Intro to Knowledge Graphs

Alina Petrova

Alina Petrova

alina.petrova@cs.ox.ac.uk

http://www.cs.ox.ac.uk/people/alina.petrova

Overview

- Graph data
- RDF graphs
  - Semantic Web
  - basic syntax
  - semantics
- Property Graphs
  - basic structure
  - computational advantages
- Use Cases
  - fantastic KGs and where to find them
  - industrial cases

Graph databases

- a basic unit is a structure:

- the first node is connected to the second node via a relation
- nodes represent entities: persons (*A_Einstein*), organizations (*Amazon*), chemical compounds (*hydrogen_peroxide*), files (*abstract.txt*), products (*Patagonia_wetsuit_B07QZJT6XL*) etc.
- … or even classes (*product*, *protein*, *customer* etc.)
- relations are usually directed (*follows*) rather than undirected (*friends_with*)

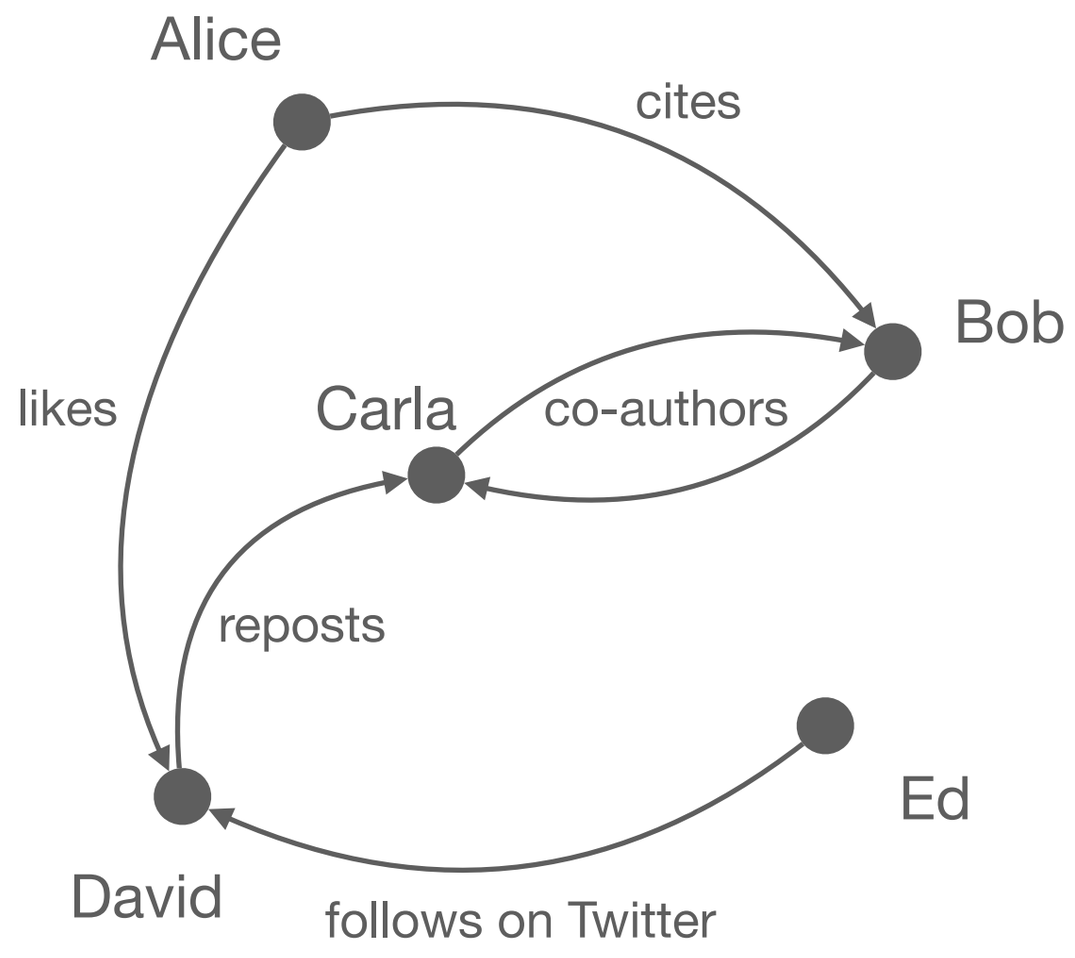
Graph data

Activity logs:

date	*# steps*	*kms*
1.01.21	4504	3.0
2.01.21	5309	4.5
3.01.21	10499	9
4.01.21	22300	17

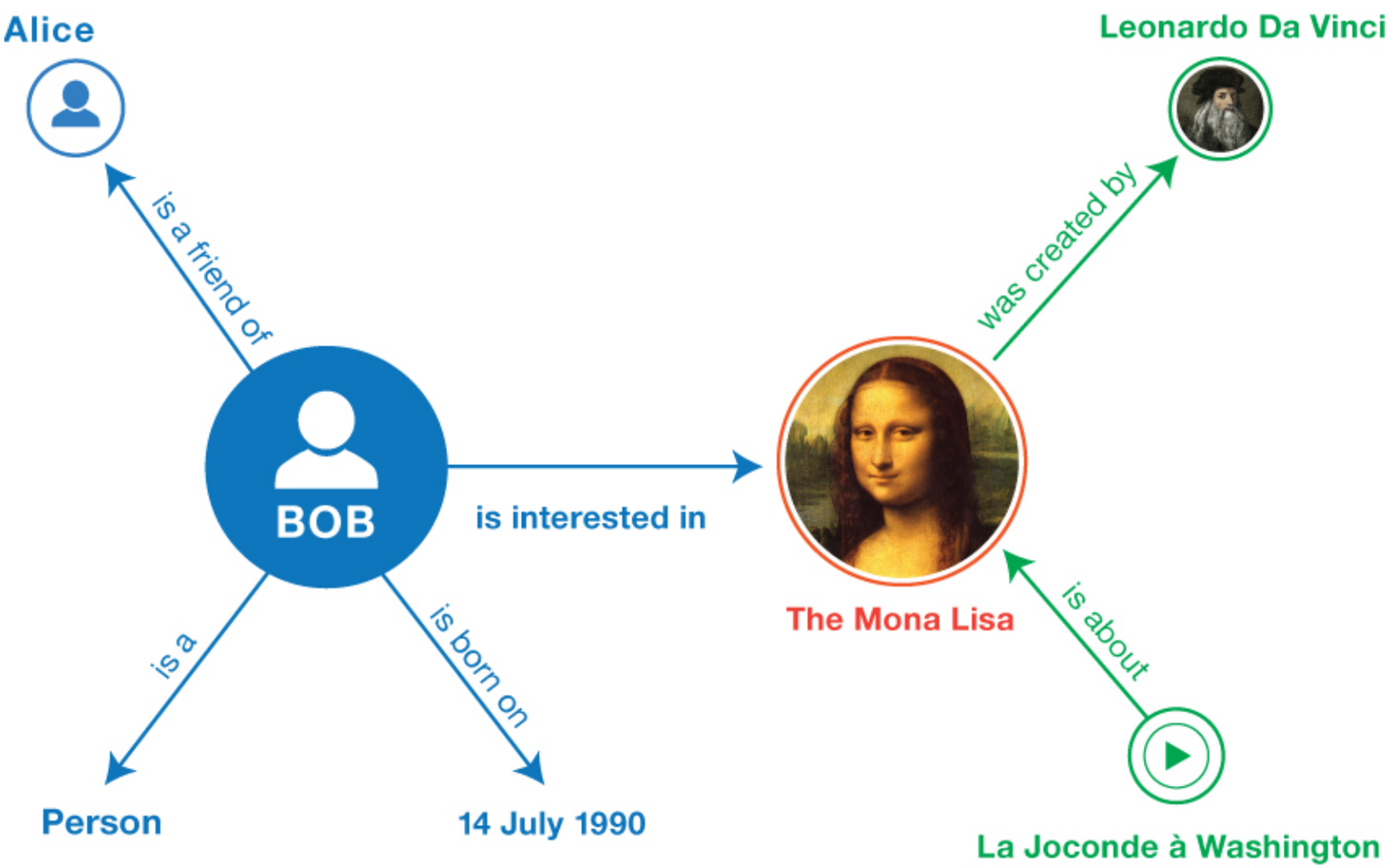
...

