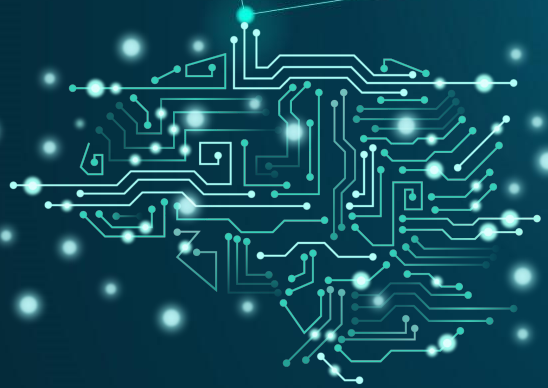


Organize Seus Dados Com Grafos De Conhecimento

Rafael Rocha

SOBRE MIM

Rafael Rocha
Analista de TI - UFMG
TPD (2006) - Fabrai-MG
Especializaçâo (2010) - PUC-MG
Mestrado (2021) PPGGOC/ECI-UFMG

Agenda

THE
DEVELOPER'S
CONFERENCE

01 Introdução

02 Grafo de Conhecimento

03 Aplicabilidade

04 Considerações Finais

Introdução

01

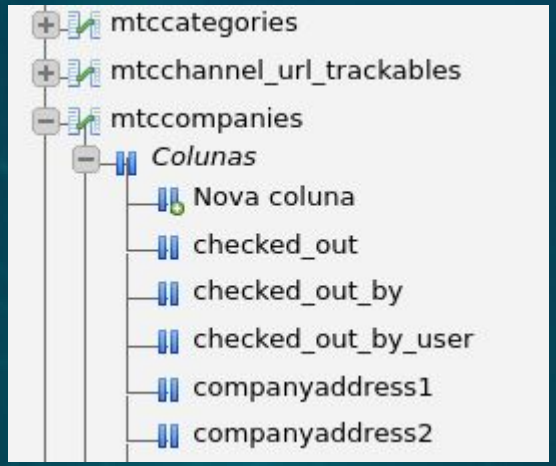
Introdução - Persistência

- Dados podem ser persistidos de modo estruturados, semi-estruturado e não estruturados.

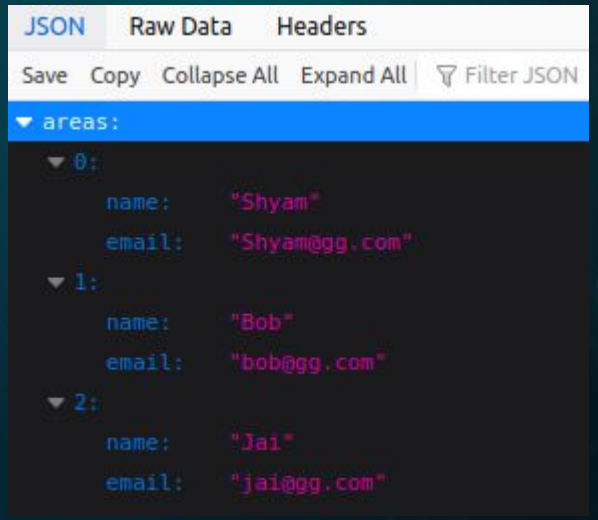
Introdução - Persistência

THE DEVELOPER'S CONFERENCE

(Semi-)Estruturados



	C	D	E	F
1	eow	cause	cause_short	date
2	NC	EOW: Sunday	January 2	2000
3	GA	EOW: Monday	January 3	2000
4	GA	EOW: Monday	January 3	2000
5	CA	EOW: Tuesday	January 11	2000
6	CA	EOW: Tuesday	January 11	2000
7	NC	EOW: Friday	January 14	2000



Introdução - Persistência

THE DEVELOPER'S CONFERENCE

- Não Estruturados

Introdução - Persistência

- Qual a semelhança desses tipos de persistências?

Introdução - Persistência

- Qual a semelhança desses tipos de persistências?
- Não explicitam formalmente significado, ou seja, conhecimento.

Introdução - Persistência

- Qual a semelhança desses tipos de persistências?
- Não explicitam formalmente significado, ou seja, conhecimento.
- Conhecimento é um **crença verdadeira justificada** (Platão - Teeteto).

Introdução - Persistência

- Qual a semelhança desses tipos de persistências?
- Não explicitam formalmente significado, ou seja, conhecimento.
- Conhecimento é um **crença verdadeira justificada** (Platão - Teeteto).
- O São Paulo joga em São Paulo hoje. Rezo para São Paulo ajudar nessa partida.

Introdução - Persistência

THE DEVELOPER'S CONFERENCE

- Qual a semelhança desses tipos de persistências?
- Não explicitam formalmente significado, ou seja, conhecimento.
- Conhecimento é um **crença verdadeira justificada** (Platão - Teeteto).
- O São Paulo joga em São Paulo hoje. Rezo para São Paulo ajudar nessa partida.
- Um ser humano sabe, mas a máquina não.
  O São Paulo **(Time de Futebol)** joga em São Paulo **(Cidade)** hoje. Rezo para São Paulo **(Santo)** ajudar nessa partida.

