

Using Virtual Knowledge Graphs for ML

Peter Hopfgartner

Ontopic

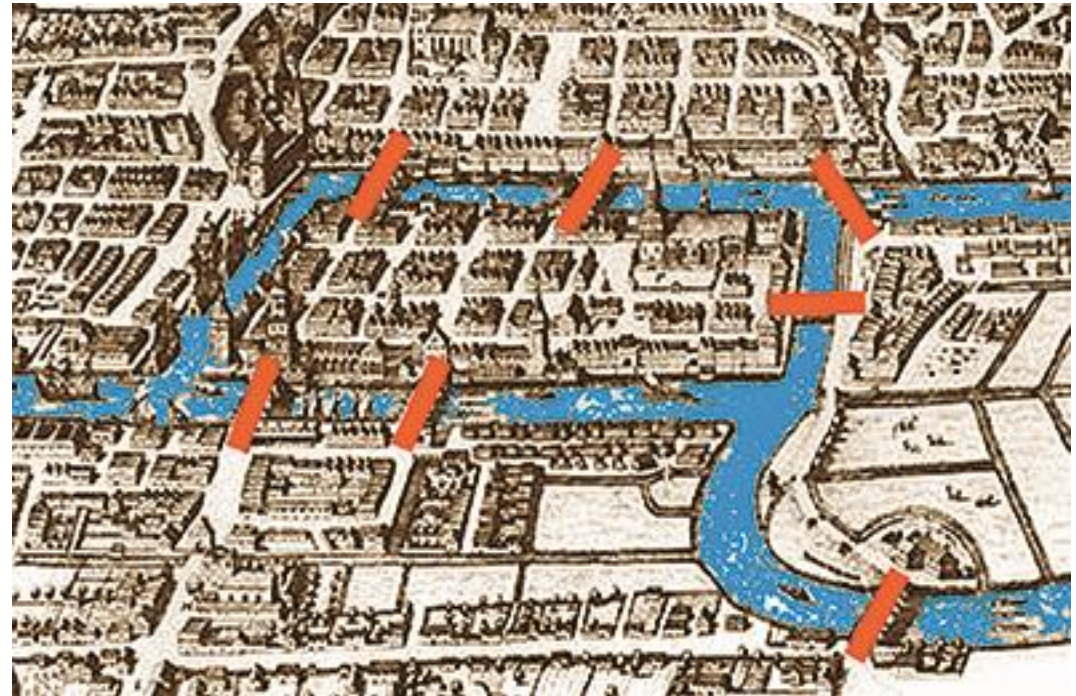
Bolzano, 3 September 2019

What is a graph?

Origin:

Euler's issues with the 7
bridges of Königsberg:

Is it possible to have a
walk through all 4 parts of
the city crossing each
bridge once and only once?

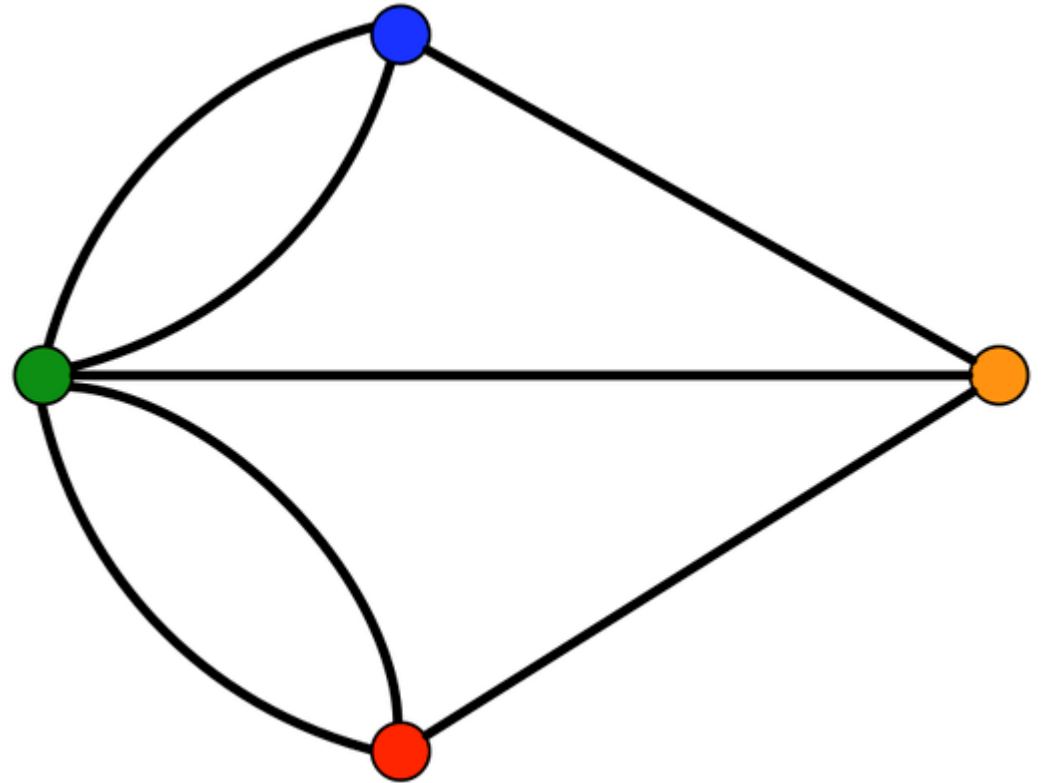


What is a graph?

