GRAPH-BASED KNOWLEDGE MODELS

JELENA JOVANOVIĆ

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

OVERVIEW

- Graphs and semantic networks for knowledge representation

- Data and knowledge graphs in the business domain

- Open data and knowledge graphs

- Gigantic Global Graph

  - Vision of the Web as a gigantic global graph (of data and knowledge)

  - Creation of gigantic knowledge bases through automated data collection from the Web
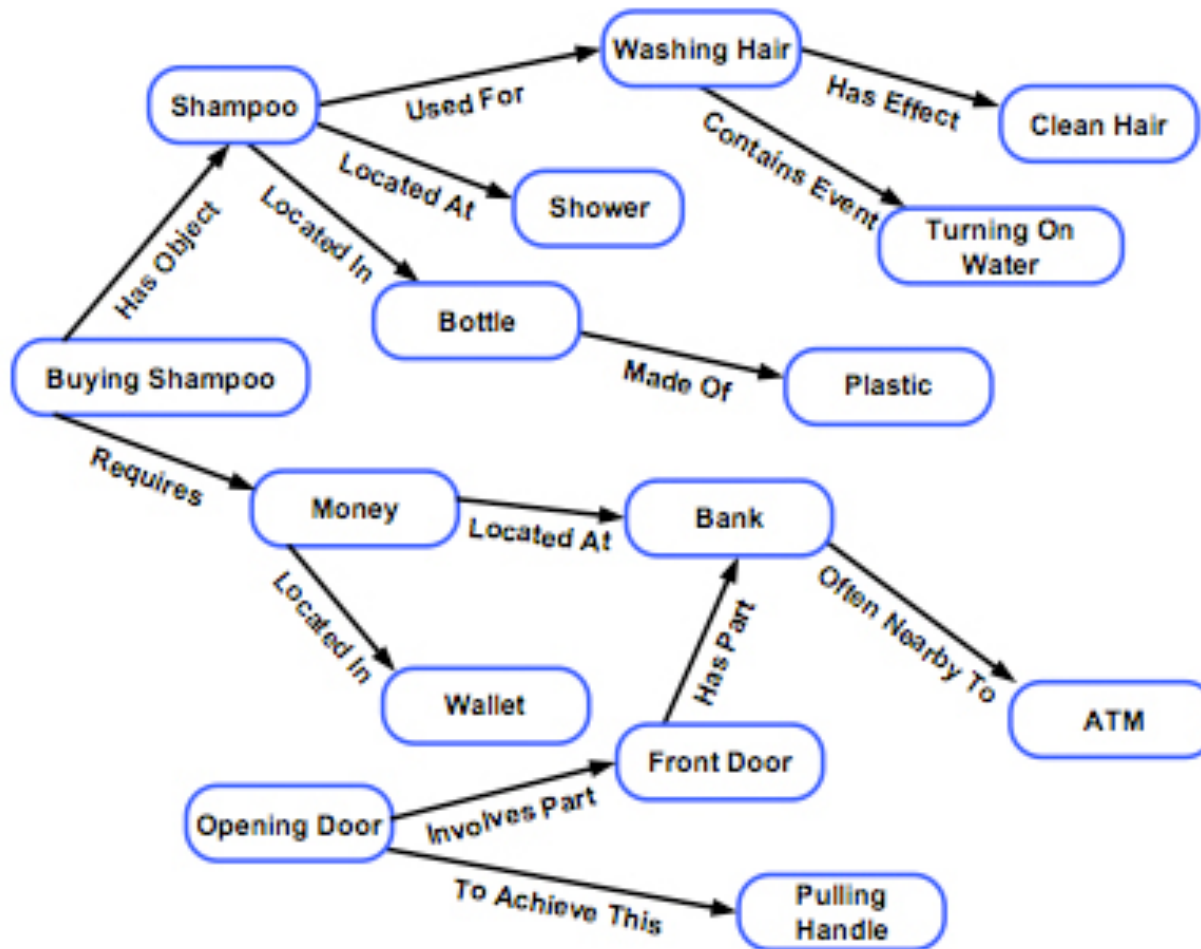
GRAPHS AND SEMANTIC NETWORKS FOR KNOWLEDGE REPRESENTATION