User interactions:



I. RDF Graphs

A versatile data model



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

Knowledge Graphs

- a data source in which entities are linked together via relations
- entities and relations have labels

- no real restrictions on the format of labels, nodes and edges
- **however**, there are 2 well-known and widely-used formats:
  - RDF graphs
  - Property Graphs

A note on Semantic Web

- Main idea: bring meaning to Web data (TBL)
  - all Web resources are represented in **uniform, unambiguous, machine-readable** format, and
  - their semantics is described by machine-readable statements.

- **URIs** (uniform resource identifier) identify everything online and offline: people, concepts, web pages (URLs are types of URIs) etc.

  T_Berners-Lee vs. www.cs.ox.ac.uk/people/tim.berners-lee

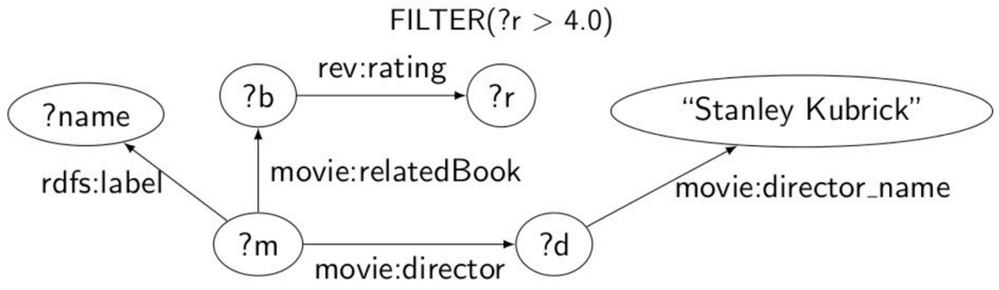
- **RDF** statements that use URIs describe those resources, link resources to each other, give context.

  (www.cs.../tim.berners-lee, works_at, en.wikipedia.org/University_of_Oxford)

RDF graphs

- come from the Semantic Web community
- have **well-defined syntax and semantics**, serialization formats (*turtle*, *rdf/xml*, *n3* etc.), native databases (*triplestores*) and query languages etc.
- **SPARQL**: SQL-like query language with formal syntax and semantics for RDF graphs
- the RDF data model and SPARQL are official W3C recommendations

```
SELECT ?name
WHERE {
   ?m rdfs:label ?name. ?m movie:director ?d.
   ?d movie:director_name "Stanley Kubrick".
   ?m movie:relatedBook ?b. ?b rev:rating ?r.
   FILTER(?r > 4.0)
}
```



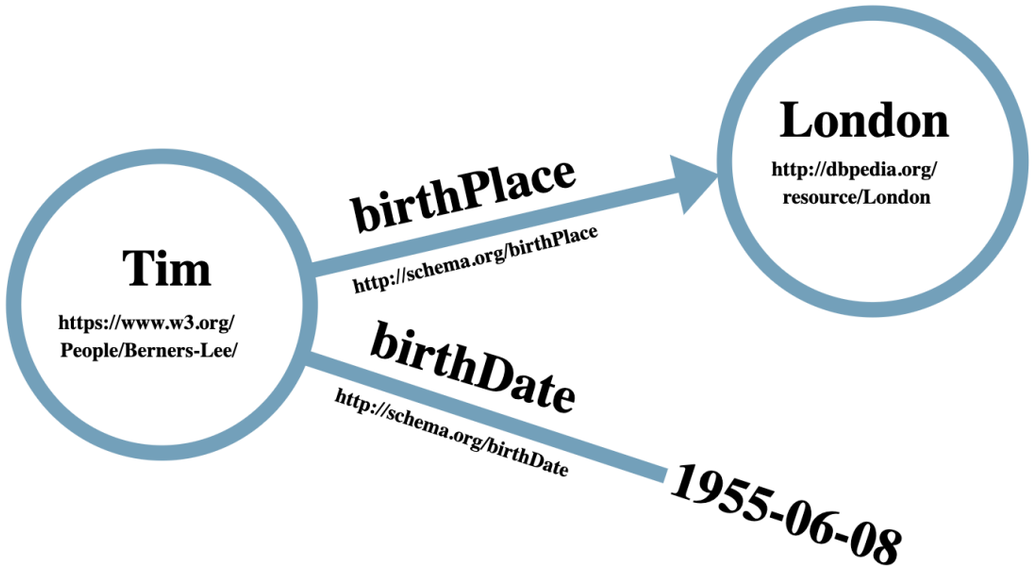
http://muratbuffalo.blogspot.com/2017/09/web-data-management-in-rdf-age.html

RDF data model

- RDF = Resource Description Framework
- data model for representing information about Web resources

- information is stored in **triples** — statements about 2 entities and a relation between them

- a triple is a tuple *(subject, predicate, object)*
- both entities and relations are **URIs**
- an object can also be a **literal**



https://ontola.io/what-is-linked-data/

(https://www.w3.org/People/Berners-Lee, https://schema.org/birthPlace, http://dbpedia.org/resource/London)
(https://www.w3.org/People/Berners-Lee, https://schema.org/birthDate, '1955-06-08')

Vocabularies

- RDF has no assumption on the form of URIs, except for they are strings,

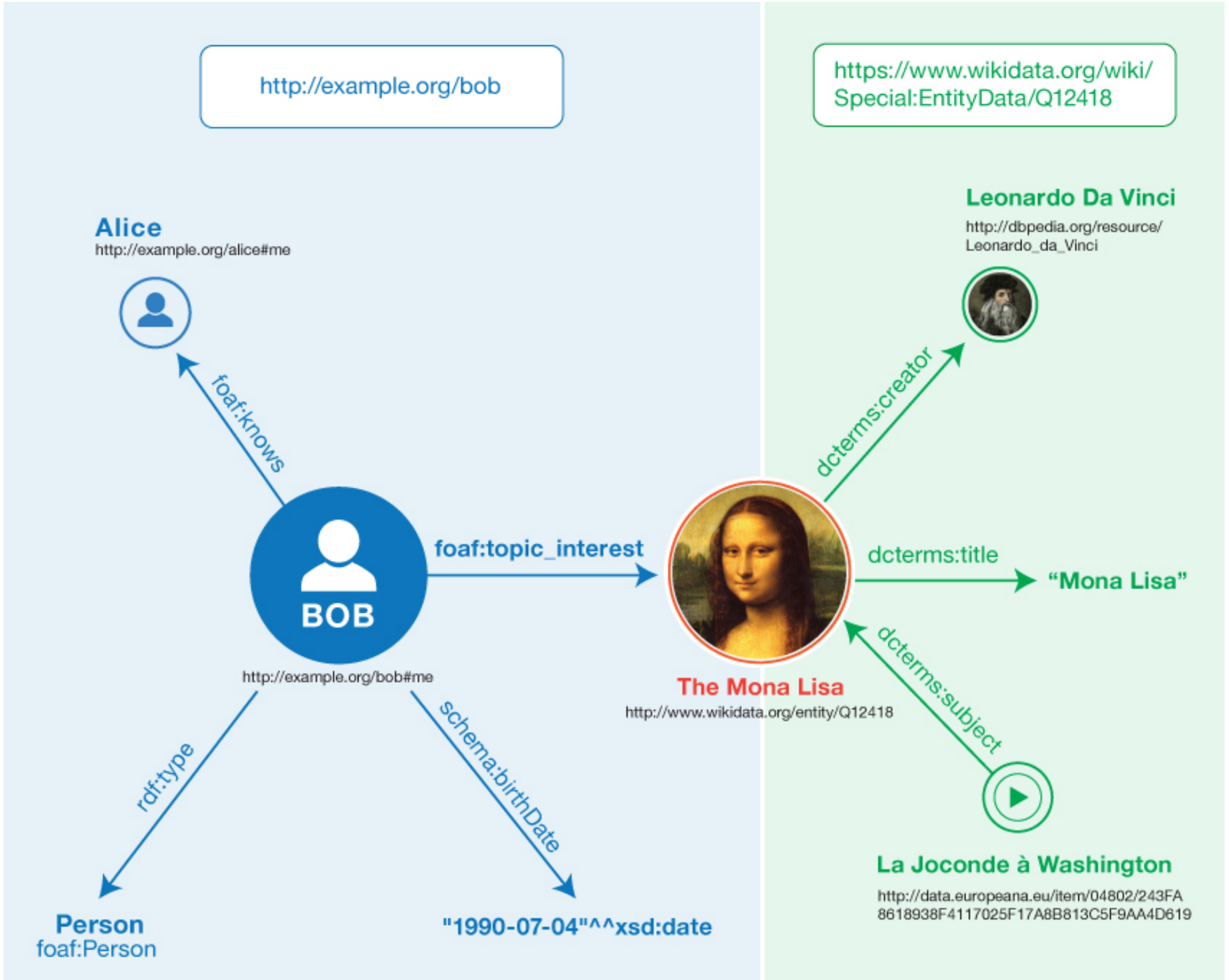
  T_Berners-Lee vs. www.cs.ox.ac.uk/people/tim.berners-lee

- but people do use **existing vocabularies**, usually a mix of them (reusability brings additional semantics, both to entities and relations!)