Heuristic definition:

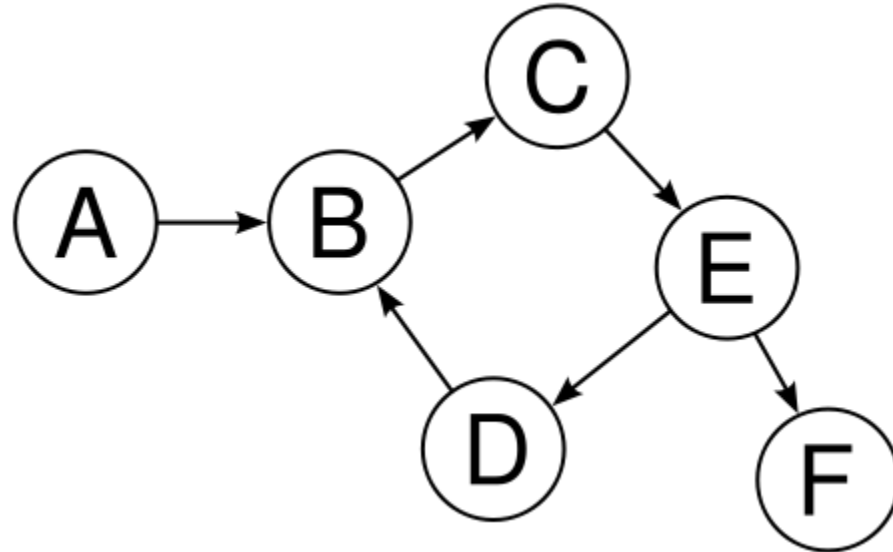
A mathematical structure

- To model pairwise relations
- Made of vertices (entities, nodes, ...)
- And edges (relations, connections, ...)



Directed Graphs

The connection between two nodes has a direction.



Knowledge Graphs

Name first
appeared at the
Google post in
2012



Official Blog

Insights from Googlers into our products,
technology, and the Google culture

Introducing the Knowledge Graph: things, not strings

May 16, 2012

Cross-posted on the [Inside Search Blog](#)

Search is a lot about discovery—the basic human need to learn and broaden your horizons. But searching still requires a lot of hard work by you, the user. So today I'm really excited to launch the Knowledge Graph, which will help you discover new information quickly and easily.

