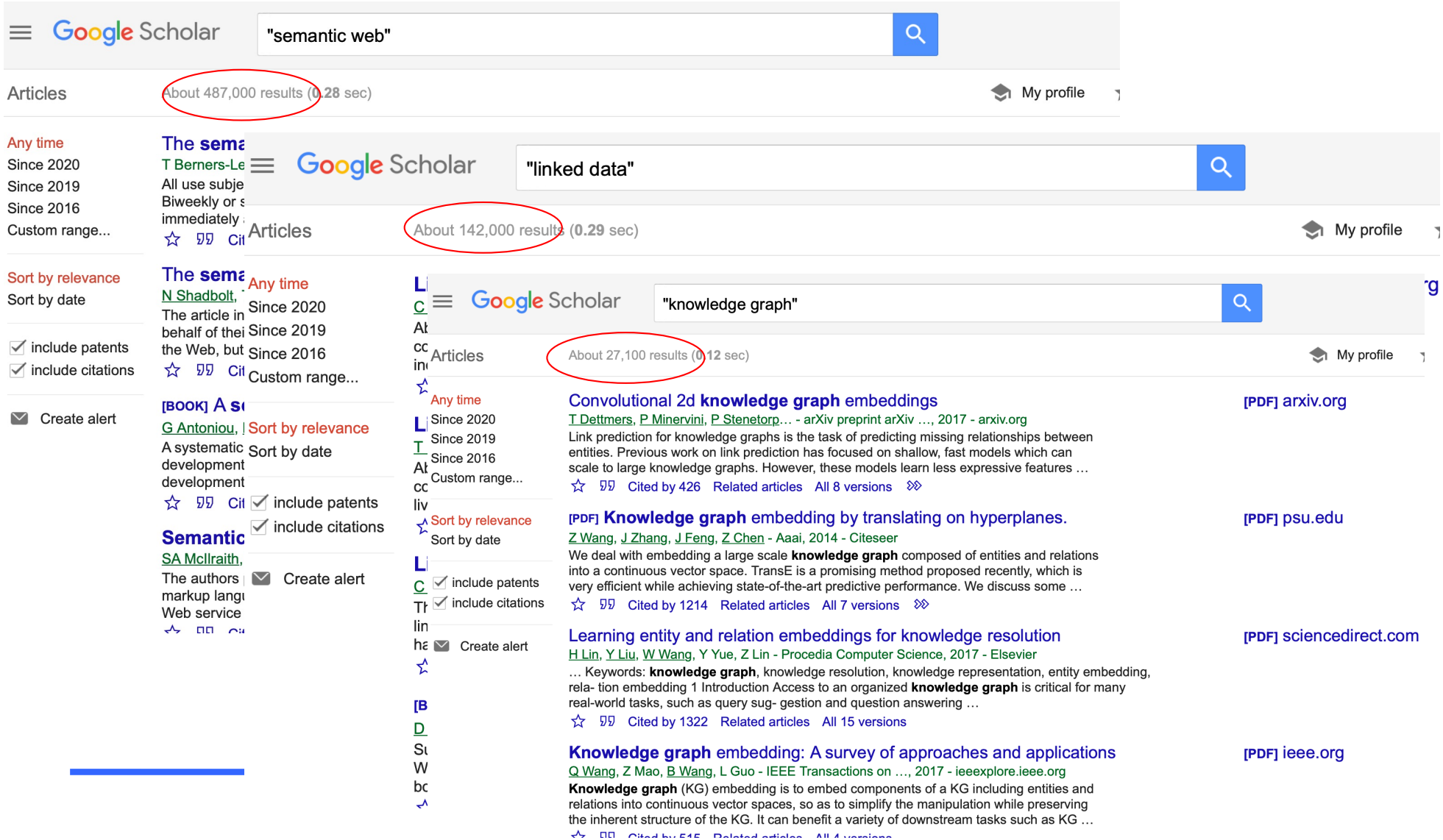
What makes a “good” paper?



What makes a “good” paper?



What makes a “good” paper?



Requirements for CS7IS1 ”good” paper

Must have Greater than 5 pages of technical content (excluding references)

Must be Published IN 2019 or AFTER 2019

Must be published in a Reputable Journal or Reputable Conference

Reputable Journals in the area:

- Journal of Web Semantics (JWS)
- Semantic Web Journal (SWJ)
- International Journal on Semantic Web and Information Systems (IJSWIS)
- ACM Transactions on XXXXX (e.g. ACM Transactions on the Web) https://dl.acm.org/journals

Reputable Conference Conference only (**not** associated workshop, doctorial consortium etc.)