Most used vocabularies:

- **FOAF** ontology — for describing people
- **DBpedia** — structured information from Wikipedia
- **schema.org** — for describing web pages
- **RDF Schema** — to describe the data itself (https://www.w3.org/2000/01/rdf-schema#)

Vocabularies and reusability



https://www.w3.org/TR/rdf11-primer/#section-trig

Classes vs instances

- some entities represent not individual entities, but **classes**
- class structure is given by a **schema** (taxonomy, ontology, vocabulary)

Three types of triples:

- **subClassOf** — schema level

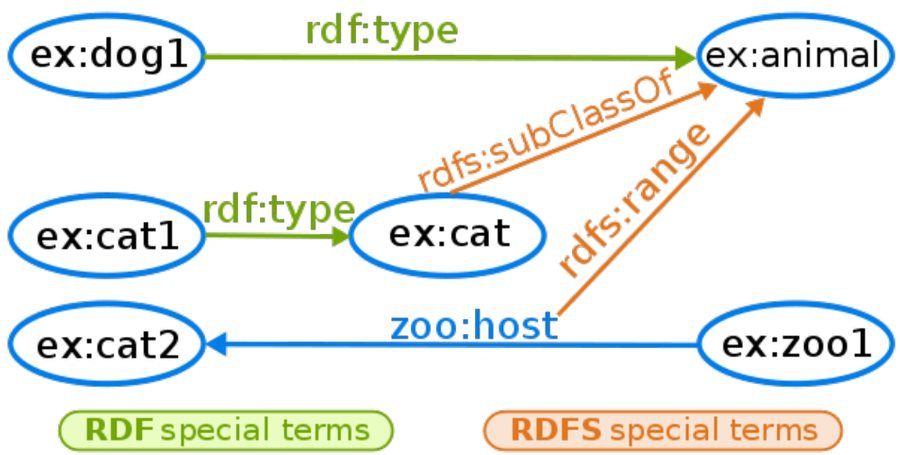
  (biologist, subClassOf, scientist)

- **typeOf** — link between instances and classes

  (Charles_Darwin, typeOf, scientist)

- other — instance level

  (Charles_Darwin, nationality, British)



https://en.wikipedia.org/wiki/RDF_Schema

Some syntactic sugar

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<#JW>
    a foaf:Person ;
    foaf:name "James Wales" ;
    foaf:mbox <mailto:jwales@bomis.com> ;
    foaf:homepage <http://www.jameswales.com> ;
    foaf:nick "Jimbo" ;
    foaf:depiction <http://www.jameswales.com/aus_img_small.jpg> ;
    foaf:interest <http://www.wikimedia.org> ;
    foaf:knows [
        a foaf:Person ;
        foaf:name "Angela Beesley"
    ] .