<https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html#>

Googles Knowledge Graphs in practice

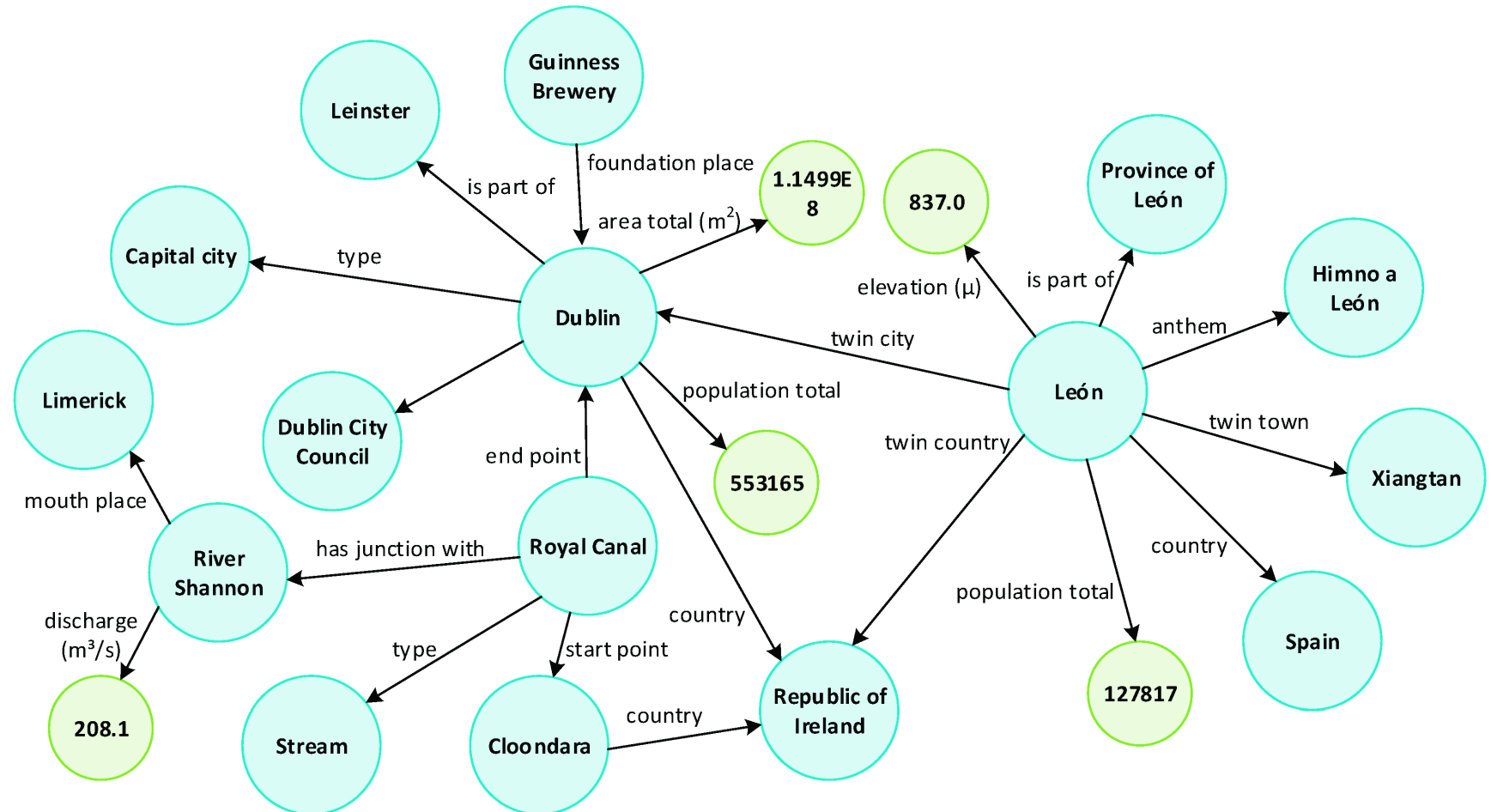
The image shows a Google search for 'Bolzano'. The search bar at the top contains the word 'Bolzano'. Below the search bar, there are navigation options: 'All', 'Images', 'Maps', 'News', 'Videos', 'More', 'Settings', and 'Tools'. The search results show 'About 58,300,000 results (0.77 seconds)'. The first result is 'Bolzano - Wikipedia' with the URL 'https://en.wikipedia.org › wiki › Bolzano'. Below this, there is a brief description: 'Bolzano is the capital city of the province of South Tyrol in northern Italy. With a population of 107,436, Bolzano is also by far the largest city in South Tyrol and ...'. Further down, there are details: 'Region: Trentino-Alto Adige/Südtirol', 'Dialing code: 0471', 'Country: Italy', and 'Demonym(s): Italian: bolzanini; German: Bozn...'. There are also links to 'Bolzano Airport', 'Timeline of Bolzano', and 'Bolzano/Bozen railway station'. Below this, there is another 'Bolzano - Wikipedia' result in Italian, with the URL 'https://it.wikipedia.org › wiki › Bolzano' and a 'Translate this page' link. This result includes pronunciation information and a list of inhabitants: 'Nome abitanti: (IT) bolzanini; (DE) Bozner', 'Patrono: Maria Assunta, Arrigo da Bolzano', and 'Lingue ufficiali: Italiano, Tedesco'. There are also links to 'Provincia autonoma di Bolzano', 'Arrigo da Bolzano', 'Duomo di Bolzano', and 'Don Bosco'. At the bottom, there is a result for 'Bolzano, Tourist Board Official Web Site' with the URL 'https://www.bolzano-bozen.it › bolzano'. The Knowledge Panel on the right side of the search results is titled 'Bolzano' and includes the subtitle 'City in Italy'. It features a photograph of the city and a map. The text in the panel describes Bolzano as a city in the South Tyrol province of north Italy, set in a valley amid hilly vineyards. It mentions that it's a gateway to the Dolomites mountain range in the Italian Alps. In the medieval city center, the South Tyrol Museum of Archaeology features the Neolithic mummy called Ötzi the Iceman. Nearby is the imposing 13th-century Mareccio Castle, and the Duomo di Bolzano cathedral with its Romanesque and Gothic architecture. The panel also lists 'Elevation: 262 m', 'Weather: 21°C, Wind W at 2 km/h, 79% Humidity', and 'Population: 106,951 (2017) Istat'. A blue arrow points from the bottom of the search results towards the Knowledge Panel.

This comes from the Google Knowledge Graph

Put knowledge into digital memories

Nodes are resources

Edges define the kind of relationship between resources



RDF – Resource Description Framework

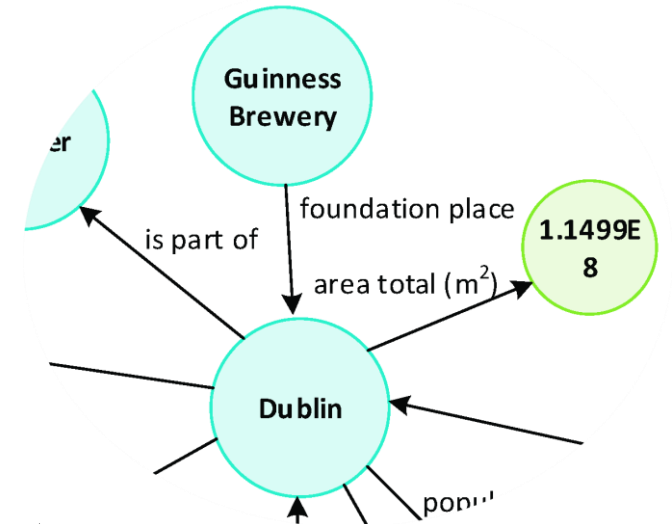
RDF is a W3C standard.

Characteristic element is the triple

Guinness Brewery – **founded in** – **Dublin**.