AN EXAMPLE OF A GRAPH OF ENTITIES AND THEIR MUTUAL RELATIONS
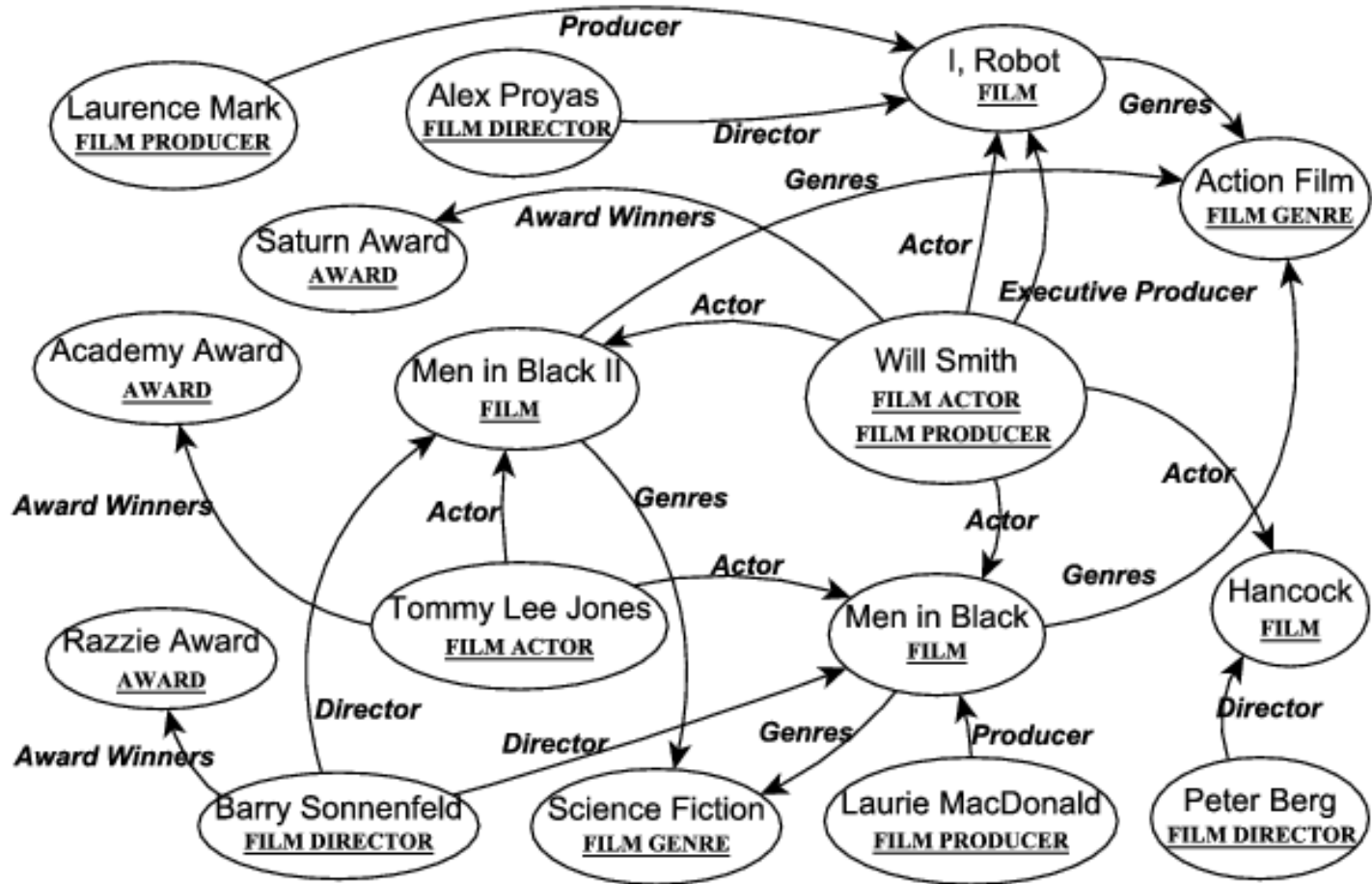
The example illustrates a tiny segment of the Freebase knowledge base

AN EXAMPLE OF A GRAPH OF COMMONSENSE KNOWLEDGE

The example illustrates a small segment of the ConceptNet knowledge base

1) META-KNOWLEDGE (CLASSES/CONCEPTS)

2) DOMAIN SPECIFIC KNOWLEDGE (DOMAIN ENTITIES)

Image source:
http://research.microsoft.com/en-us/groups/osa/krr.aspx
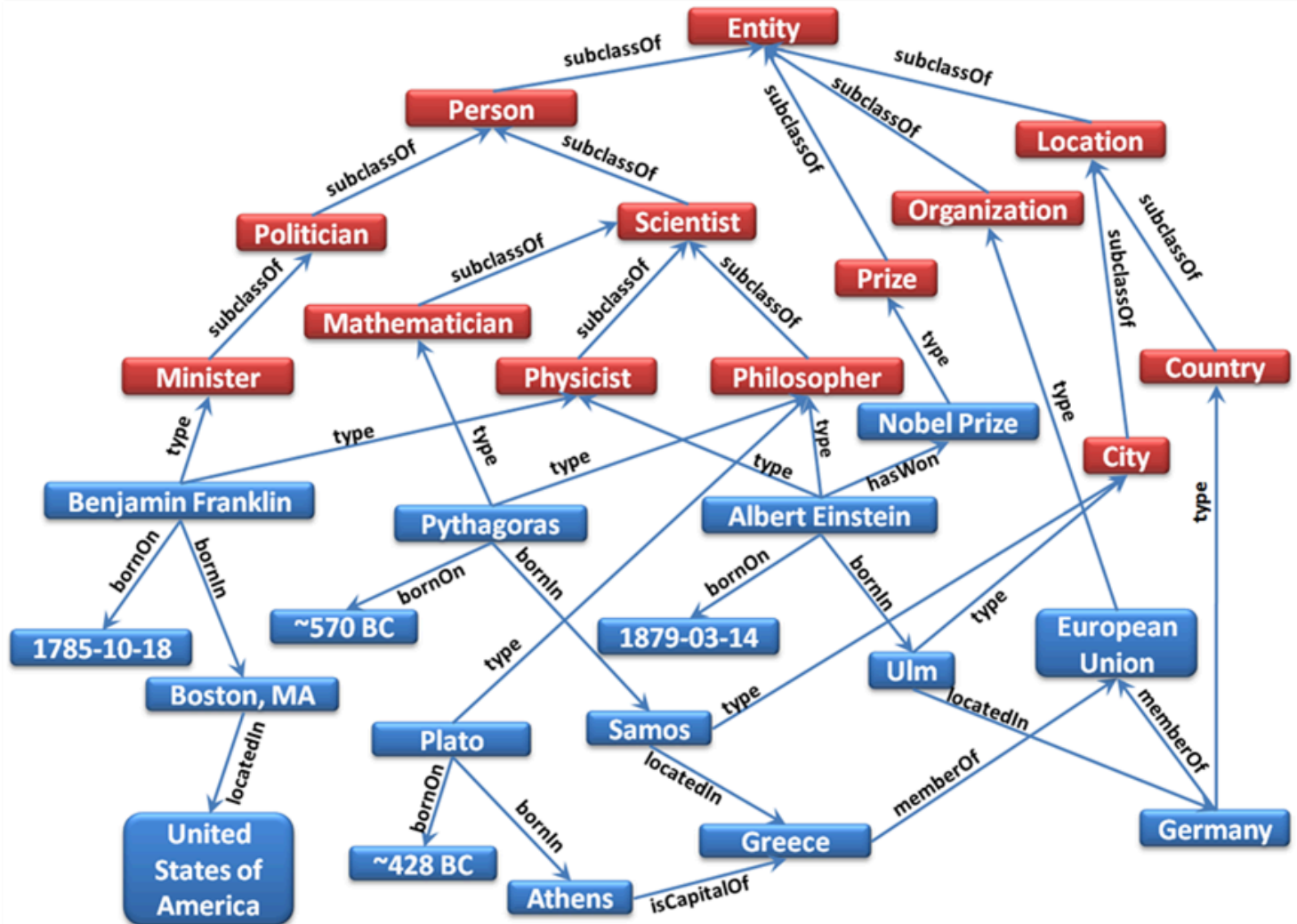
DATA AND KNOWLEDGE GRAPHS IN THE BUSINESS WORLD