<http://www.wikimedia.org>
    rdfs:label "Wikimedia" .
```

https://en.wikipedia.org/wiki/FOAF_(ontology)

Problems…

- An RDF graph is a set of triples.
- It could be represented as a relational database with a single *Triples* table:

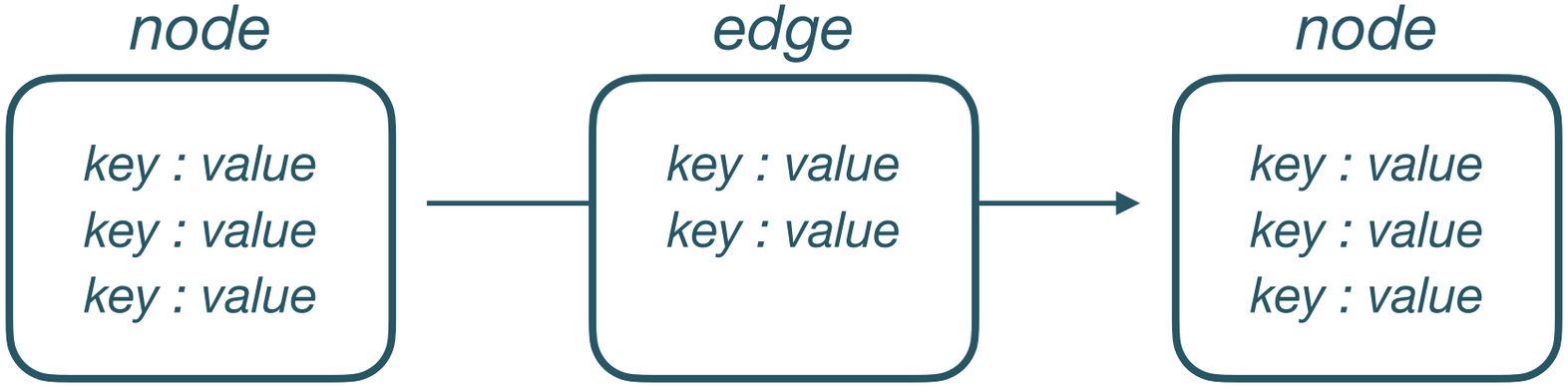
subject	*predicate*	*object*
James Wales	type of	person
James Wales	nickname	Jimbo
James Wales	works at	Wikimedia
Wikimedia	type of	encyclopedia

- SQL, and hence SPARQL, lacks native support for edges and paths.
- A path of *n* hops ~~ *n-1* **joins** of the *Triples* table with itself. It gets tricky after 3-4 hops even on small datasets…
- Obviously, modern triplestores use all sorts of optimizations, but they are still inherently **index-based**.

II. Property Graphs

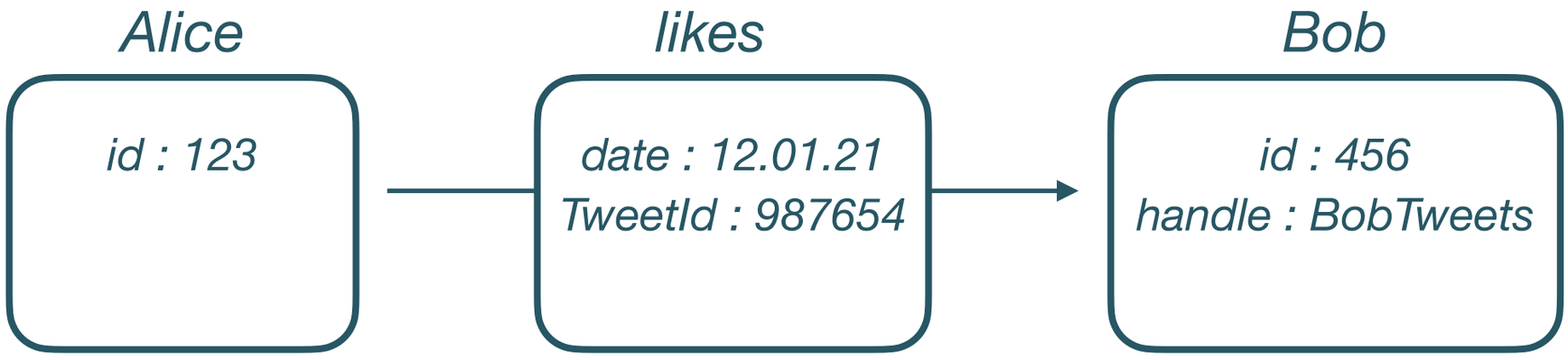
Property graphs

- property graphs (PGs) or labelled property graphs (LPGs)
- KGs in which entities and relations **have structure,** but they are **not formally defined** (*yet*)



- PGs are more generic than RDF graphs;
  they lack common semantics, but are native to graphs as they use pointers.

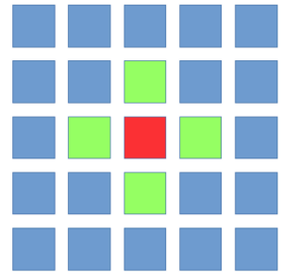
Property graphs



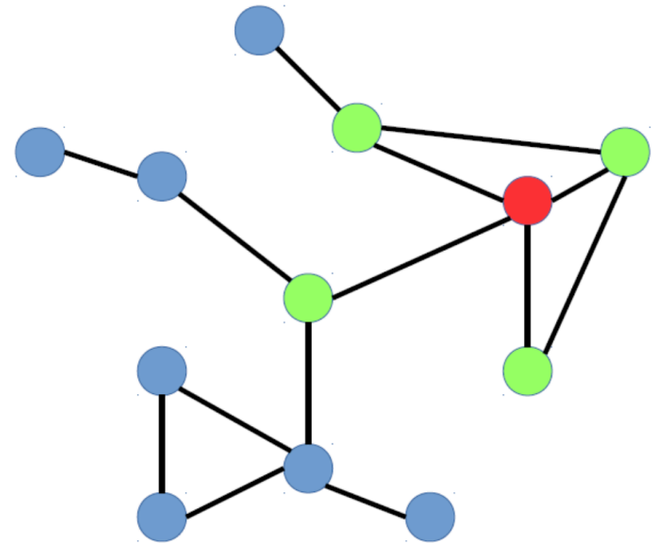
key : value pairs do not have formal semantics (*yet*), but they

- give context
- can qualify relations
- facilitate adding and modifying data (no schema)

Algorithmic advantages



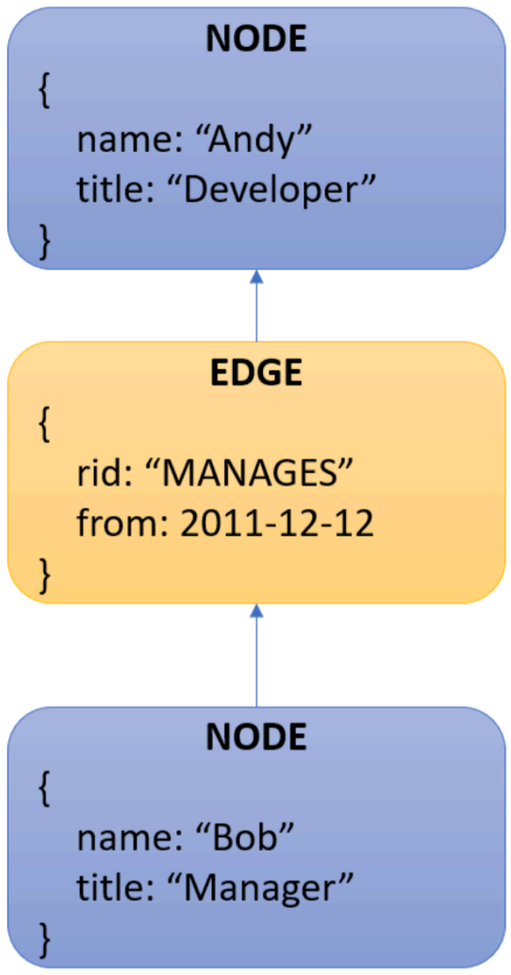
Connections in PGs are directly stored per node,
no scan through the data needed.
→ PGs support:



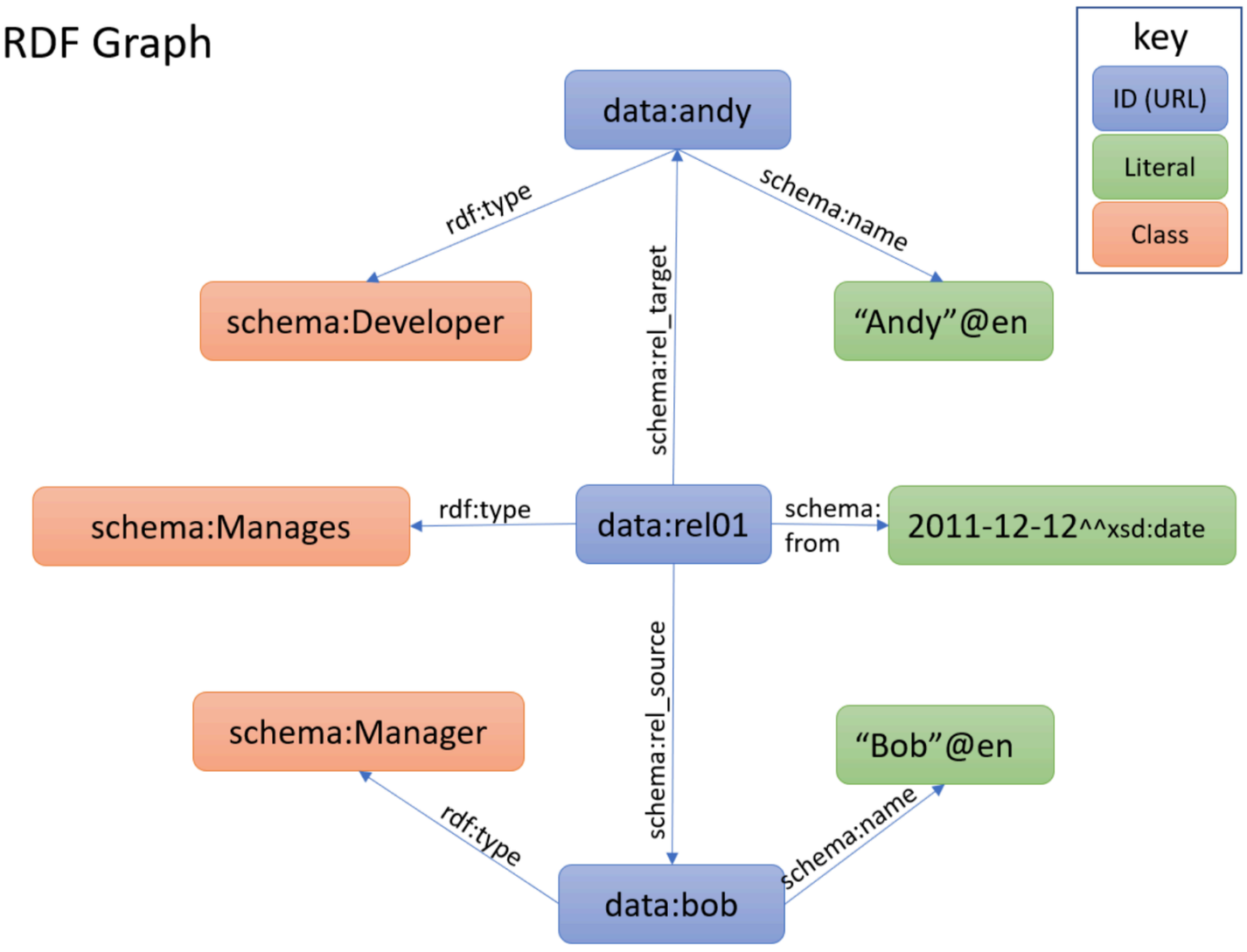
- path algorithms
- centrality algorithms
- similarity algorithms and neighborhood comparison etc.

RDF graphs vs. property graphs

Labelled Property Graph



RDF Graph



https://medium.com/terminusdb/graph-fundamentals-part-2-labelled-property-graphs-ba9a8edb5dfe

RDF graphs vs. property graphs

- Both types of graphs emphasize the interlinks between entities rather than volumes of data about each entity.
- **PGs** are best suited for path analysis and traversal; their original purpose was to query large volumes of graph data.
- **RDF graphs** are best suited for creating and analyzing complex graph patterns with rich formal semantics; their original purpose was data exchange.
- Implementations differ: RDF stores are index-based; PG stores are pointer-based.

Technological stacks

RDF graphs:

- RDF format
- *SPARQL, Graql, Vadalog* for querying
- triplestores for storing the data: *RDFox*, *GraphDB* (by OntoText), *Grakn*, *Stardog*, *Apache Fuseki* etc.

Property graphs:

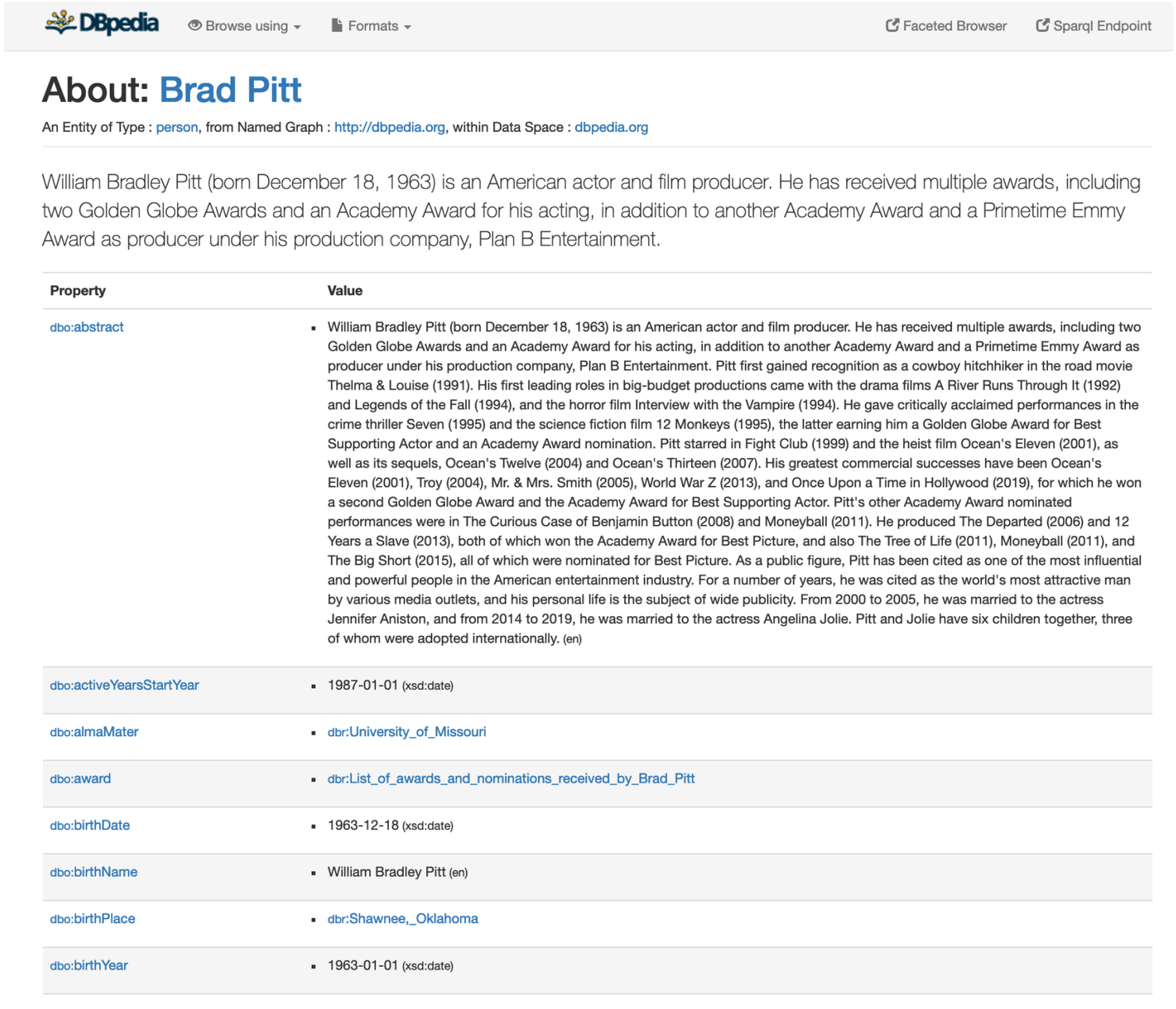
- likely json format
- *GraphQL, Cypher, Gremlin* for querying
- graph databases fot storing: *Neo4J*, *TigerGraph*

III. KGs and Use Cases

Famous KGs: general domain

- **DBpedia**