- International Semantic Web Conference (ISWC)
- Extended Semantic Web Conference (ESWC)
- World Wide Web Conference
- International Conference on Semantic Systems (SEMANTICS)
- International Conference on Semantic Computing (ICSC)
- ACM or IEEE sponsored conference (not just indexed or published by them.. Will mention in the title)

How to book

1. Check relevant **Blackboard Module Assignment Area "Individual Research Task" Discussion Board**, to make sure nobody else has already booked the same paper **or** exactly same application area **or** exactly same technology topic

2. "Propose your paper" by posting the FULL reference/citation to the Blackboard Discussion Board BEFORE **Sunday September 18th** at the very latest
   - Posting **MUST** include:
     a) Full citation for the paper and ****online link to the paper****
     b) Tag whether it is an application paper or technology paper
     c) Tag whether it is Semantic Web or Linked Data or Knowledge Graph paper

3. ****Wait ** until I have "approved"** your proposal via my reply to your post on the Discussion Board before going onto the next step

Presentation Structure

Remember you are recording your voice giving the presentation, so there is no need to overload slides with text; and do use diagrams/visuals to help illustrate points

Slide 1: Your name; your course**; **Full title of paper; full citation; Hyperlink to the paper**

Slide 2: Area/Motivation of problem being addressed

Slide 3: Overview of solution being proposed

Slide 4, 5, 6: Go into detail, include evaluation undertaken if available

Slide 7: Your Reflection on the paper

a) What the authors considered novel or the major contribution

b) What you yourself considered interesting

c) What the paper you think tells us about the state of the art in Semantic Web or Linked Data or Knowledge Graphs

** course=

MSc (Data Science or **Intelligent Systems** or **Future Networks** or **Augmented & Virtual Reality)** or **MAI** or **MSc ICS Year 5;**

Individual Research Task by Declan O'Sullivan
Course = XXXXXXX

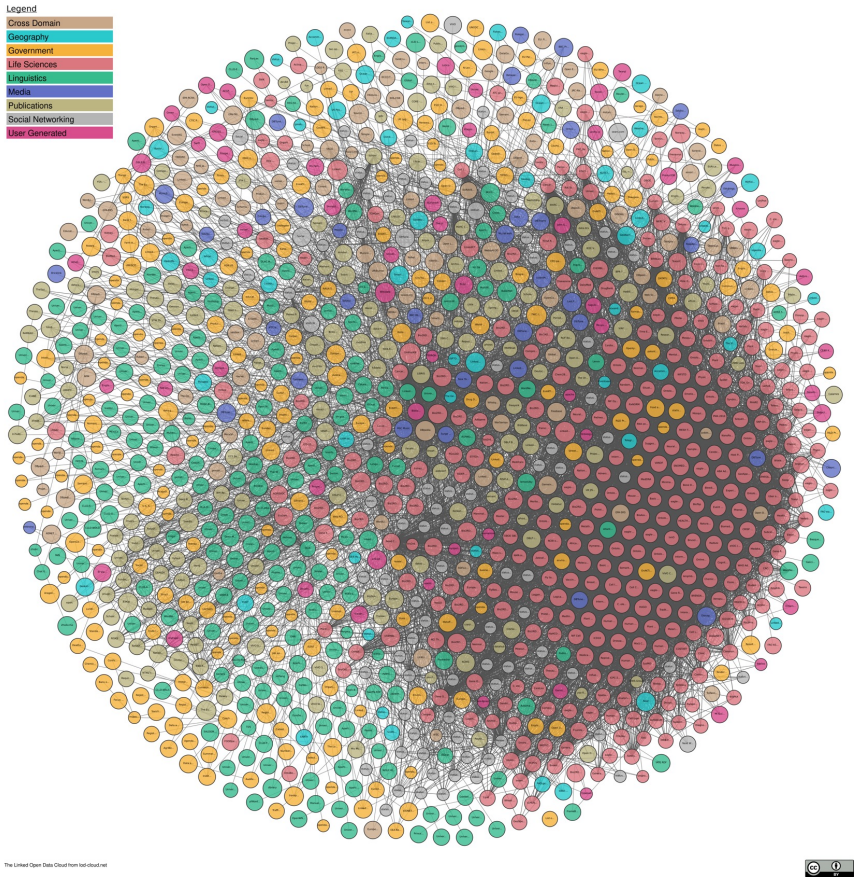
Luzzu—a methodology and framework for linked data quality assessment