Subject (S) – Predicate (P) – Object (O)

Dublin – **total area (m²)** – **1.1499E8**

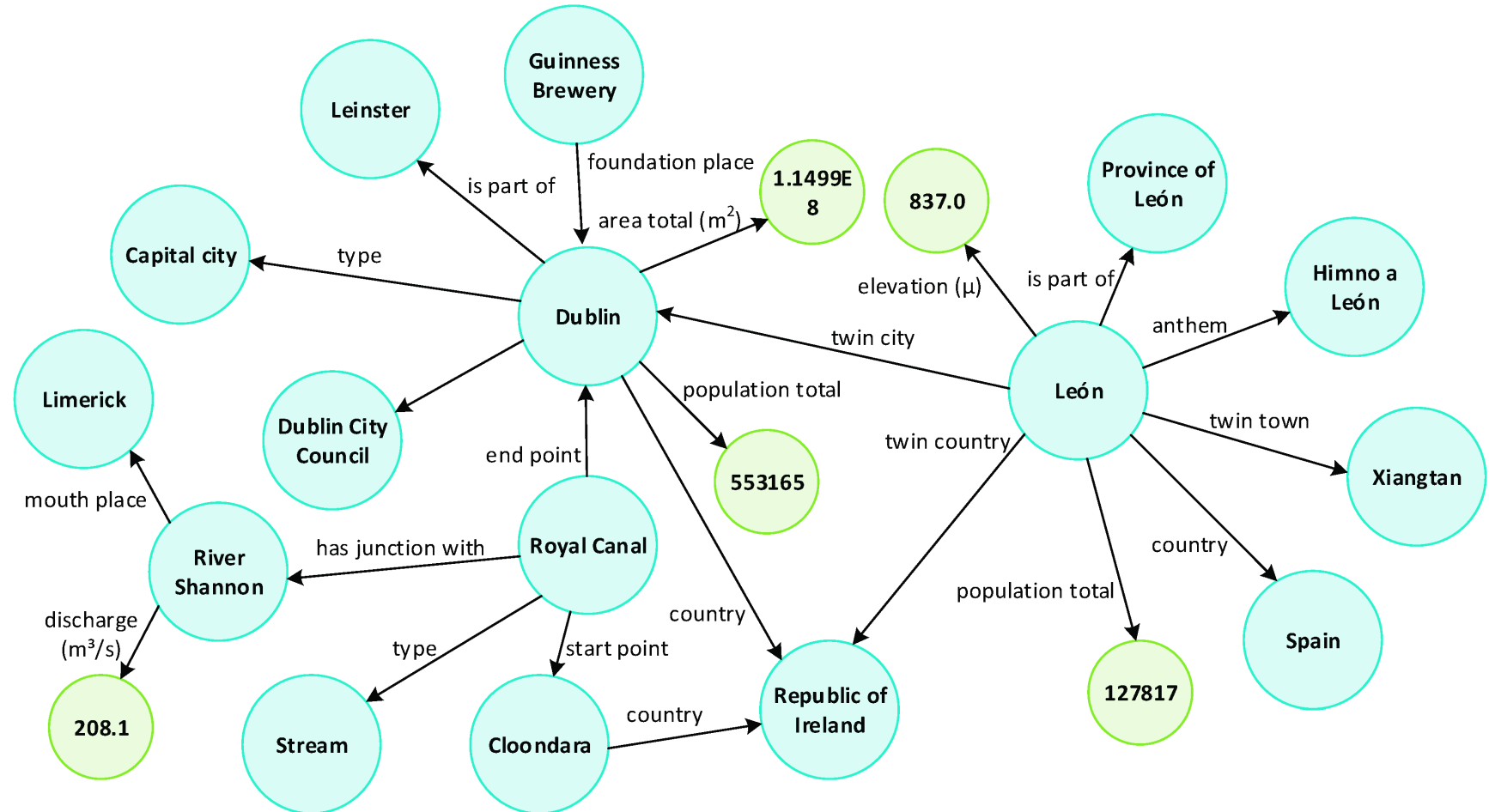


Graphs are very flexible

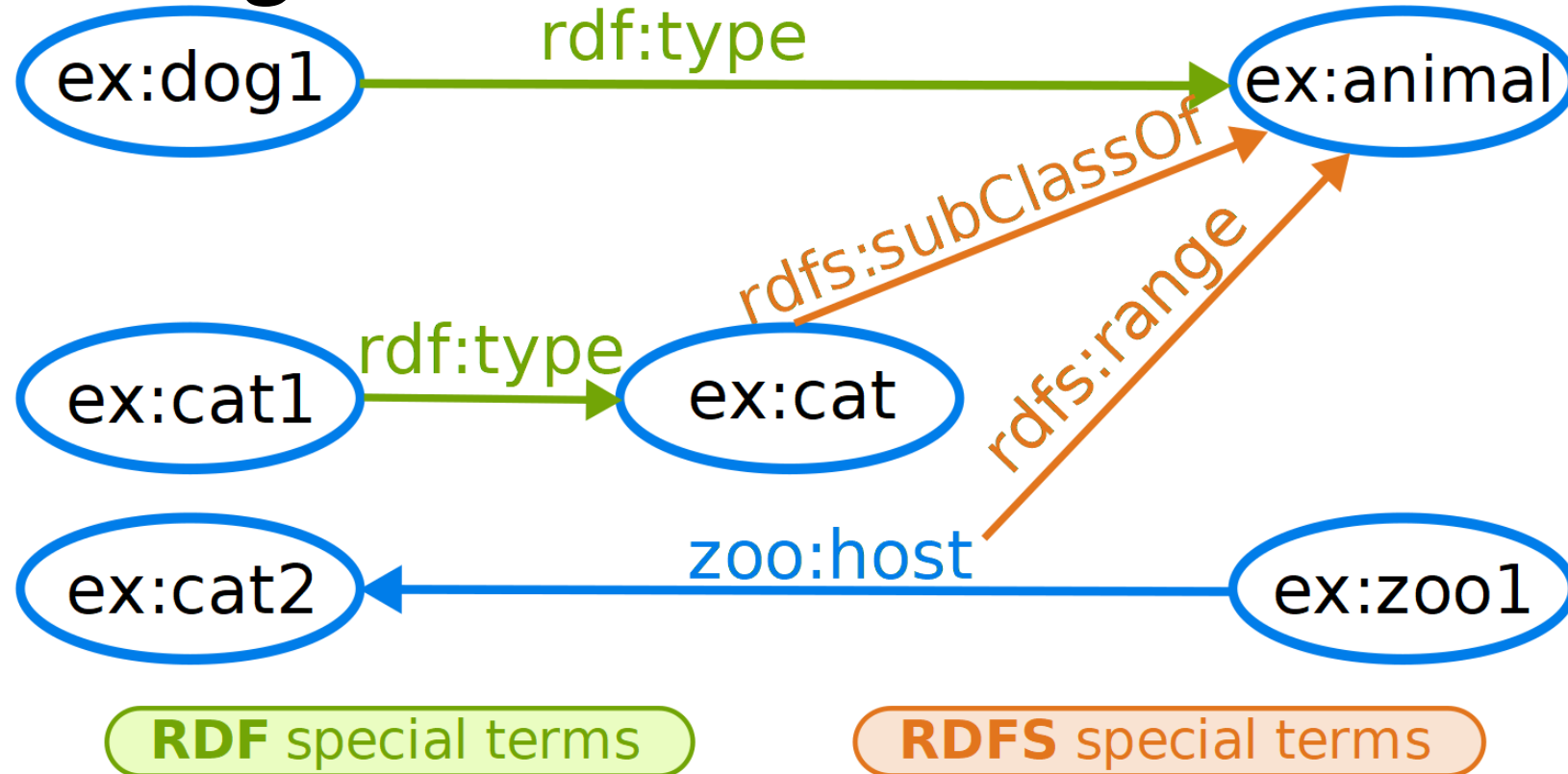
Yet,
Knowledge is
not only
tons of
information.

=>

Some
structure is
needed.