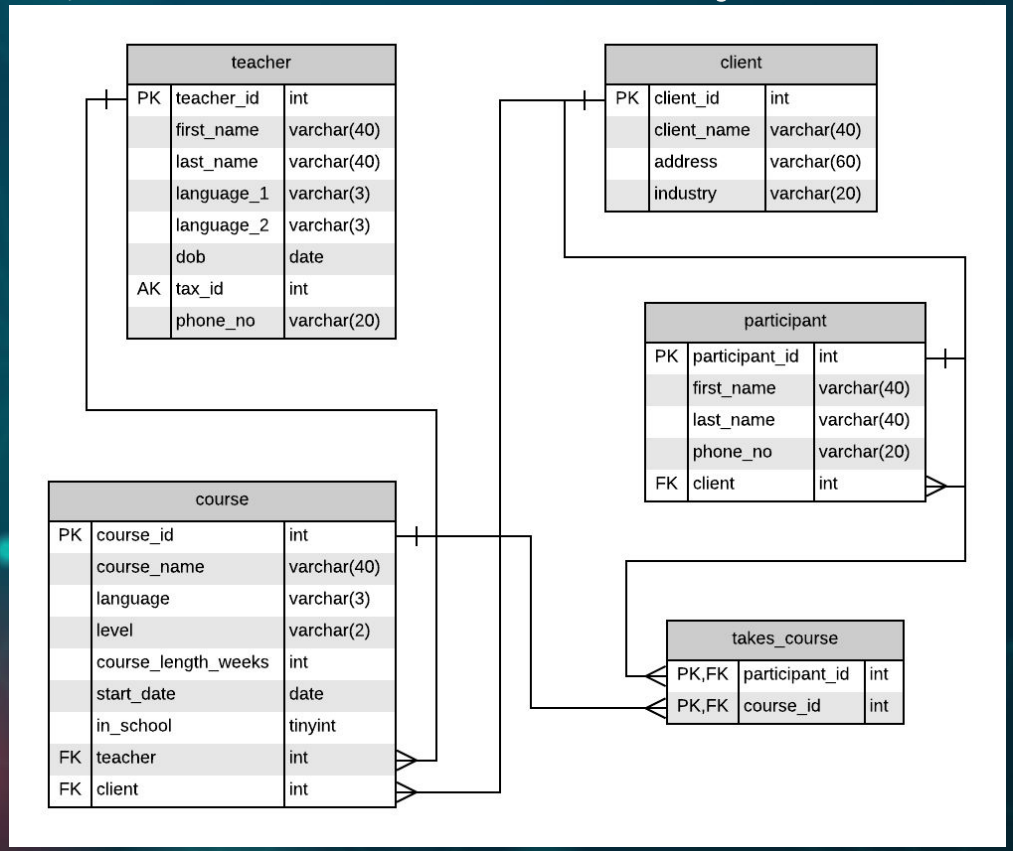
Introdução - Problema

- Como representar esse conhecimento em banco de dados relacional?

Introdução - Problema

THE DEVELOPER'S CONFERENCE

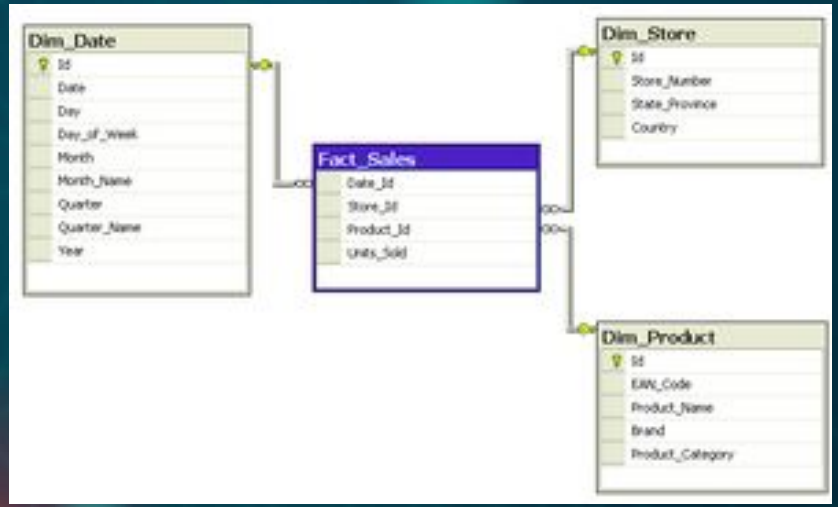
- Representar com normalização [TOWARDS, 2021].



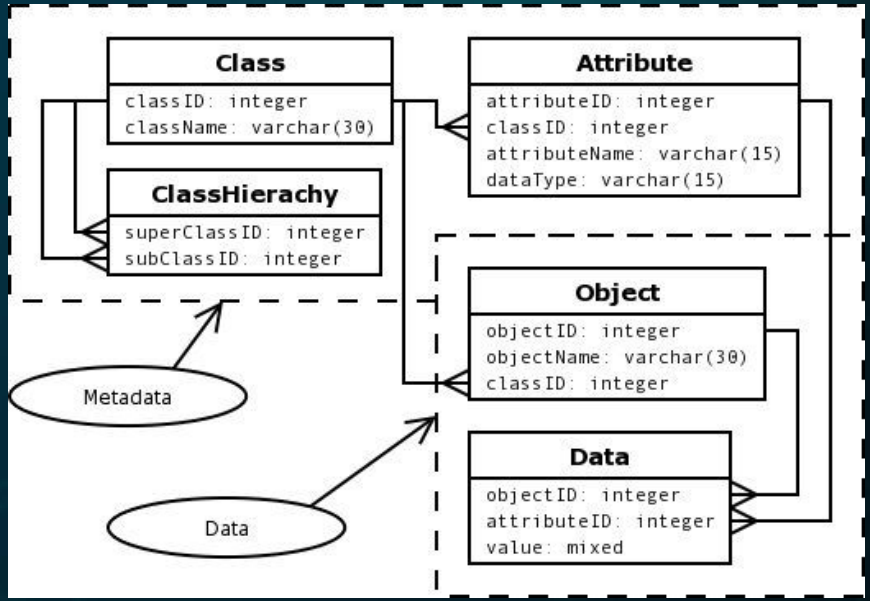
Introdução - Problema

- Representar sem normalização.