- YAGO
- OpenCYC
- Microsoft Concept Graph
- DiffBot

etc.



DBpedia Browse using Formats Faceted Browser Sparql Endpoint

About: Brad Pitt

An Entity of Type : person, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

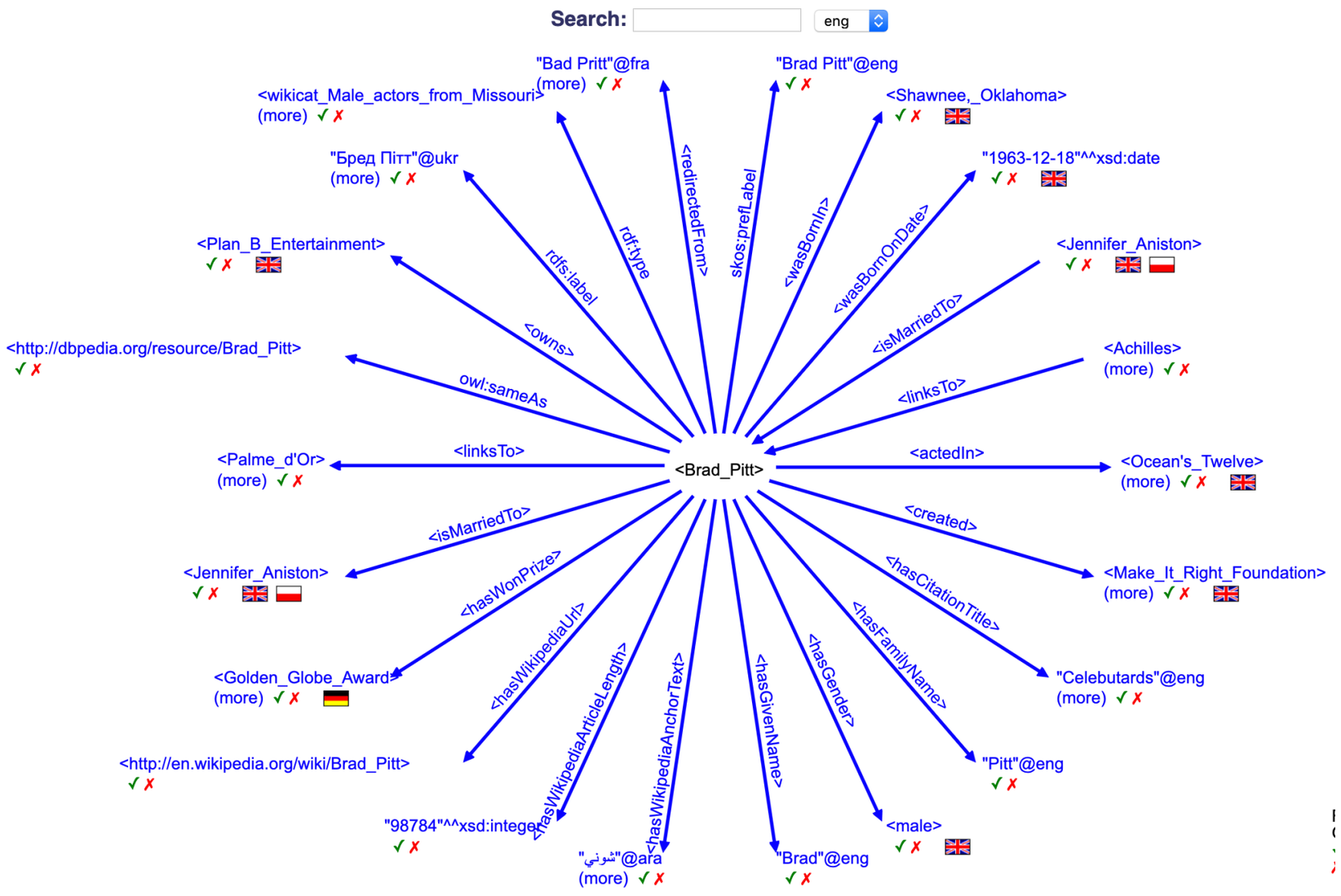
William Bradley Pitt (born December 18, 1963) is an American actor and film producer. He has received multiple awards, including two Golden Globe Awards and an Academy Award for his acting, in addition to another Academy Award and a Primetime Emmy Award as producer under his production company, Plan B Entertainment.

Property	Value
dbo:abstract	▪ William Bradley Pitt (born December 18, 1963) is an American actor and film producer. He has received multiple awards, including two Golden Globe Awards and an Academy Award for his acting, in addition to another Academy Award and a Primetime Emmy Award as producer under his production company, Plan B Entertainment. Pitt first gained recognition as a cowboy hitchhiker in the road movie Thelma & Louise (1991). His first leading roles in big-budget productions came with the drama films A River Runs Through It (1992) and Legends of the Fall (1994), and the horror film Interview with the Vampire (1994). He gave critically acclaimed performances in the crime thriller Seven (1995) and the science fiction film 12 Monkeys (1995), the latter earning him a Golden Globe Award for Best Supporting Actor and an Academy Award nomination. Pitt starred in Fight Club (1999) and the heist film Ocean's Eleven (2001), as well as its sequels, Ocean's Twelve (2004) and Ocean's Thirteen (2007). His greatest commercial successes have been Ocean's Eleven (2001), Troy (2004), Mr. & Mrs. Smith (2005), World War Z (2013), and Once Upon a Time in Hollywood (2019), for which he won a second Golden Globe Award and the Academy Award for Best Supporting Actor. Pitt's other Academy Award nominated performances were in The Curious Case of Benjamin Button (2008) and Moneyball (2011). He produced The Departed (2006) and 12 Years a Slave (2013), both of which won the Academy Award for Best Picture, and also The Tree of Life (2011), Moneyball (2011), and The Big Short (2015), all of which were nominated for Best Picture. As a public figure, Pitt has been cited as one of the most influential and powerful people in the American entertainment industry. For a number of years, he was cited as the world's most attractive man by various media outlets, and his personal life is the subject of wide publicity. From 2000 to 2005, he was married to the actress Jennifer Aniston, and from 2014 to 2019, he was married to the actress Angelina Jolie. Pitt and Jolie have six children together, three of whom were adopted internationally. (en)
dbo:activeYearsStartYear	▪ 1987-01-01 (xsd:date)