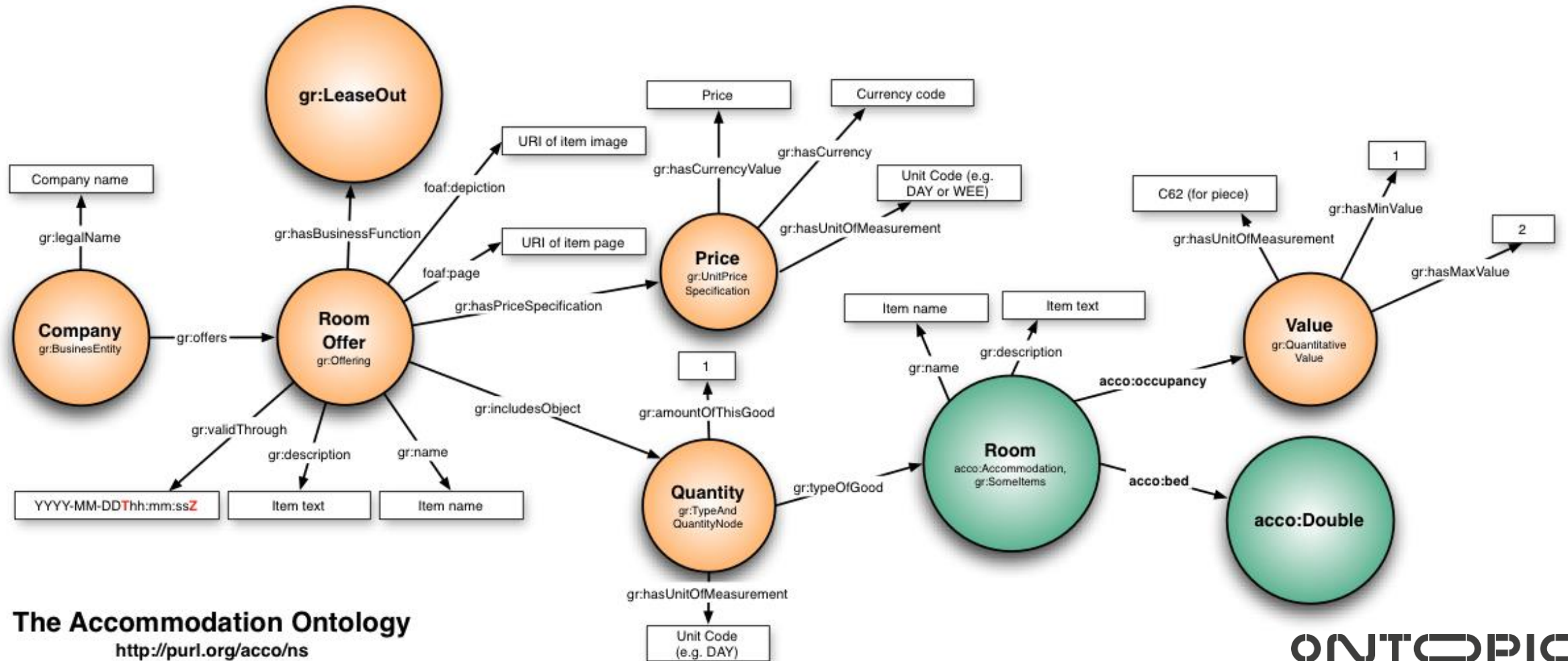


RDFS: bring some structure to it



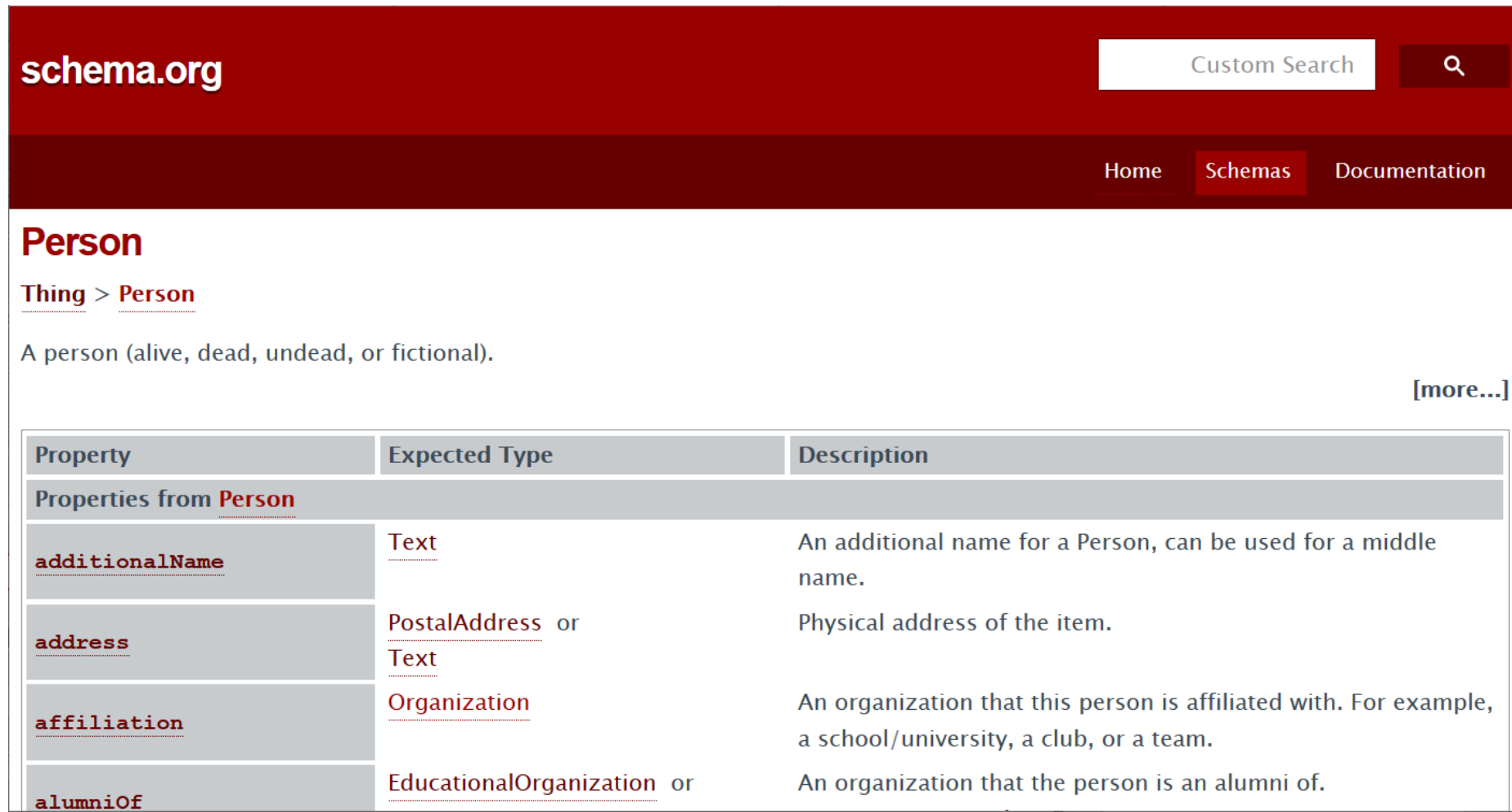
Classes and properties are called vocabularies

More expressive ontology languages: OWL (Web Ontology Language)



The importance of common vocabularies

Common vocabularies allow sharing of information



The screenshot shows the schema.org website interface. At the top, there is a dark red header with the 'schema.org' logo on the left, a search bar with the text 'Custom Search' and a magnifying glass icon on the right, and navigation links for 'Home', 'Schemas', and 'Documentation' in the center. Below the header, the page title is 'Person' in a large, bold, dark red font. Underneath the title, there is a breadcrumb trail: 'Thing > Person'. A descriptive sentence follows: 'A person (alive, dead, undead, or fictional).'. To the right of this sentence is a '[more...]' link. Below the description is a table with three columns: 'Property', 'Expected Type', and 'Description'. The table lists several properties of the 'Person' class, including 'additionalName', 'address', 'affiliation', and 'alumniOf', each with its expected type and a brief description.