DW-Modelo Estrela [STAR, 2021]



Estrutura OO [ANHØJ, 2003]



Introdução - Solução

- Em todos os casos a interoperabilidade de significado (ou semântica) está em "contrato" codificado nos sistemas.

Introdução - Solução

- Em todos os casos a interoperabilidade de significado (ou semântica) está em “contrato” codificado nos sistemas.
- Representam conhecimento, mas com limitações.

Introdução - Solução

- Em todos os casos a interoperabilidade de significado (ou semântica) está em "contrato" codificado nos sistemas.
- Representam conhecimento, mas com limitações.

- Para esse contexto é necessário os **grafos de conhecimento**.

Grafo De Conhecimento 02

Grafo De Conhecimento

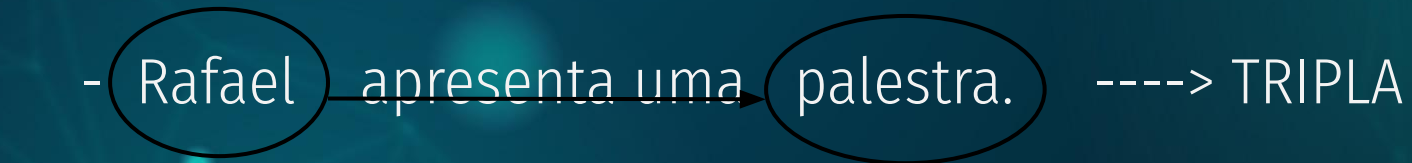
- A representação básica de conhecimento é através de uma tripla. Uma tripla possui sujeito; predicado; objeto.

-Rafael apresenta uma palestra.

Grafo De Conhecimento

THE DEVELOPER'S CONFERENCE

- A representação básica de conhecimento é através de uma tripla. Uma tripla possui sujeito; predicado; objeto.



Grafo De Conhecimento

- Base de conhecimento <> grafo de conhecimento
- “Coisas” no lugar de strings e números (literais)
- Free Schema
- Facilidade de representar dados complexos
- Base matemática
    - Busca em profundidade
    - Busca em largura e distâncias
    - etc

Grafo De Conhecimento

- Os grafos podem ser nos padrões World Wide Web Consortium(W3C); Open Source; Proprietário.
- Eles são armazenados em bancos de dados orientados a grafos (NoSQL).
- Amazon Neptune (W3C/Proprietário); Neo4J (Proprietário); OrientDB (Open Source/Proprietário); AnzoGraph DB (W3C/Proprietário); OpenLink Virtuoso (W3C); GraphDB (W3C)...
- Banco de dados de triplas é chamado de triplestore.

Grafo De Conhecimento - W3C

THE DEVELOPER'S CONFERENCE

Resource Description Framework

O RDF teve esboço inicial em 1997 [RDF, 1997] com vocabulário base para representar o conhecimento.

Web Ontology Language

O OWL teve esboço inicial em 2002 [OWL, 2002]. Adição de vocabulário ontológico.

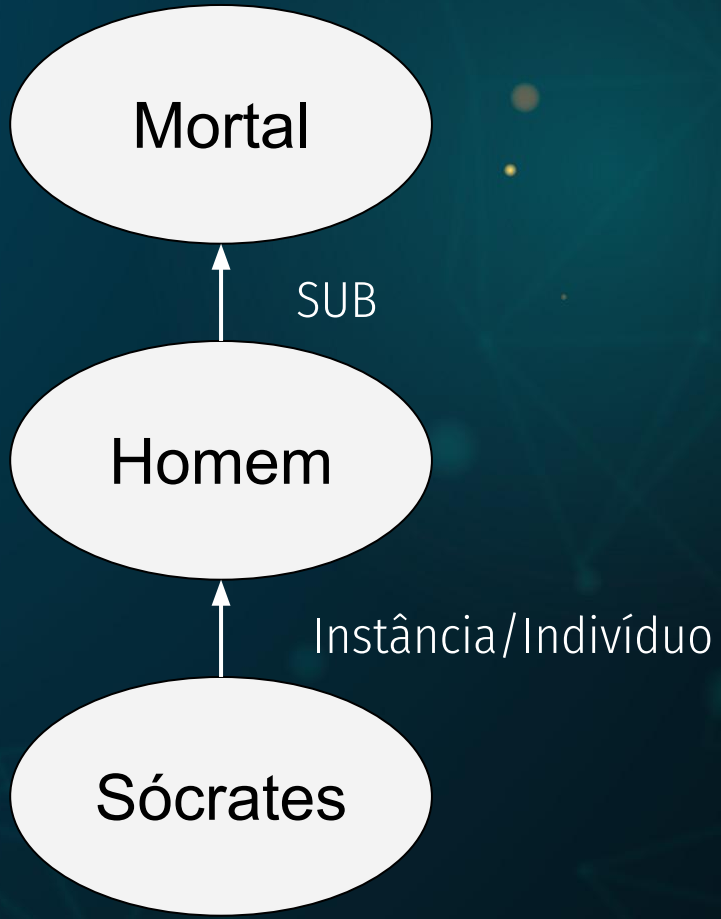
RDF Schema

O RDFS teve esboço inicial em 1999 [RDFS, 1999]. Adição de vocabulário taxonômico.