dbo:almaMater	▪ dbr:University_of_Missouri
dbo:award	▪ dbr:List_of_awards_and_nominations_received_by_Brad_Pitt
dbo:birthDate	▪ 1963-12-18 (xsd:date)
dbo:birthName	▪ William Bradley Pitt (en)
dbo:birthPlace	▪ dbr:Shawnee,_Oklahoma
dbo:birthYear	▪ 1963-01-01 (xsd:date)

Famous KGs: general domain

- DBpedia
- **YAGO**
- OpenCYC
- MS Concept Graph
- DiffBot

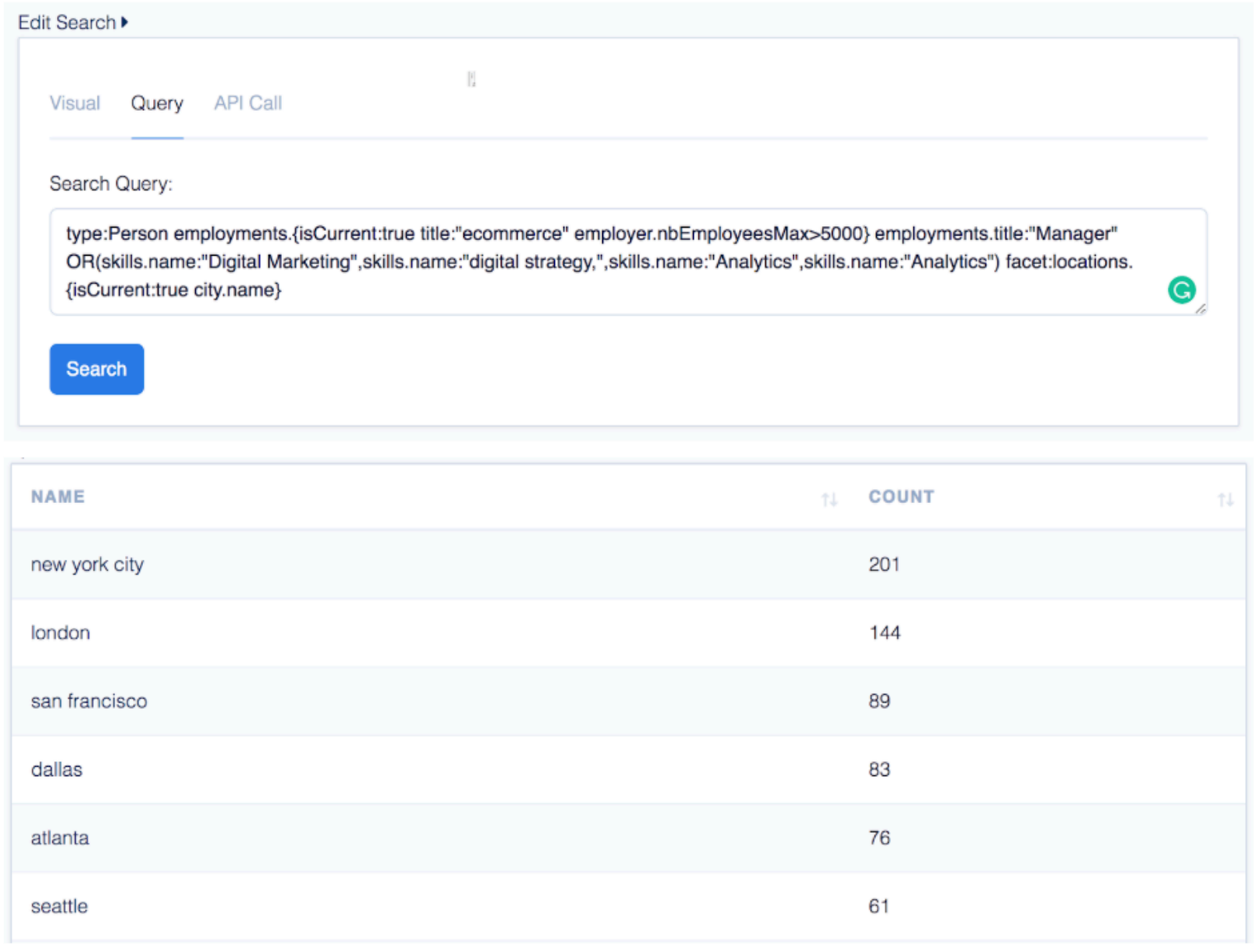
etc.



Famous KGs: general domain

- DBpedia
- YAGO
- OpenCYC
- MS Concept Graph
- **DiffBot**

etc.



Famous KGs: domain-dependent

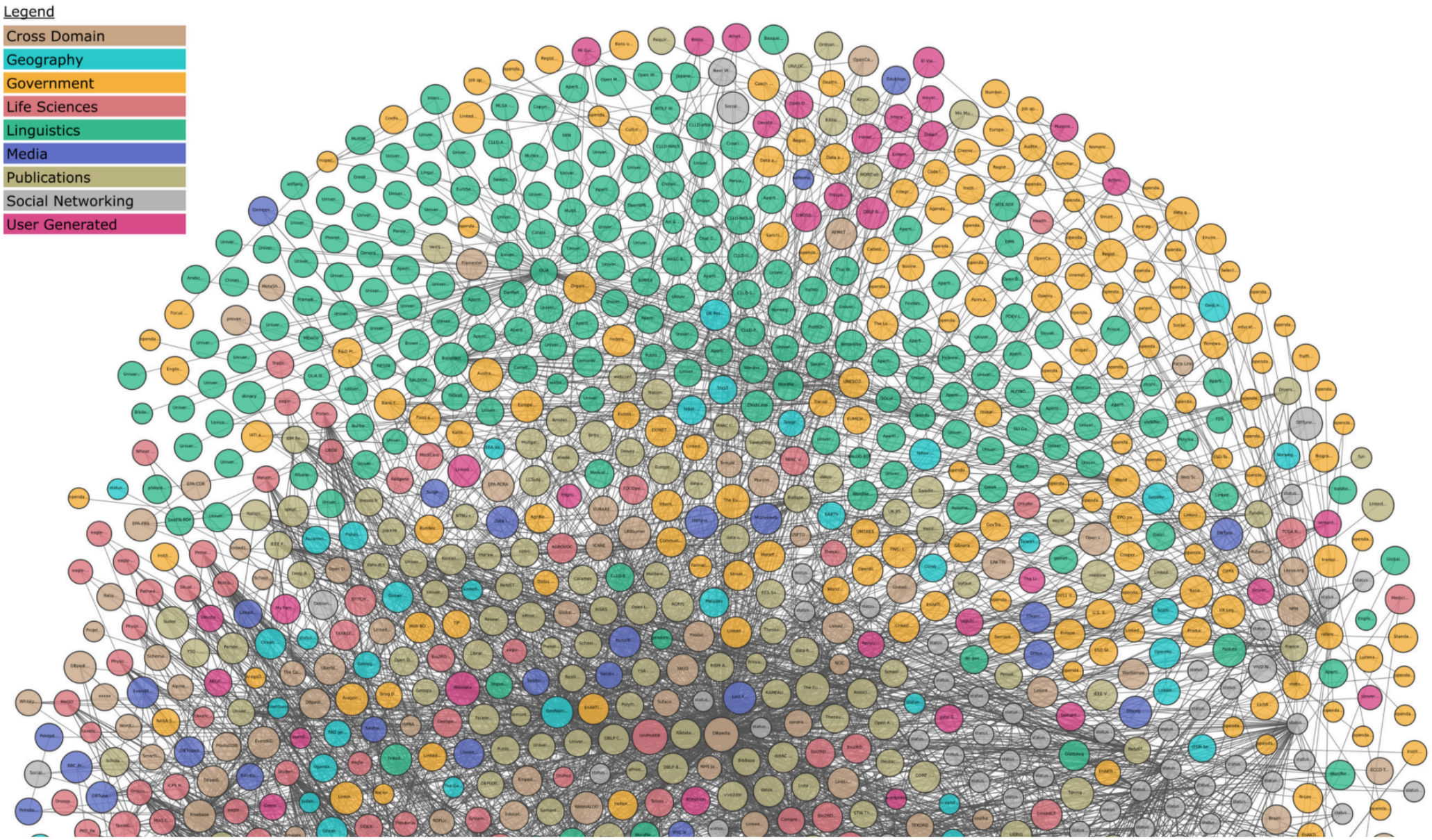
1. biomedical domain: *Bio2RDF*, *BioGrakn*, *Fraunhofer bio-KG*, an abundance of COVID19 KGs

2. scholarly KGs: *Microsoft Academic Knowledge Graph* and *DBLP*

3. enterprise/proprietary KGs: Google's *Knowledge Graph*, Amazon's *Product Graph*, Facebook's *Social Graph*, *Bing KG*, Bloomberg, Ebay, Uber, LinkedIn etc.

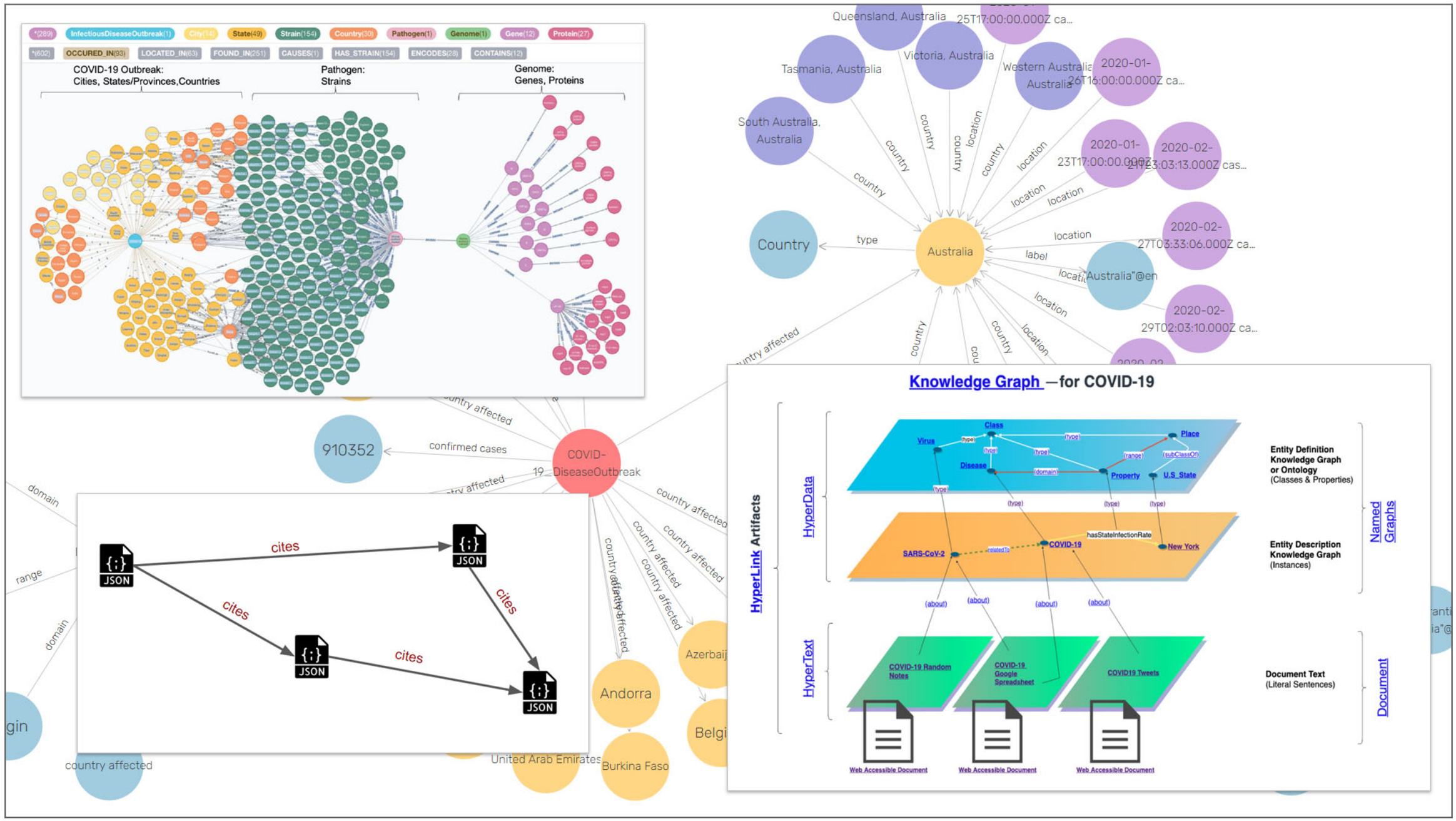
… and much, much more: https://lod-cloud.net/

The Linked Open Data cloud



https://lod-cloud.net

Knowledge graphs on COVID19



https://thegraphlounge.com/knowledge-graphs-in-the-fight-against-covid-19/

Cybersecurity KGs

Research
Cybersecurity—Article

A Practical Approach to Constructing a Knowledge Graph for Cybersecurity

Yan Jia, Yulu Qi, Huaijun Shang, Rong Jiang, Aiping Li *

School of Computer Science, National University of Defense Technology, Changsha 410073, China

ARTICLE INFO

Article history:
Received 10 December 2017
Revised 21 December 2017
Accepted 7 January 2018
Available online 9 February 2018

Keywords:
Cybersecurity
Knowledge graph
Knowledge deduction

ABSTRACT

Cyberattack forms are
are always challenging
fields. At present, it is
graph with cybersecur
cybersecurity knowled
we extract entities an
deduced by calculating
ognizer (NER) is also u
that the Stanford NER
ognizer in the cyberse

CSKB: A Cyber Security Knowle on Knowledge Grap

Kun Li, Huachun Zhou(✉), Zhe Tu, and

School of Electronic and Information En
Beijing Jiaotong University, Beijing 100
{kun_li,hchzhou,zhe_tu,bhfeng}@

Abstract. The access of massive terminal devices has brought new security risks to the existing Internet, so traditional cybersecurity data sets are difficult to reflect the modern and complex network attack environment. Therefore, how to realize the standardization and integration of cybersecurity data, so as to continuously store and update malicious traffic information under massively connected terminals, has become a critical issue to be solved urgently. Therefore, based on the knowledge graph, we built a standardized cybersecurity ontology, and introduced the implementation process of the cybersecurity knowledge base (CSKB) from five stages of knowledge acquisition, knowledge fusion/extraction, know-ledge storage, knowledge inference, and knowledge update, aiming at providing a reliable basis for real-time cybersecurity protection solutions. Experiments prove that the knowledge stored in CSKB can effectively realize the specification and integration of securi