GOOGLE'S KNOWLEDGE GRAPH

GOOGLE'S KNOWLEDGE GRAPH

"*…Google users will able to browse through the company's* **'knowledge graph**,*' or its ever-expanding* **database** *of information* **about 'entities'** *– people, places and things –* **the 'attributes' of those entities** *and* **how** *different* **entities are connected** *to one another*."

What Google's Search Changes Might Mean for You

Wall Street Journal, March 14, 2012

GOOGLE'S KNOWLEDGE GRAPH

"…Every piece of information that we crawl, index, or search is analyzed in the context of Knowledge Graph."

"…Almost all the structured data from all of our products like Maps and Finance and Movies and Music are all in the Knowledge Graph, so we can reasonably say that everything we know about is in this canonical form."

How a Database of the World's Knowledge Shapes Google's Future

MIT Technology Review, January 27, 2014

GOOGLE'S KNOWLEDGE GRAPH

"[Google Now] works by using machine learning algorithms to determine what you're doing, then matches this understanding with information stored in what the company calls the Google Knowledge Graph—a database of semantic data describing more than 1 billion people, places, and things. 'To be able assist you,' says Aparna Chennapragada, who oversees Google Now, 'we have to understand the world.' "

Startup Unleashes Its Clone of Google's 'Knowledge Graph'

Wired, April 6, 2014

FACEBOOK'S ENTITY GRAPH

"Facebook is building a rich stock of knowledge that could make its software smarter and boost the usefulness of its search engine…

*…Entities such as colleges and employers are **learned from data** typed **in profile pages**; businesses, movies, fictional characters, and other concepts are **learned from fan pag**es created by Facebook users. … **analyzing** many **employment histories** on the site allows Facebook's search engine to know that a search for "software engineers" should also return people who say they are "coders."*

Facebook Nudges Users to Catalog the Real World

MIT Technology Review, February 27, 2013

BING'S KNOWLEDGE AND ACTION GRAPH

"*Bing has over a billion **entities** (people, places, and things) and the number is growing every day. For those entities, we have over 21 billion associated **facts**, 18 billion links to **key actions** and over 5 billion **relationships between entities**.*

Millions of Bing users around the globe use this rich information every day, in bing.com, Cortana, Xbox, Office and more"

"*knowledge and action graph will be available to developers via a new API*"

Bing announces availability of the knowledge and action graph API
Bing Blogs, 20 August 2015

MICROSOFT'S CONCEPT GRAPH

*"The **Microsoft Concept Graph** is a **massive graph of concepts** – more than 5.4 million and growing – that machine-learning algorithms are culling from billions of web pages and years' worth of anonymized search queries."*

"The technology has potential applications that range from keyword advertising and search enhancement to the development of human-like chatbots."

[Microsoft researchers release graph that helps machines conceptualize](#)
Microsoft official blog, Nov 1, 2016

YAHOO'S KNOWLEDGE GRAPH

*"Spark [a semantic search assistance tool] takes a large **entity graph** as input, … consisting of the most important entities, their most important related entities, and their respective types. This entity graph is drawn from a larger Yahoo! Knowledge Graph, a unified **knowledge base** that provides **key information** about all the **entities** we care about, and **how they relate** to each other."*

Entity Recommendations in Web Search
The 12th International Semantic Web Conference, Oct 2013

LINKEDIN:
PROFESSIONAL GRAPH => ECONOMIC GRAPH

"[Economic graph is] *a digital mapping of the global economy, comprised of a profile for every professional, company, job opportunity, the skills required to obtain those opportunities, every higher education organization, and all the professionally relevant knowledge associated with each of these entities*"

"*With these elements in place, we can connect talent with opportunity at massive scale*"

Announcing The LinkedIn Economic Graph Challenge

Oct 14, 2014

OPEN DATA AND KNOWLEDGE GRAPHS

DBPEDIA

Machine interpretable version of Wikipedia

Data collected from Wikipedia are:

- *Structured*: represented in the form {subject-predicate-object} triplets suitable for further processing

- *Semantically annotated*: semantics of each triplet element is explicitly defined => it can be directly interpreted by a computer program