Property	Expected Type	Description
Properties from Person		
<u>additionalName</u>	Text	An additional name for a Person, can be used for a middle name.
<u>address</u>	PostalAddress or Text	Physical address of the item.
<u>affiliation</u>	Organization	An organization that this person is affiliated with. For example, a school/university, a club, or a team.
<u>alumniOf</u>	EducationalOrganization or	An organization that the person is an alumni of.

Google structured data

```

1 <!doctype html>
2 <html lang="en">
3   <head>
4     <!-- Required meta tags -->
5     <meta charset="utf-8">
6     <meta name="viewport" content="width=device-width, initial-
scalew=1, shrink-to-fit=no">
7     <meta name="description" content="Virtual Knowledge Graph,
OBDA, Ontology Based Data Access, Data Integration">
8     <meta name="author" content="Peter Hopfgartner, Ontopic">
9     <link rel="stylesheet" href="bs/css/bootstrap.min.css" >
10    <link href="css/ontopic.css" rel="stylesheet">
11    <title>Ontopic - The Virtual Knowledge Graph Company</title>
12    <script type="application/ld+json">
13    {
14      "@context" : "https://schema.org",
15      "@type" : "Organization",
16      "name" : "Ontopic",
17      "url" : "https://ontopic.biz",
18      "slogan": "The Virtual Knowledge Graph Company",
19      "address" : {
20        "@type" : "PostalAddress",
21        "streetAddress" : "via Alessandro Volta, 13/A",
22        "addressCountry" : "Italy"

```

Organization		0 ERRORS	0 WARNINGS	^
@type	Organization			
name	Ontopic			
url	https://ontopic.biz/			
slogan	The Virtual Knowledge Graph Company			
logo	https://ontopic.biz/img/logo-corto-01.png			
address				
@type	PostalAddress			
streetAddress	via Alessandro Volta, 13/A			
addressLocality	Bolzano			
postalCode	39100			
addressCountry				
@type	Country			
name	Italy			
contactPoint				
@type	ContactPoint			



What's the the thing with „virtual“?

Classical RDF data is kept in *triple stores*.

For analysing data from enterprise databases, this means:

- Another copy of the data
- Latency for data
- Latency for changes in the schema
- Does not scale well with ontologies

What's the thing with „virtual“?

Virtual Knowledge Graphs: data is kept in the original databases.

Mapping:

The recipe for describing how the data in the database is linked to the graph is called „*mapping*“.

Ontop does exactly this.

Mapping the Knowledge Graph to data sources

Edit Mapping

Mapping ID:

Target (Triples Template):
<http://musicbrainz.org/artist/{gid}#_> a :SoloMusicArtist .

Source (SQL Query):
SELECT *
FROM artist
WHERE artist.type = 1

Test SQL Q... (100 rows)

Edit Mapping

Mapping ID:

Target (Triples Template):
<http://musicbrainz.org/artist/{gid}#_> :member_of
<http://musicbrainz.org/artist/{band}#_> .

Source (SQL Query):
SELECT a1.gid, a2.gid AS band
FROM artist a1
INNER JOIN l_artist_artist ON a1.id = l_artist_artist.entity0
INNER JOIN link ON l_artist_artist.link = link.id
INNER JOIN link_type ON link_type = link_type.id
INNER JOIN artist a2 on l_artist_artist.entity1 = a2.id
WHERE link_type.gid='5be4c609-9afa-4ea0-910b-12ffb71e3821'
AND link.ended=FALSE

Test SQL Q... (100 rows)

Magic numbers

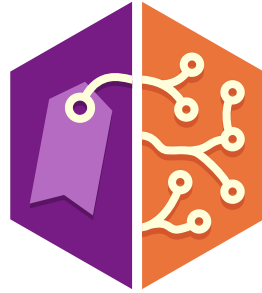
Complex queries

SPARQL: The query language for RDF

- Carries many ideas from SQL for relational databases.
- Operates on triples, not on tables
- Uses HTTP as a transfer protocol

Practical Example

Data from MusicBrainz



Ontology by



Mapping : <https://github.com/metabrainz/MusicBrainz-R2RML>

SPARQL: An example

```
# Simple example
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX mo: <http://purl.org/ontology/mo/>

SELECT *
WHERE {
    ?artist rdf:type mo:MusicArtist;
           foaf:name "Herbert Pixner" .
}
```

artist

[http://musicbrainz.org/artist/0c7785bb-1d2e-4384-a572-7e69954508e9#_>](http://musicbrainz.org/artist/0c7785bb-1d2e-4384-a572-7e69954508e9#_)

SPARQL: More infos

```
# local artists' friends
PREFIX : <http://purl.org/ontology/mo/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX mo: <http://purl.org/ontology/mo/>

SELECT *
WHERE {
    ?artist rdf:type mo:MusicArtist;
        foaf:name ?name ;
        foaf:based_near ?area .
    ?area rdfs:label ?area_name .
    ?artist foaf:name "Herbert Pixner" ;
}
```

artist	name	area	area_name
< <a <="" href="http://musicbrainz.org/artist/0c7785bb-1d2e-4384-a572-7e69954508e9#_></td><td>Herbert Pixner" td=""><td><</td><td>" italy"<="" td=""></td>	< </td><td>" italy"<="" td="">		