Grafo De Conhecimento - W3C

- Propriedades taxonômicas

Todo homem é mortal.
Sócrates é homem.
Logo, Sócrates é mortal.



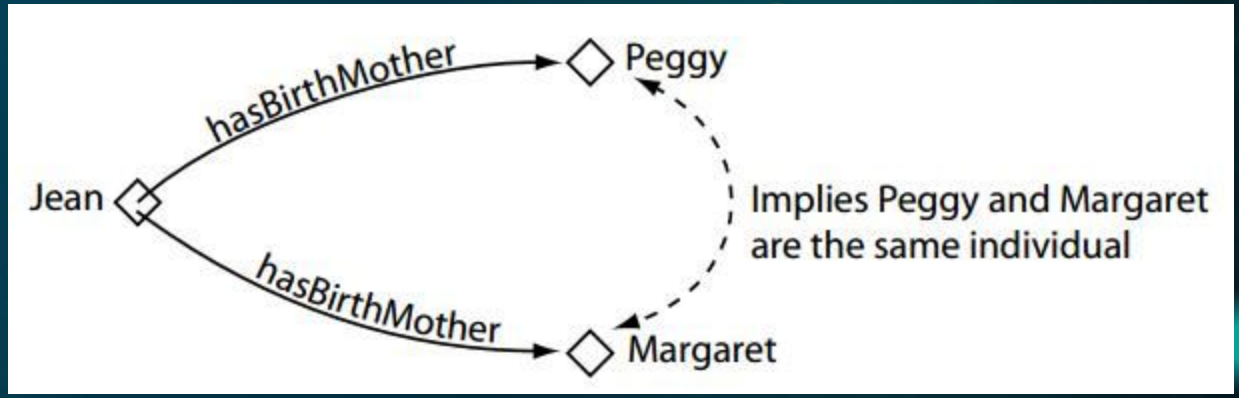
Grafo De Conhecimento - W3C

- OWL é linguagem para ontologia formal
- Ontologia é conceitualização de entidades e seus relacionamentos
- Formal pois possuir rigor claro e explícito