DBPEDIA

Knowledge representation in DBpedia
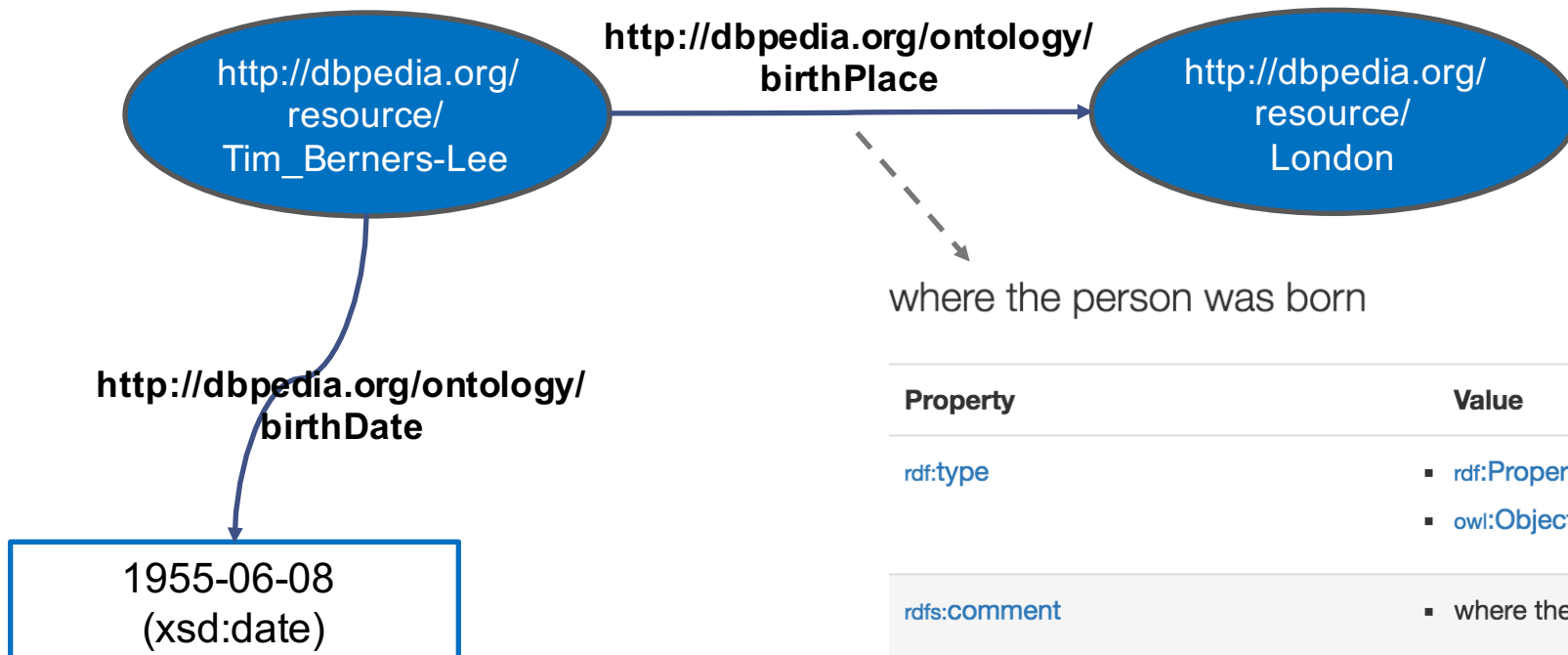
Property	Value
rdf:type	• rdf:Property • owl:ObjectProperty
rdfs:comment	• where the person was born (en)
rdfs:domain	• dbo:Person
rdfs:label	• birth place (en)
rdfs:range	• dbo:Place

DBPEDIA

http://en.wikipedia.org/wiki/San_Francisco

Country	🇺🇸 United States
State	California
Founded	June 29, 1776
Incorporated	April 15, 1850[9]
Founded by	José Joaquin Moraga Francisco Palóu
Named for	St. Francis of Assisi

Government
· Type	Mayor-council
· Body	Board of Supervisors
· Mayor of San Francisco	Ed Lee
· Board of Supervisors	**Supervisors** [show]
· California State Assembly[10][11]	Tom Ammiano (D) Phil Ting (D)
· CA State Senate[12]	Mark Leno (D)
· United States House of Representatives[13][14]	Nancy Pelosi (D) Barbara Lee (D) Jackie Speier (D)

Area[15]
| · Consolidated city-county | 231.89 sq mi (600.6 km²) |
| · Land | 46.87 sq mi (121.4 km²) |

```
<http://dbpedia.org/resource/San_Francisco>
   db:country dbpedia:United_States ;
   ...
   db:foundingDate "1776-6-29"^^xsd:date ;
   dbpprop:namedFor
       dbpedia:Francis_of_Assisi ;
   db:governmentType
       dbpedia:Mayor-council_government ;
   ...
```

WIKIDATA

"***Wikipedia's data*** *is buried in 30 million Wikipedia articles in 287 languages from which* **extraction is** *inherently* **very difficult**."

"*Population numbers for Rome, for example, can be found in English and Italian articles about Rome but also in the English article "Cities in Italy." The* **numbers are all different**."

Main objectives of the WikiData project:

- Turn Wikipedia data into a machine interpretable format, suitable for direct processing
- Sustain the accuracy and 'freshness' of the data

Wikidata: A Free Collaborative Knowledgebase.
D. Vrandečić, M. Krötzsch. Communications of the ACM 57:10, 2014

WIKIDATA

"*Wikidata is a project of the Wikimedia Foundation: a **free**, **collaborative**, **multilingual**, **secondary database**, collecting **structured data** to provide support for Wikipedia, Wikimedia Commons, the other Wikimedia projects, and well beyond that*"