Keywo
security

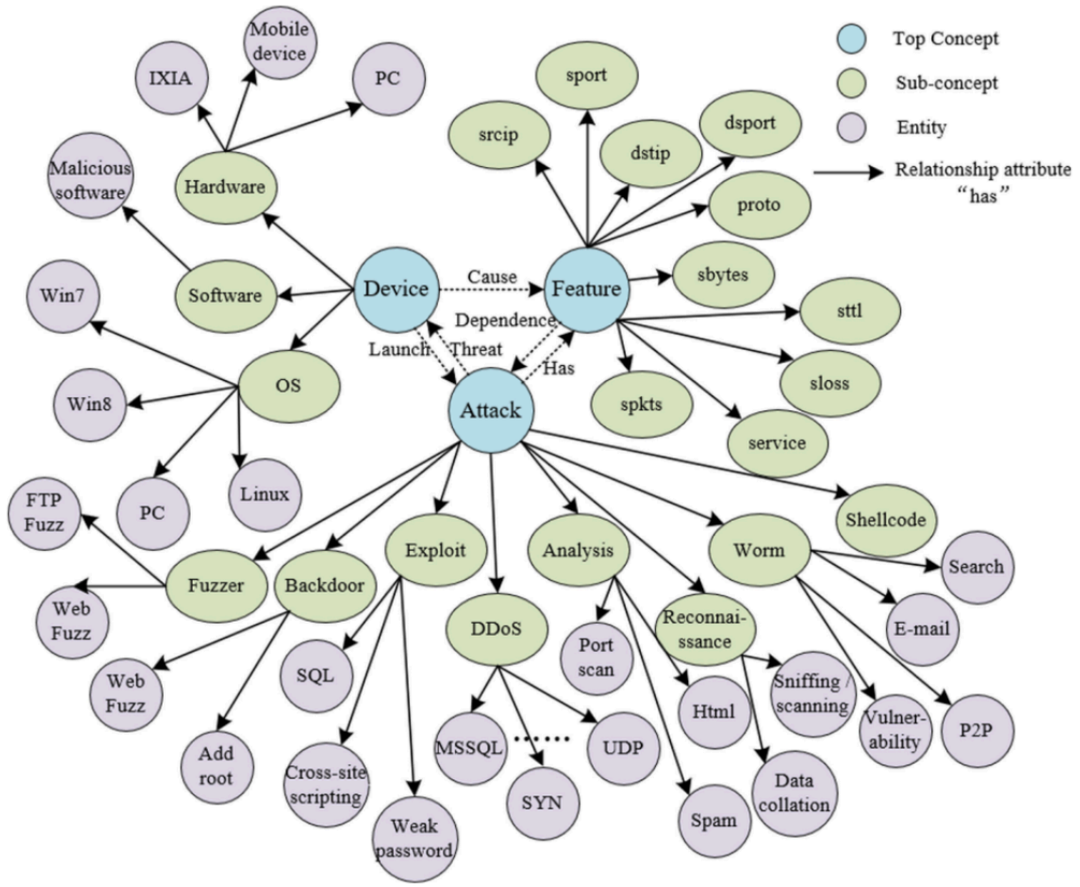


Fig. 2. Cybersecurity ontology

Developing an Ontology for Cyber Security Knowledge Graphs

Michael Iannacone
iannaconemd@ornl.gov
Oak Ridge National Laboratory

Shawn Bohn
shawn.bohn@pnnl.gov
Pacific Northwest National Laboratory

Grant Na
grant.nakam
Pacific North
Labo

John Gerth
gerth@graphics.stanford.edu
Stanford University

Kelly Huffer
testakm@ornl.gov
Oak Ridge National Laboratory

Robert Bridges
bridgesra@ornl.gov
Oak Ridge National Laboratory

Erik Ferragut
ferragutem@ornl.gov
Oak Ridge National Laboratory

John G
jgoodall@
Oak Ridg
Laboratory

ABSTRACT

In this paper we describe an ontology developed for a cyber security knowledge graph database. This is intended to provide an organized schema that incorporates information from a large variety of structured and unstructured data sources, and includes all relevant concepts within the domain. We compare the resulting ontology with previous efforts, discuss its strengths and limitations, and describe areas for future work.

and more economically important, the amou
information has been increasing rapidly, leadi
in managing and using this information. Th
some notable successes in creating structured
some domain entities (e.g. vulnerability datab
much domain information is only available i
Where structured data sources are available,
ever representation is convenient, without an
structure, contents, or names of entities. G
needed in the organization of this cyber sec

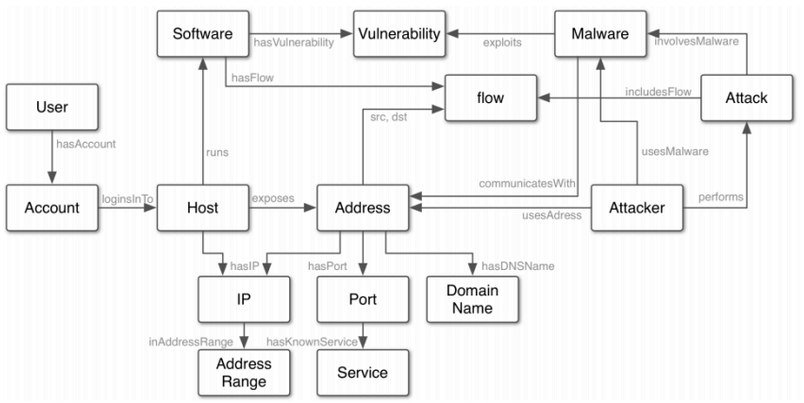


Figure 1: Entities and Relations in the STUCCO Ontology.

Review on the Application of Knowledge Graph in Cyber Security Assessment

Kai Zhang[1], Jingju Liu[1,*]

[1] College of Electromagnetic Countermeasure, National University of Defense Technology, Hefei, China

*Corresponding author e-mail: jingjul@aliyun.com

The development of artificial intelligence technology has advanced by leaps
nds and made significant progress in many areas. Many researchers have
apply artificial intelligence technology to the cyber security domain.
ge graphs can describe the concepts, entities and their relationships in the
world in a structured way. Applying knowledge graph to the cyber security
can organize, manage, and utilize massive amounts of information in
ce in a better way. In this paper, the common cyber security assessment
nd their shortcomings is summarized, the research progress of ontology-
owledge representation is discussed, thus leading to a conclusion that