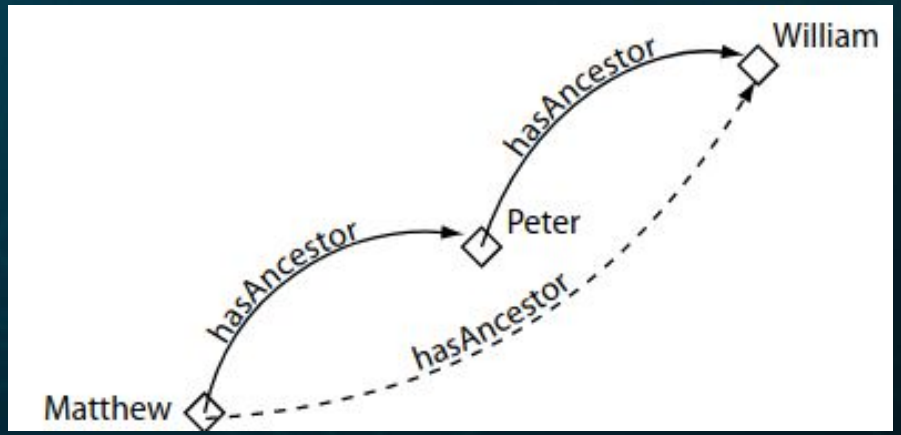
Grafo De Conhecimento - W3C

THE DEVELOPER'S CONFERENCE

- Propriedades ontológicas [HORRIDGE, 2009]
- Propriedade funcional



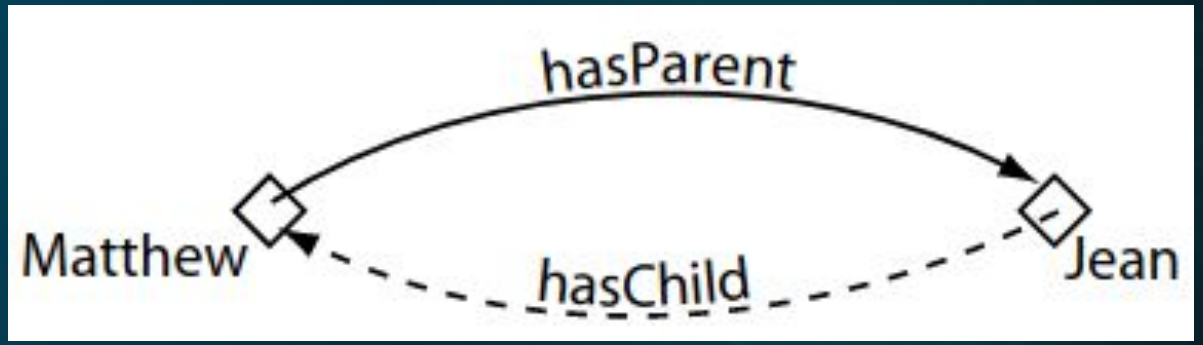
- Propriedade transitiva



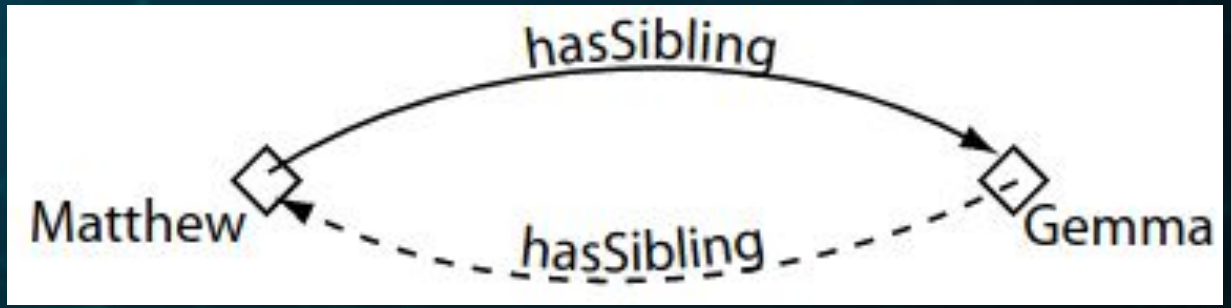
Grafo De Conhecimento - W3C

THE DEVELOPER'S CONFERENCE

- Propriedade inversa



- Propriedade simétrica



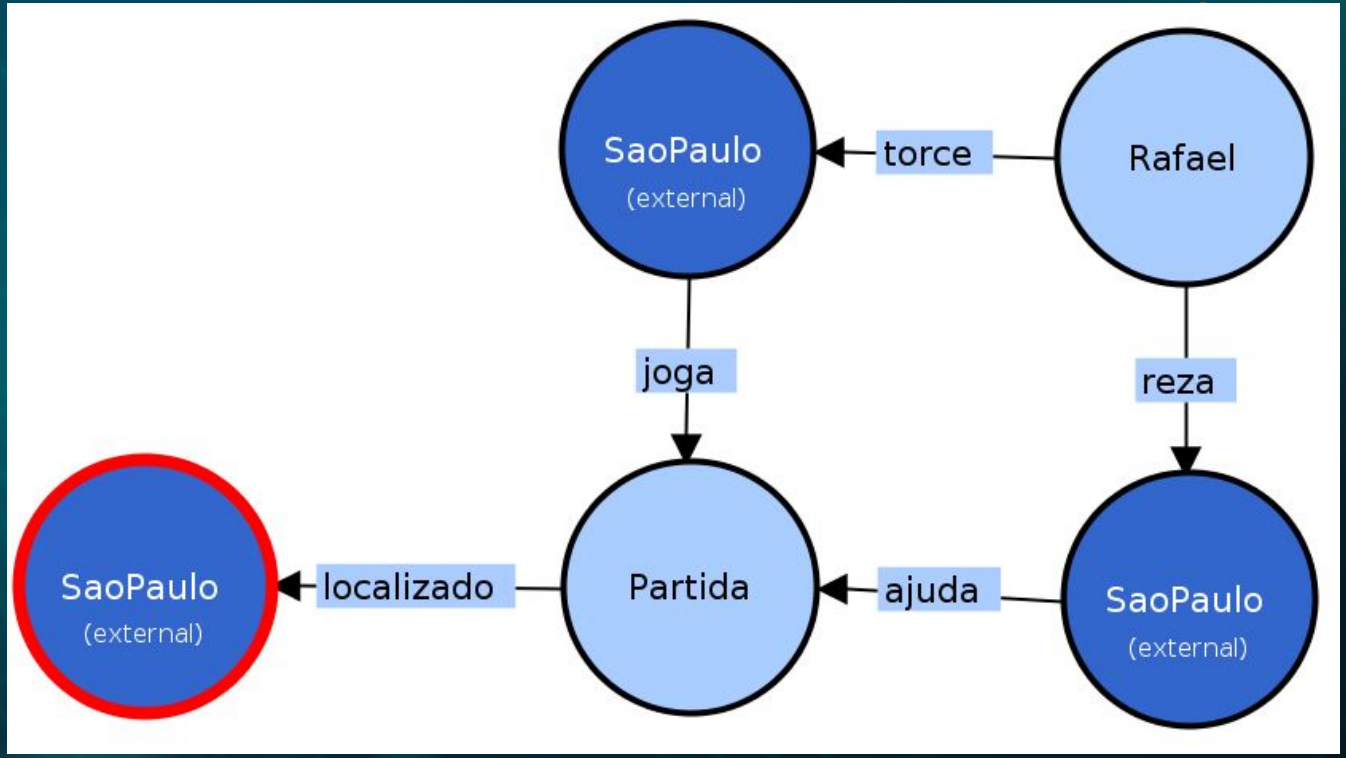
Grafo De Conhecimento - W3C

- Existem outras propriedades ontológicas.
- Ontologias de fundamentação
  - -Unified Foundational Ontology (UFO)
  - -Basic Formal Ontology (BFO)
- Ontologia de domínio
  - Open Biological and Biomedical Ontologies (OBO)
- Serialização dos grafos de conhecimento

Grafo De Conhecimento - W3C

THE DEVELOPER'S CONFERENCE

- O São Paulo (Time de Futebol) joga em São Paulo (Cidade) hoje. Rezo para São Paulo (Santo) ajudar nessa partida.