"*A secondary database: Wikidata can record not just statements, but also their sources, thus reflecting the **diversity of knowledge** available and supporting the notion of **verifiability**"*
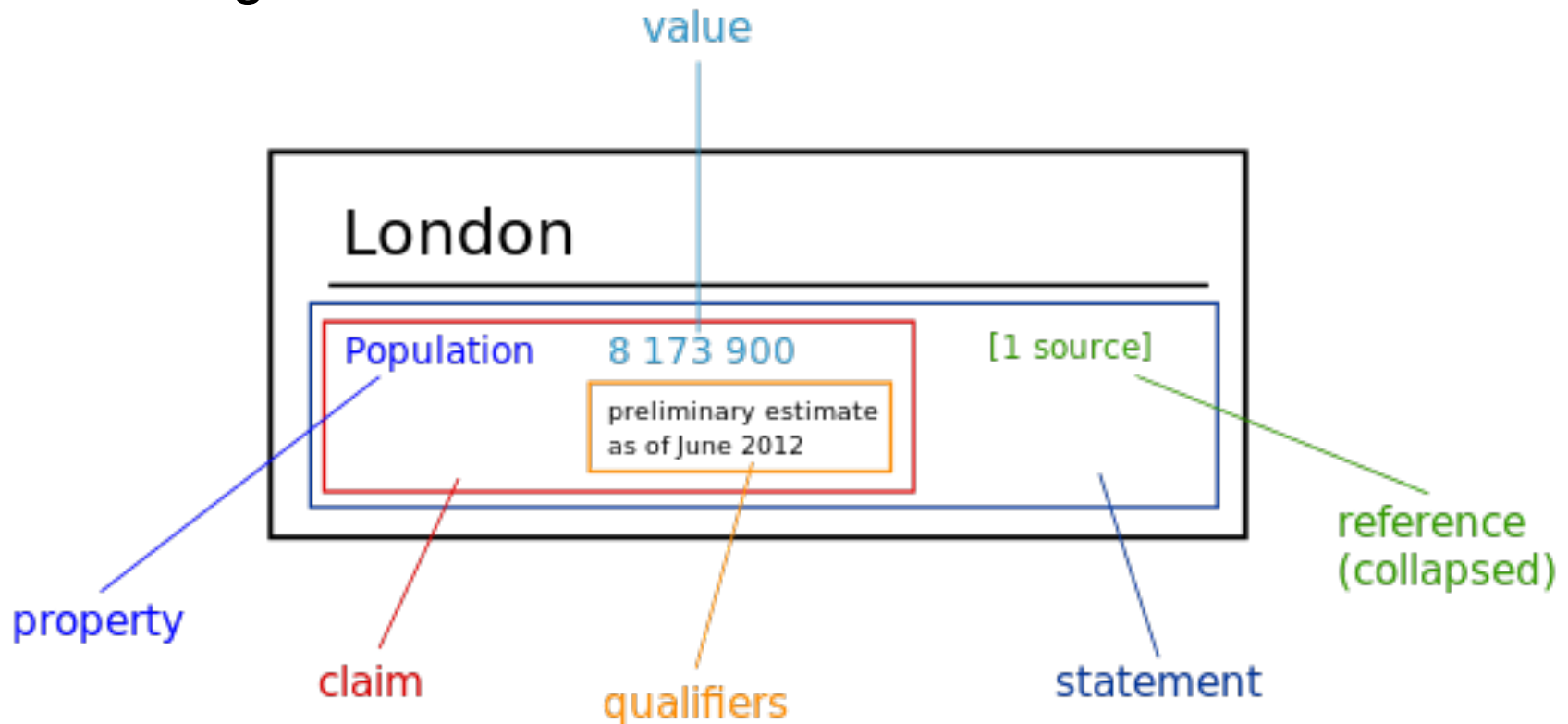
"*Collecting structured data: [to] allow easy **reuse of** that **data** by Wikimedia projects and third parties, and enable computers to easily process and "understand" it.*"