-based knowledge representation can completely and accurately represent the
knowledge of heterogeneous systems in the cyber security domain. Then we
the concept of knowledge graph, summarize the application progress of
ge graphs in the cyber security domain, and discuss directions of future

KG construction

- manual (*OpenCYC*)
- crowdsourced (*DBpedia*, *Freebase*)
- (semi-)automatic (*DiffBot KG*):
  - curated vs unsupervised
  - from tables, infoboxes (*YAGO*), free text (*DiffBot KG*)
  - pattern-based vs machine learning-based

Use Cases

- EDA, discovering new patterns and dependencies
- link prediction
- NLP: providing common-sense knowledge, context to almost any task
- page ranking, logistics, semantically-rich social media analysis, customer analysis, fraud detection, recommendation

- **China Mobile**: data on 900 mio mobile phones; detects fraudulent activity using graph-based features
- **Amazon**: KG(s?) used in product recommendation, Alexa's question answering, in Amazon Music and Prime Video
- **Grakn**: drug discovery, precision medicine
- **'Panama Papers'**: tax and money fraud detected by journalists in 11.5 bio emails from Mossack Fonseca

Takeaways

- **Knowledge graphs** are a versatile data model for domains in which relations between entities play a major role.
- Originally KGs **stemmed from Semantic Web**; modern KGs are a marriage of SW and graph databases.
- There is a huge ‘cloud’ of **open-source KGs** on almost any topic.
- Most of them are **in RDF** (or RDF-compatible) format, although **property graphs** are on the rise. Enterprise KGs often use their own, proprietary formats.
- There is an increasing number of **native tools for KG** storage, efficient querying and analytics.
- KGs are used across industries in a variety of **applications**, from EDA to recommendations, to anomaly detection, to explainable predictions.

Images Used

- https://unsplash.com/photos/ZiQkhI7417A
- https://www.w3.org/TR/rdf11-primer/
- https://twobithistory.org/2018/05/27/semantic-web.html
- http://muratbuffalo.blogspot.com/2017/09/web-data-management-in-rdf-age.html
- https://ontola.io/what-is-linked-data/
- https://en.wikipedia.org/wiki/RDF_Schema
- https://en.wikipedia.org/wiki/FOAF_(ontology)
- https://cvw.cac.cornell.edu/mpiadvtopics/neighborhood
- https://medium.com/terminusdb/graph-fundamentals-part-2-labelled-property-graphs-ba9a8edb5dfe
- https://blog.diffbot.com/turn-existing-customer-data-into-fresh-marketing-opportunities-with-knowledge-graph/
- https://lod-cloud.net
- https://thegraphlounge.com/knowledge-graphs-in-the-fight-against-covid-19/
- https://rogersmovienation.com/2019/10/19/netflixable-the-laundromat-takes-the-cute-approach-in-explaining-the-panama-papers-scandal/