SPARQL: Connections matter

```
# local artists' friends
PREFIX : <http://purl.org/ontology/mo/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX mo: <http://purl.org/ontology/mo/>

SELECT DISTINCT ?name ?name2
WHERE {
  ?artist rdf:type mo:MusicArtist;
    foaf:name ?name ;
    foaf:based_near ?area;
    foaf:made ?something .
  ?area rdfs:label ?area_name .
  ?artist2 foaf:made ?something ;
    foaf:based_near ?area2;
    foaf:name ?name2 .
  ?area2 rdfs:label ?area_name2 .
  FILTER ( ?area_name = "Bolzano" ||
    ?area_name = "Trentino-Alto Adige" ||
    ?area_name = "Italy" ||
    ?area_name = "Austria" )
}
```

name	name2	
Andreas Fulterer	G.G. Anderson	
Wicked & Bonny	Wicked & Bonny	
Andreas Fulterer	Bernhard Brink	
Dominik Plangger	Papermoon	
Andreas Fulterer	Die Paldauer	
Andreas Fulterer	Andy Borg	
Voices of Decay	Voices of Decay	
Andreas Fulterer	Oliver Haidt	
Kinderchor der Kantorei Leonhard Lechner	Barry Faldner	
Kinderchor der Kantorei Leonhard Lechner	Orchestra Haydn di Bolzano e Trento	
Andreas Fulterer	Michael Morgan	
...	...	

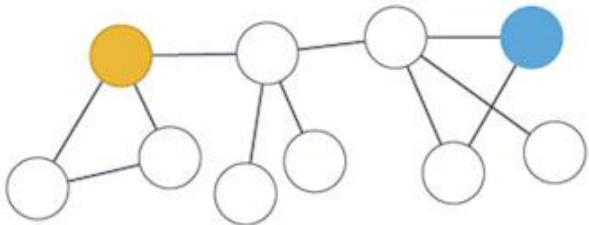
Cooperate with others

Results from the graph can be used as attributes for your input vectors.

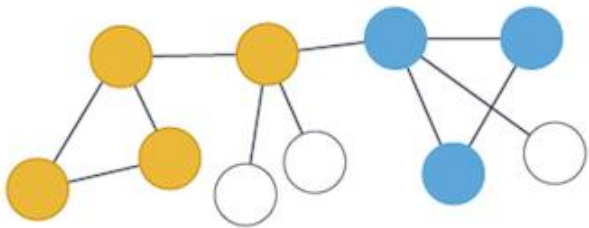
Specialized software for Big Data, e.g. with Apache Spark:



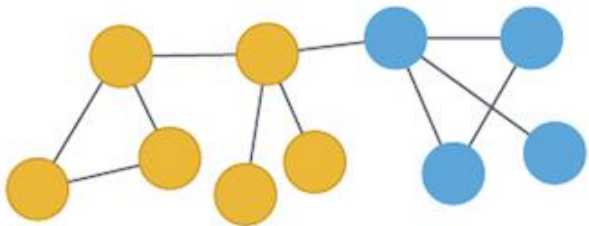
e.g. Label Propagation



Pass 1



Pass 2



URI	Name	Label
http://musicbrainz.org/artist/bfcc6d75-a6a5-4bc6-8282-47aec8531818#_	Cher	0
http://musicbrainz.org/artist/1aacd39b-8731-4923-a37d-884e2176ef93#_	Giorgia	0
http://musicbrainz.org/artist/a641ad4b-4b45-4a7b-b076-711f4775094d#_	Patsy Kensit	0
http://musicbrainz.org/artist/9072df14-b61e-42e2-b4f4-6bbb7fdb5586#_	Tina Turner	0
http://musicbrainz.org/artist/70ea63ea-70dc-4b63-951a-2c249d2b3b0a#_	Ricky Martin	0
http://musicbrainz.org/artist/0102e395-e4bd-476e-9a57-80e8335ba64a#_	Sonologist	1
http://musicbrainz.org/artist/0548426a-7265-4671-8a08-19ca2baa47e2#_	Juda	3
http://musicbrainz.org/artist/062fceb6-81ab-4769-a6c2-a4d866350cd0#_	Elepharmers	5
http://musicbrainz.org/artist/08a5240a-0a45-43cd-af95-49b9a7d6bece#_	Lorenzo Campani	8
http://musicbrainz.org/artist/5bb63765-2282-4f6c-b823-20b95956fbef#_	Dynamic Base	9
http://musicbrainz.org/artist/d0e3329d-f909-490b-b755-dfa31f446eaf#_	Giorgia Angiuli	9
http://musicbrainz.org/artist/098acbe6-c428-4ae2-9cd0-3ba80162befb#_	Corona	9

Conclusions

- Virtual Knowledge Graphs can be easily integrated with your existing IT landscape
- SPARQL is very flexible for data preparation, e.g. based on topological properties
- These can be elaborated with 3rd party software, like Apache Spark, for calculating connection based attributes for improved ML

Thank you for attending

Still some questions?

Ontop is on GitHub

<https://github.com/ontop/ontop>