Tag: **Linked Data** paper
Tag: **Technology** paper

J. Debattista, S. Auer, and C. Lange. 2016. Luzzu—a methodology and framework for linked data quality assessment. **Journal of Data and Information Quality (JDIQ**) 8, 1, 4**.**

https://dl-acm-org.elib.tcd.ie/doi/pdf/10.1145/2992786

Area/Motivation



- Increasing number of Datasets published as Linked Open Data
- Increasingly difficult to provide to consumers indicators of data quality
- Outputs from SoA Data Quality tools not easily machine processable to analyse or rank datasets

Overview of Solution proposed

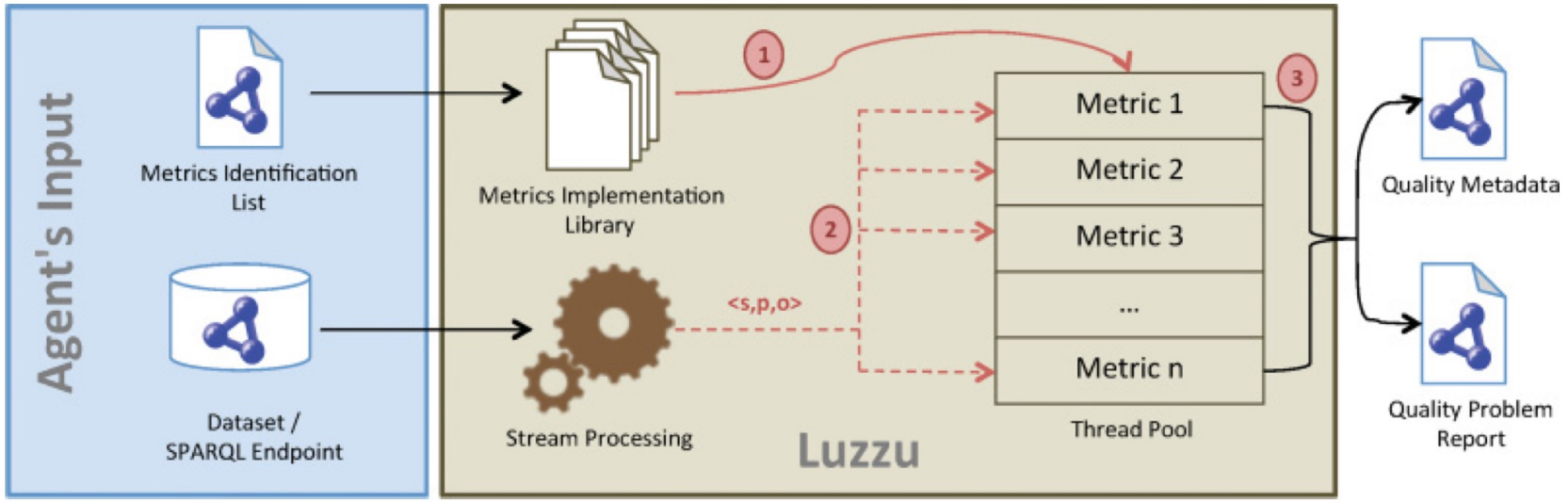
Article proposes:

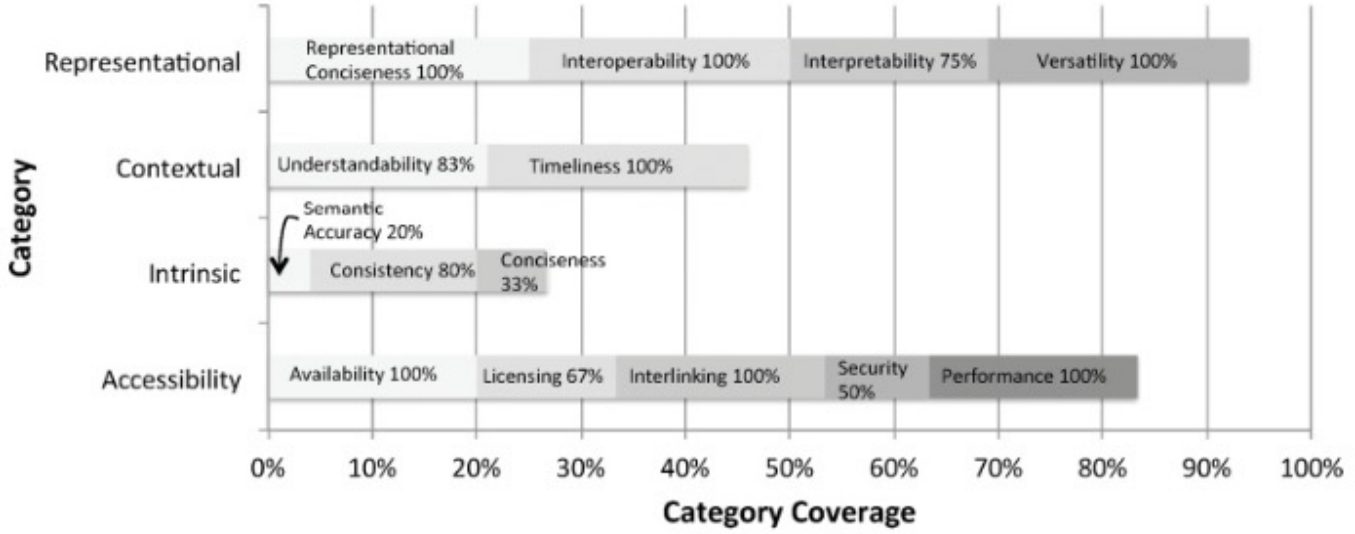
1. A conceptual methodology for assessing Linked Data quality
2. Luzzu, a quality assessment framework for Linked Data
3. Evaluation of the framework's applicability to a number of statistical linked open datasets against relevant quality metrics