Grafo De Conhecimento - W3C

THE DEVELOPER'S CONFERENCE

- As triplas em XML permite conhecimento compartilhado com a máquina.
- Classes/vértices

```
<owl:Class rdf:about="https://thedevconf.com/tdc/2021/HOC-4320#Partida"/>
<owl:Class rdf:about="https://thedevconf.com/tdc/2021/HOC-4320#Rafael"/>
<owl:Class rdf:about="https://thedevconf.com/tdc/2021/HOC-4320/cidade#SaoPaulo"/>
<owl:Class rdf:about="https://thedevconf.com/tdc/2021/HOC-4320/santo#SaoPaulo"/>
<owl:Class rdf:about="https://thedevconf.com/tdc/2021/HOC-4320/time#SaoPaulo"/>
```

- Relacionamentos/Arestas

```
<owl:ObjectProperty rdf:about="https://thedevconf.com/tdc/2021/HOC-4320#joga">
    <rdfs:domain rdf:resource="https://thedevconf.com/tdc/2021/HOC-4320/time#SaoPaulo"/>
    <rdfs:range rdf:resource="https://thedevconf.com/tdc/2021/HOC-4320#Partida"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="https://thedevconf.com/tdc/2021/HOC-4320#localizado">
    <rdfs:domain rdf:resource="https://thedevconf.com/tdc/2021/HOC-4320#Partida"/>
    <rdfs:range rdf:resource="https://thedevconf.com/tdc/2021/HOC-4320/cidade#SaoPaulo"/>
</owl:ObjectProperty>
```

Grafo De Conhecimento - W3C

THE DEVELOPER'S CONFERENCE

- Turtle é outra forma de serializar triplas.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix ex: <http://example.org/stuff/1.0/> .

<http://www.w3.org/TR/rdf-syntax-grammar>
  dc:title "RDF/XML Syntax Specification (Revised)" ;
  ex:editor [
    ex:fullname "Dave Beckett";
    ex:homePage <http://purl.org/net/dajobe/>
  ] .
```

Grafo De Conhecimento - SPARQL

- SPARQL Protocol and RDF Query Language (SPARQL) é utilizado para recuperar as triplas.
- Inspirado no Structured Query Language (SQL).
- Recuperação de grafos de conhecimento no DBPedia.

```
select ?pais ?PIBpercapita
where {
?pais dct:subject dbc:Countries_in_South_America.
?pais rdf:type dbo:Country.
?pais dbp:gdpNominalPerCapita ?PIBpercapita.
}
order by DESC(xsd:float(?PIBpercapita))  LIMIT 100
```

Grafo De Conhecimento - SPARQL

THE DEVELOPER'S CONFERENCE

```
select ?pais ?PIBpercapita
where {
  ?pais  dct:subject                dbc:Countries_in_South_America.
  ?pais  rdf:type                   dbo:Country.
  ?pais  dbp:gdpNominalPerCapita    ?PIBpercapita.
}
order by DESC(xsd:float(?PIBpercapita))  LIMIT 100
```

Tripla

Tripla

Tripla

Sujeito

Predicado

Objeto

Grafo De Conhecimento - SPARQL

THE DEVELOPER'S CONFERENCE

```
select ?pais ?PIBpercapita
where {
        ?pais  dct:subject                    dbc:Countries_in_South_America.
        ?pais  rdf:type                       dbo:Country.
        ?pais  dbp:gdpNominalPerCapita ?PIBpercapita.
}
order by DESC(xsd:float(?PIBpercapita))  LIMIT 100
```

As variáveis

variável

Grafo De Conhecimento - SPARQL

THE DEVELOPER'S CONFERENCE

- A consulta executada no DBPedia.

pais	PIBpercapita
http://dbpedia.org/resource/Uruguay	"17819.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Chile	"15854.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Argentina	"9887.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Guyana	"8649.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Suriname	"6881.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Colombia	"6744.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Brazil	"6450.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Paraguay	"6230.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Ecuador	"6155.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Peru	"5845.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Bolivia	"3823.0"^^<http://dbpedia.org/datatype/usDollar>
http://dbpedia.org/resource/Venezuela	"2299.0"^^<http://dbpedia.org/datatype/usDollar>

Grafo De Conhecimento - Named Graph

THE DEVELOPER'S CONFERENCE

- O named graph tem vários propósitos.
- As triplas são inseridas em um grafo nomeado.
- Exemplo no triplestore GraphDB

Grafo De Conhecimento - Named Graph

THE DEVELOPER'S CONFERENCE