Wikidata:Introduction
August 2015

WIKIDATA

Data and knowledge representation in WikiData knowledge base

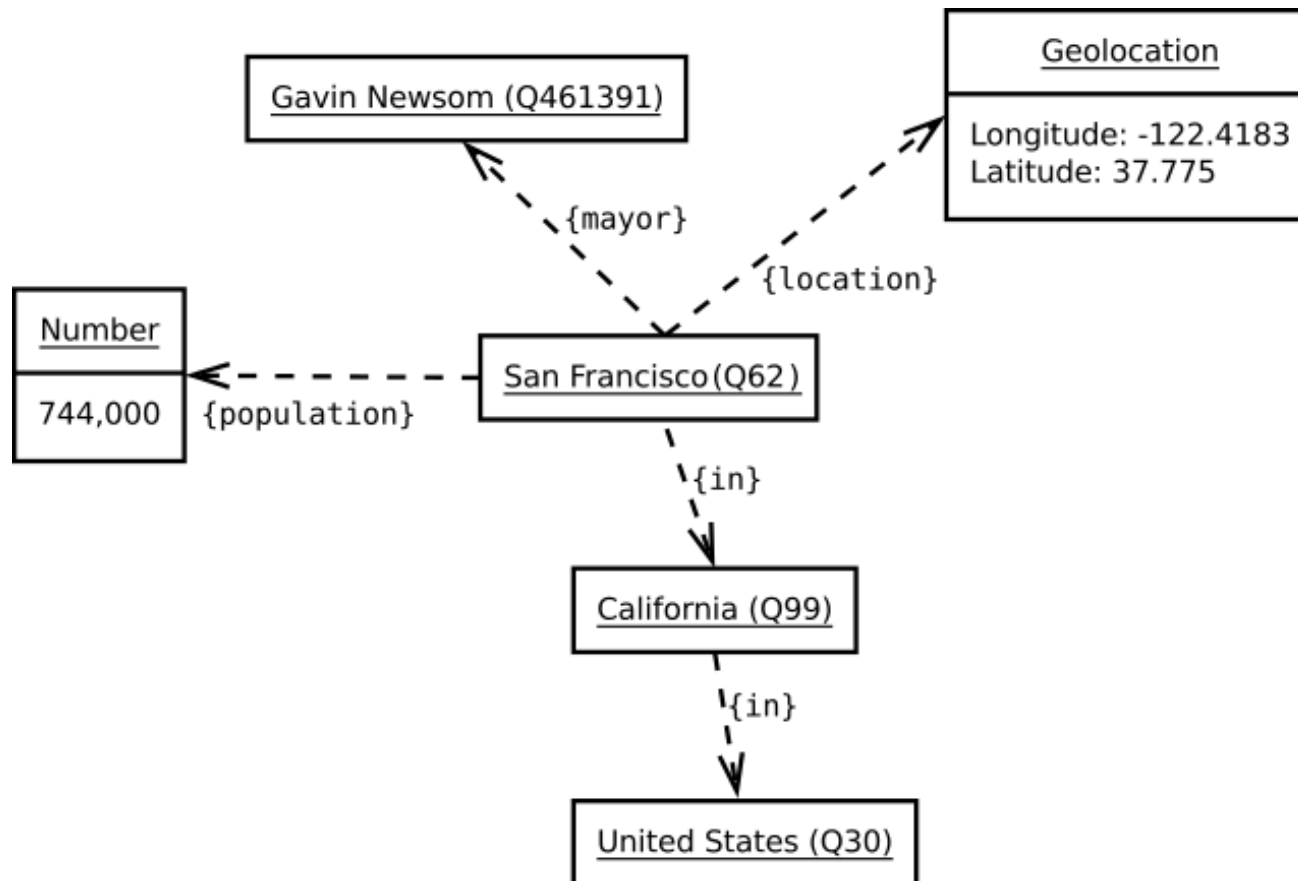# WIKIDATA

Data and knowledge representation in WikiData knowledge base

GIGANTIC GLOBAL GRAPH

GIGANTIC GLOBAL GRAPH (1):

VISION OF THE WEB AS A

GIGANTIC GLOBAL REPOSITORY (GRAPH)

OF DATA AND KNOWLEDGE

BY SIR TIM BERNERS-LEE

GIGANTIC GLOBAL GRAPH (1)

International Information Infrastructure (III)

- network/graph of computers known as *Internet* or *Net*
- "*It isn't the cables, it is the computers which are interesting*"

World Wide Web (WWW)

- network/graph of documents known as *Web*
- "*It isn't the computers, but the documents which are interesting*"

Gigantic Global Graph (GGG)

- network/graph of entities (resources) and data that describe the entities
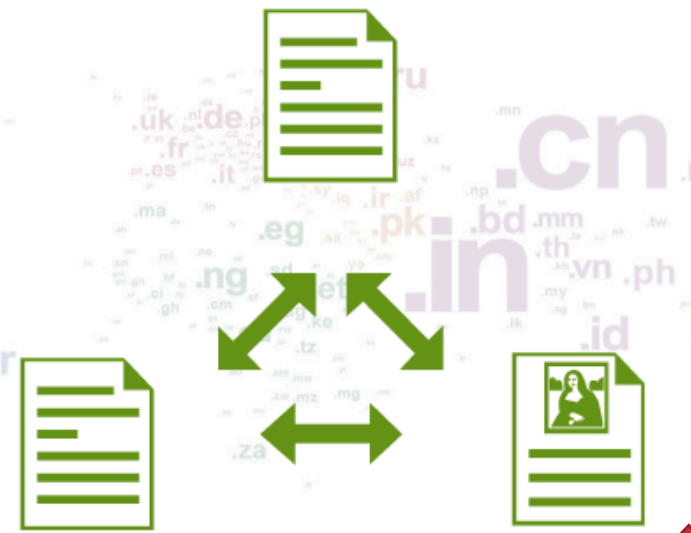- "*It's not the documents, it is the things they are about which are important*"

GIGANTIC GLOBAL GRAPH (1)

"…when I book a flight it is the flight that interests me. Not the flight page on the travel site, or the flight page on the airline site, but the URI (issued by the airlines) of the flight itself. …

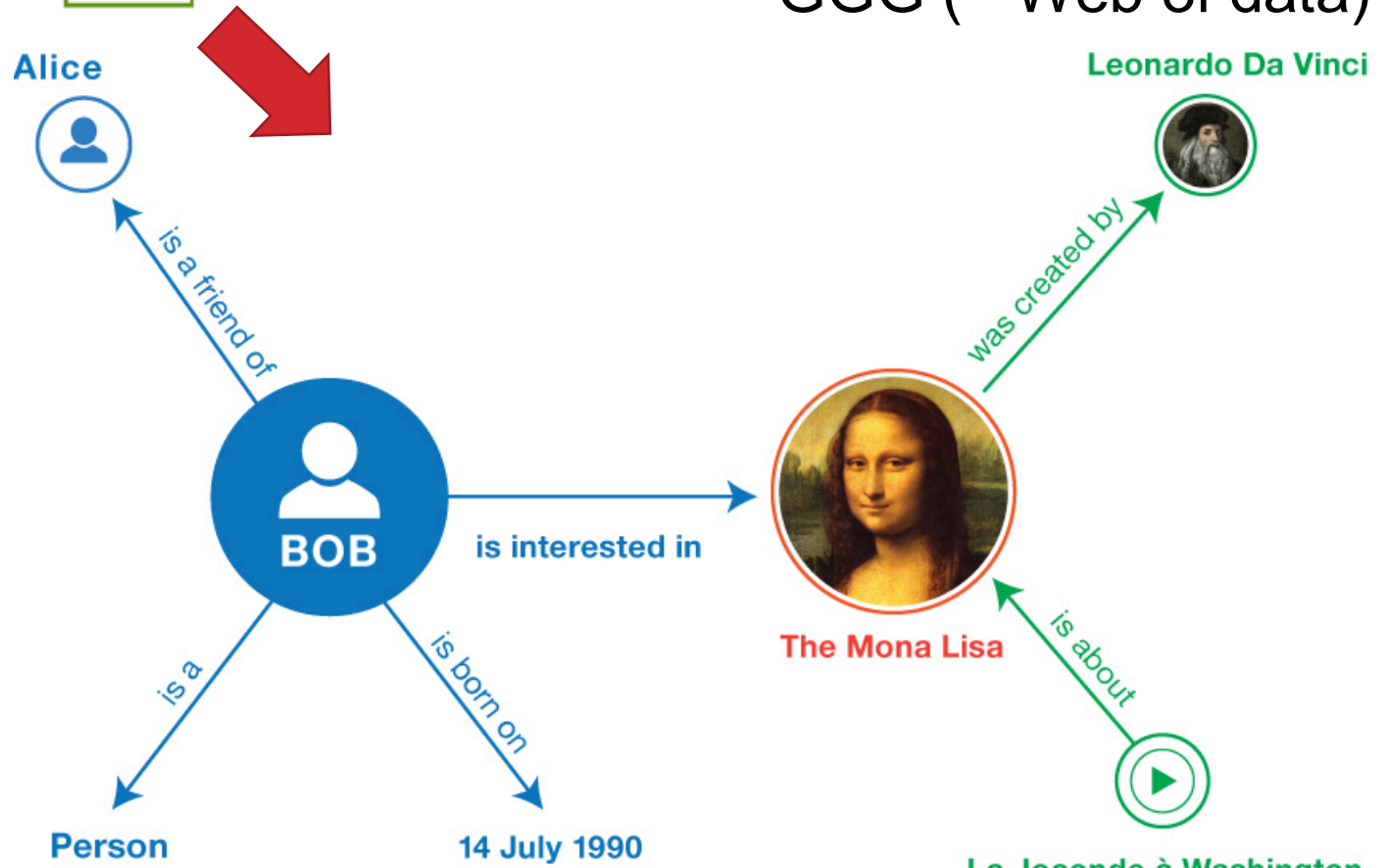*And whichever device I use … it will access a situation-appropriate view of an integration of everything I know about that flight from different sources.*