Quality Assessment – A Conceptual Methodology

1. Identify Quality Measures for the task at hand
   - What are the important characteristics of my task?
2. Re-use or define quality metrics
3. Prepare the quality assessment
   a) Access point of dataset in question
   b) External Resources such as gold standard
4. Running the quality assessment
5. Assessment representation
   a) Immediate use
   b) Mid-to-long term use

Luzzu – A Quality Assessment Framework for Linked Data





Evaluation

Table I. Quality Evaluation Results for 270a Datasets

		Datasets								
		BFS	ECB	FAO	FRB	IMF	OECD	Transparency	UIS	World Bank
Availability	Estimated No Misreported Content Types (A4)	97.30%	91.67%	95.41%	96.11%	88.96%	96.82%	99.67%	88.96%	80.43%
	End Point Availability (A1)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	RDF Availability (A2)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Performance	Scalability of Data Source (P4)	99.01%	98.87%	99.00%	100.00%	98.89%	96.41%	97.38%	97.92%	97.67%
	Low Latency (P2)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	High Throughput (P3)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Correct URI Usage (P1)	99.96%	100.00%	100.00%	100.00%	100.00%	100.00%	0.12%	99.99%	100.00%
Licencing	Machine Readable License (L1)	25.00%	94.23%	78.57%	52.63%	62.50%	97.16%	25.00%	66.67%	42.86%
	Human Readable License (L2)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Interoperability	re-use of Existing Terms (IO1)	17.81%	11.01%	33.12%	19.18%	28.99%	3.93%	56.36%	32.54%	24.15%
	re-use of Existing Vocabularies (IO2)	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6.25%
Provenance	Provision for Basic Provenance Information	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Identifying the Origin of the Data	74.43%	48.82%	30.94%	79.09%	22.76%	73.15%	0.00%	67.16%	0.00%
Representational Conciseness	No Prolix RDF (RC2)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Short URIs (RC1)	99.95%	99.99%	99.99%	96.77%	99.99%	97.94%	95.53%	100.00%	38.17%
Versatility	Different Serialisations (V1)	33.33%	1.92%	7.69%	5.55%	14.28%	0.71%	25.00%	12.50%	14.28%
	Multiple Language Usage (V2)	4	1	3	1	1	2	1	2	1
Interpretability	Low Blank Node Usage (IN4)	80.38%	99.97%	99.98%	99.97%	99.98%	99.92%	89.55%	99.87%	99.50%
	Defined Classes and Properties (IN3)	90.91%	22.92%	82.18%	92.00%	59.12%	7.55%	90.32%	64.50%	34.16%
Consistency	Misplaced Classes or Properties (CS2)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Estimated No Entities As Members of Disjoint Classes (CS1)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Correct Usage of OWL Datatype Or Object Properties (CS3)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

9 Financial Statistical Linked Data datasets

27 Metrics

1 Billion Triples

Reflection

a) What the authors considered novel or the major contribution
   - Conceptual methodology and public implementation

b) What you yourself considered interesting
   - What is quality is different for everyone
   - Need to characterize what important for fitness for intended use

c) What the paper you think tells us about the state of the art in Semantic Web or Linked Data or Knowledge Graphs
   - Need for systematic data quality assessment of Linked Open Data datasets – continuous improvement

IMPORTANT NOTE

NO CS7IS1 SESSION TOMORROW (ONLY)

Tuesday 13th Sept at 10am

USE the time instead to

1. Progress SDT 1 and SDT 2 tasks
2. Explore what paper you want to book for Individual Research Task