```
PREFIX tdc: <https://thedevconf.com/tdc#>
INSERT DATA
{
  GRAPH <https://thedevconf.com/tdc/2020> {
      <https://thedevconf.com/tdc/2020/palestra/hoc-0000> tdc:autor "Fulano" ;
      tdc:palestra "Hello World" .
  }
  GRAPH <https://thedevconf.com/tdc/2021> {
      <https://thedevconf.com/tdc/2021/palestra/hoc-4320> tdc:autor "Rafael Rocha" ;
      tdc:palestra "Organize seus dados com grafos de conhecimento" .
  }
}
```

Grafo De Conhecimento - Named Graph

Named Graph

Triplas

	g	s	p	o
1	https://thedevconf.com/tdc/2020	https://thedevconf.com/tdc/2020/palestra/hoc-0000	https://thedevconf.com/tdc#autor	"Fulano"
2	https://thedevconf.com/tdc/2020	https://thedevconf.com/tdc/2020/palestra/hoc-0000	https://thedevconf.com/tdc#palestra	"Hello World"
3	https://thedevconf.com/tdc/2021	https://thedevconf.com/tdc/2021/palestra/hoc-4320	https://thedevconf.com/tdc#palestra	"Organize seus dados com grafos de conhecimento"
4	https://thedevconf.com/tdc/2021	https://thedevconf.com/tdc/2021/palestra/hoc-4320	https://thedevconf.com/tdc#autor	"Rafael Rocha"

Aplicabilidade 03

Aplicabilidade

- Organizar o conhecimento com reaproveitamento de vocabulário.
- Facilidade na integração (semântica) de dados.
- Manter proveniência dos dados.

  Quem gerou/alterou. Qual fonte/dispositivo. Onde.
- Racionador semântico (semantic reasoner) é um software que gera deduções a partir de uma base dados. Mantém a coesão com os axiomas ontológicos e axiomas criados por você.

Aplicabilidade - Mapeamento

- O mapeamento gera grafos de conhecimento a partir de outros formatos.
- Cada ferramenta possui uma abordagem distinta.

Aplicabilidade - Mapeamento

THE DEVELOPER'S CONFERENCE

- R2RML [R2RML, 2021]

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix ex: <http://example.com/ns#>.

<#TriplesMap1>
    rr:logicalTable [ rr:tableName "EMP" ];
    rr:subjectMap [
        rr:template "http://data.example.com/employee/{EMPNO}";
        rr:class ex:Employee;
    ];
    rr:predicateObjectMap [
        rr:predicate ex:name;
        rr:objectMap [ rr:column "ENAME" ];
    ].
```

Aplicabilidade - Mapeamento

THE DEVELOPER'S CONFERENCE

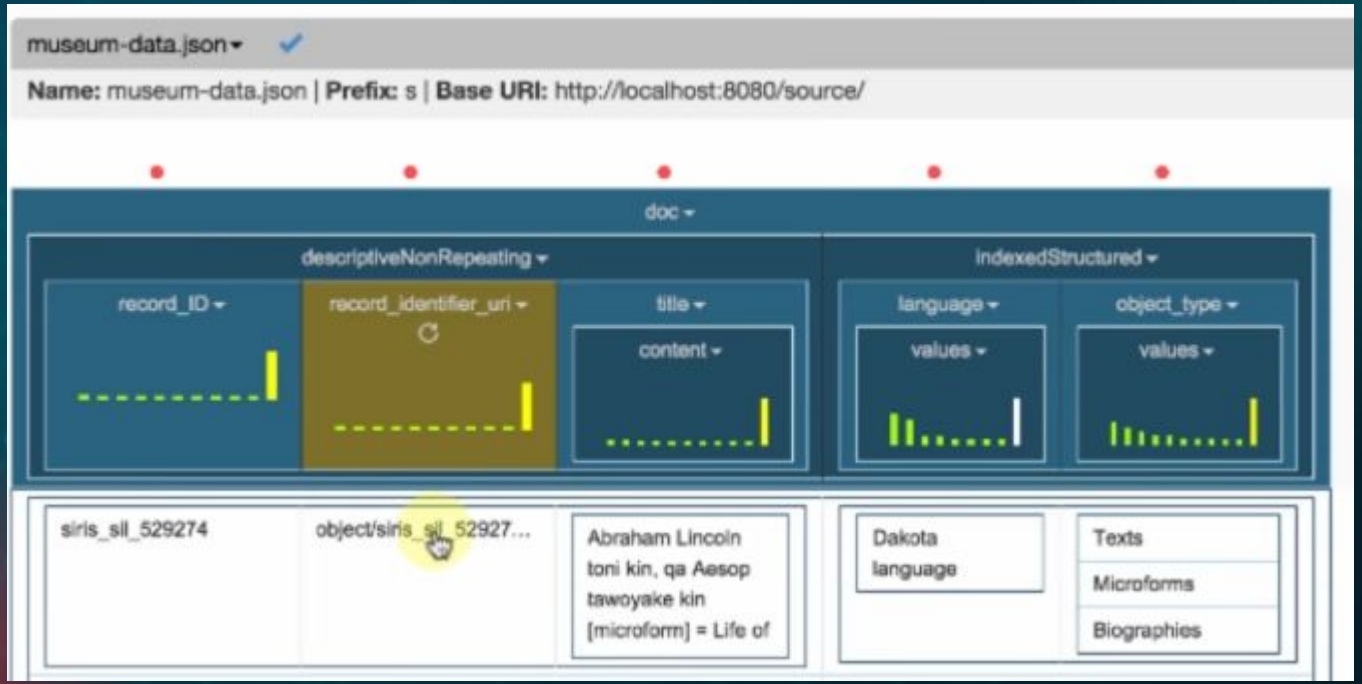
- RML [RML, 2021]

```
<#AirportMapping> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "Airport.csv" ;
    rml:referenceFormulation ql:CSV
  ];
  rr:subjectMap [
    rr:template "http://airport.example.com/{id}";
    rr:class transit:Stop
  ];

  rr:predicateObjectMap [
    rr:predicate transit:route;
    rr:objectMap [
      rml:reference "stop";
      rr:datatype xsd:int
      ]
    ];

  rr:predicateObjectMap [
    rr:predicate wgs84_pos:lat;
    rr:objectMap [
      rml:reference "latitude"
    ]
  ];
```