The task of booking and taking the flight will … be primary things in my awareness, the websites involved will be secondary things, and the network and the devices tertiary."

TBL's blog post on GGG: http://dig.csail.mit.edu/breadcrumbs/node/215

WWW (= Web of documents)

GGG (= Web of data)

Alice

Leonardo Da Vinci

BOB
- is a friend of → Alice
- is interested in → The Mona Lisa
- is a → Person
- is born on → 14 July 1990

The Mona Lisa
- was created by → Leonardo Da Vinci
- is about → La Joconde à Washington

WEB OF DOCS VS. WEB OF DATA

	Web of Documents	Web of Data
Designed for	Human consumption	Humans 'served' by computer programs
Primary objects	Documents (including multimedia)	Resources* (and descriptions of resources)
Links between	Documents	Resources
Degree of structure	Fairly low	High
Semantics of content and links	Implicit	Explicit
Analogy	A global file system	A global database

* Resource is anything that can be uniquely identified (has its URI); e.g., a resource is Belgrade, and its description is DBpedia entry for Belgrade

WEB OF (OPEN) DATA

Linked Datasets as of August 2014

Publications
Life Sciences
Cross-Domain
Social Networking
Geographic
Government
Media
User-Generated Content
Linguistics

Source: http://lod-cloud.net/

Animation illustrating the evolution of LOD: http://goo.gl/49p9Eh

GIGANTIC GLOBAL GRAPH (2)

Gigantic graph-based knowledge bases that

- contain structured data extracted from Web pages
- continuously grow and evolve so that their content properly reflects the data and knowledge of the Web

Features:

- based on automated learning systems
- combine different Machine Learning methods to assure continuous improvement of the data/knowledge extraction process
- subject of extensive research (both in industry and academia) aimed at the improvement of the knowledge extraction process

READ THE WEB

Research project at the Carnegie Mellon University

- http://rtw.ml.cmu.edu/rtw/

Objectives:

- Develop a never-ending machine learning system for extracting structured information from unstructured Web pages
- The development of a gigantic structured KB that
    - reflects the factual content of the Web,
    - continually grows in terms of both predicates and instances,
   - could be useful to many AI efforts

NEVER ENDING LANGUAGE LEARNER (NELL)

NELL is an implementation of the Read the Web approach

It performs 2 tasks each day, 7 days per week:

- *Reading task*: extract new instances (of known categories and relations) from texts on the Web, and thus extend the KB

- *Learning task*: learn to 'read' better each day, as evidenced by the ability to extract more information more accurately
  - the learning components continuously retrain themselves using the growing KB as a source of training examples

NELL knowledge fragment
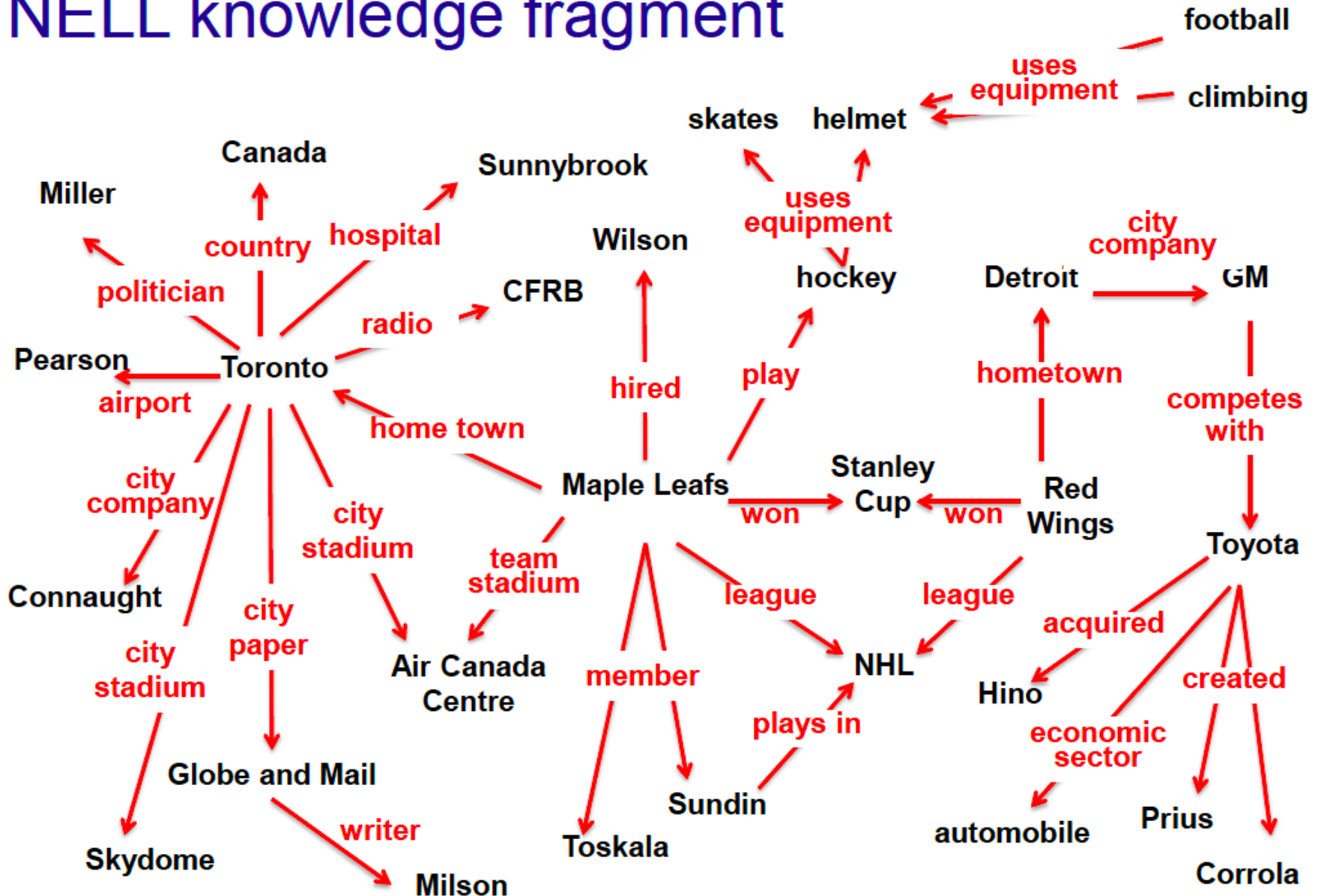
One can follow NELL while it 'reads',

and help it *learn* to 'read' better

Recently-Learned Facts twitter   Refresh

instance	iteration	date learned	confidence	
estonians is an ethnic group	956	23-oct-2015	97.9	👍 👎
jane_krakowski is an architect	955	20-oct-2015	100.0	👍 👎
cups spoons is an item found in the kitchen	956	23-oct-2015	99.6	👍 👎
voight_park is a zoo	959	07-nov-2015	90.5	👍 👎
michael_mayer is american	955	20-oct-2015	100.0	👍 👎
bob is a U.S. politician who holds the office of president	959	07-nov-2015	99.2	👍 👎
karl_rove works for fox	956	23-oct-2015	93.8	👍 👎
david_toseland died in the country england	960	23-nov-2015	100.0	👍 👎
cavaliers is a sports team that won the nba_finals	955	20-oct-2015	98.4	👍 👎
logan was born in chicago_south	960	23-nov-2015	99.6	👍 👎

Source: web site of the Read the Web project: http://rtw.ml.cmu.edu/rtw/

GOOGLE'S KNOWLEDGE VAULT (KV)

Envisioned as a probabilistic knowledge base which would contain all the factual knowledge of the Web, and would grow and evolve as the Web grows and evolves

Knowledge representation:

- relies on {subject-predicate-object} triplets (like DBpedia),
- each triplet has its *confidence score,* which represents the estimated probability of the triplet's validity / accuracy

The stored knowledge includes:

- facts automatically extracted from Web pages (uncertain, unverified knowledge)
- knowledge gathered from existing knowledge bases (verified, validated knowledge)

GOOGLE'S KNOWLEDGE VAULT (KV)

Comparison of KV with other state-of-the-art knowledge bases

Name	# Entity types	# Entity instances	# Relation types	# Confident facts (relation instances)
Knowledge Vault (KV)	1100	45M	4469	271M
DeepDive [32]	4	2.7M	34	7M[a]
NELL [8]	271	5.19M	306	0.435M[b]
PROSPERA [30]	11	N/A	14	0.1M
YAGO2 [19]	350,000	9.8M	100	4M[c]
Freebase [4]	1,500	40M	35,000	637M[d]
Knowledge Graph (KG)	1,500	570M	35,000	18,000M[e]

Table 1: Comparison of knowledge bases. KV, DeepDive, NELL, and PROSPERA rely solely on extraction, Freebase and KG rely on human curation and structured sources, and YAGO2 uses both strategies. Confident facts means with a probability of being true at or above 0.9.

Source: X. Dong et al. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14) (link)

DIFFBOT'S GLOBAL INDEX

"…in recent months Diffbot has been analyzing websites to build its index at a rate of up to 15 million pages a day.

Its Global Index now contains more than 600 million objects (this can be anything from a celebrity to an Ikea chair model) and 19 billion facts.

'Our approach is fairly radical in that there's no human behind the curtain'

Diffbot … [is] enhancing other search engines including Microsoft's Bing and DuckDuckGo, and powering apps for companies such as Cisco and AOL"

Diffbot Challenges Google Supremacy With Rival Knowledge Graph

Xconomy, June 4, 2015

RECOMMENDED VIDEOS, ARTICLES

- [article] Diffbot Bests Google's Knowledge Graph To Feed The Need For Structured Data ([link](#))

- [video] Mike Tung, DiffBot CEO, Turning the Web into a Structured Database ([link](#))

- [video] Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion ([link](#))
  - this lecture explains the overall idea and concept of the KV, the principles its knowledge collection process is based upon, and the like

- [video] Tom Mitchell, Never-Ending Learning to Read the Web ([link](#))

- [video] From Structured Data to Knowledge Graph, Google I/O 2013 ([link](#))