Aplicabilidade - Mapeamento

THE DEVELOPER'S CONFERENCE

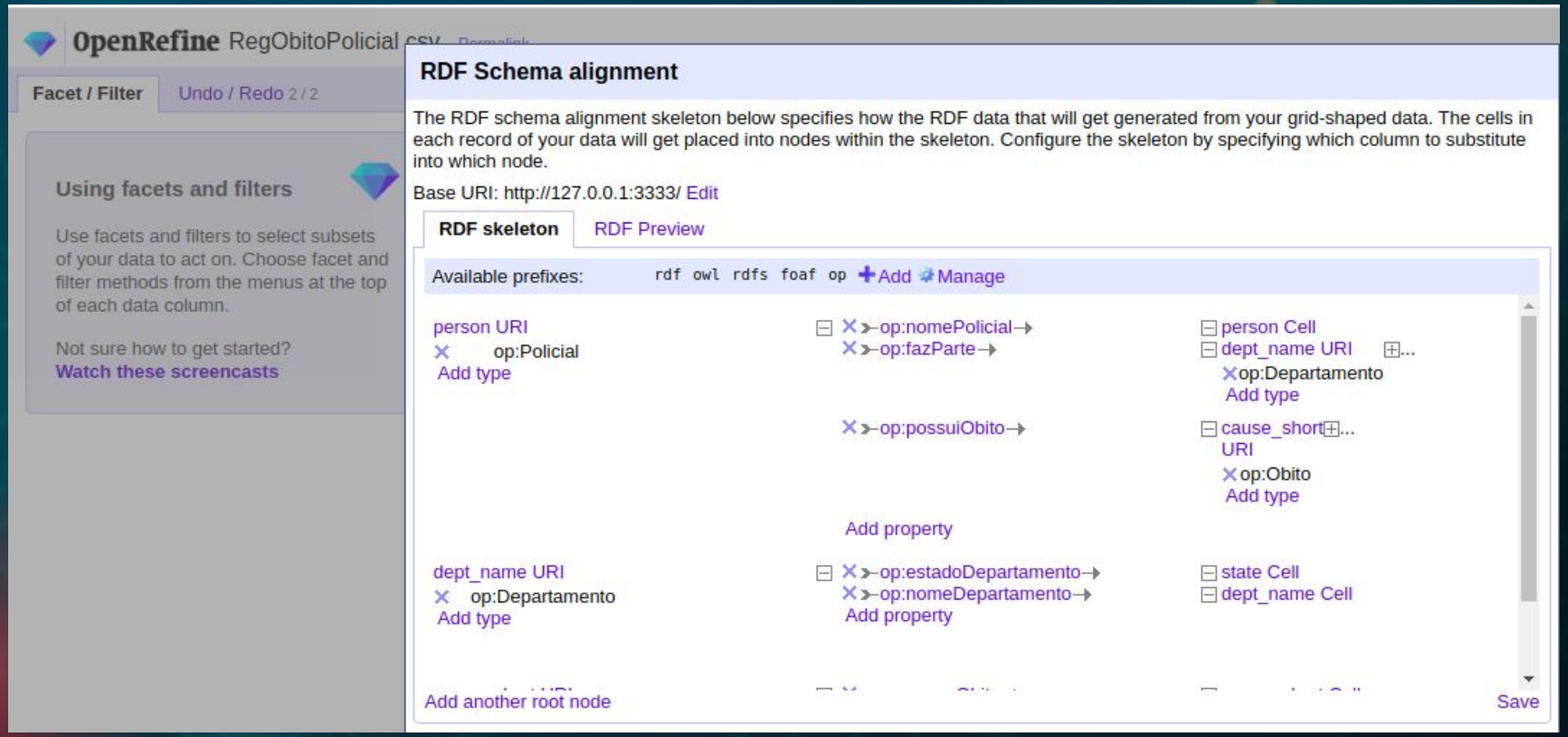
- Karma [KARMA, 2021]



Aplicabilidade - Mapeamento

THE DEVELOPER'S CONFERENCE

- OpenRefine+RDF Extension [OPEN, 2021]



Aplicabilidade - Linguagens

THE DEVELOPER'S CONFERENCE

R SPARQL [SPARQLR, 2021]

```
library(SPARQL) # SPARQL querying package
library(ggplot2)

# Step 1 - Set up preliminaries and define query
# Define the data.gov endpoint
endpoint <- "http://services.data.gov/sparql"

# create query statement
query <-
"PREFIX  dgp1187: <http: data-gov.tw.rpi.edu="" vocab="" p="" 1187="">
SELECT ?ye ?fi ?ac
WHERE {
?s dgp1187:year ?ye .
?s dgp1187:fires ?fi .
?s dgp1187:acres ?ac .
}"

# Step 2 - Use SPARQL package to submit query and save results to a data frame
qd <- SPARQL(endpoint,query)
df <- qd$results
```

Aplicabilidade - Linguagens

THE DEVELOPER'S CONFERENCE

⇢ Python SPARQL [SPARQLP, 2021]

```
>>> import sparql
```

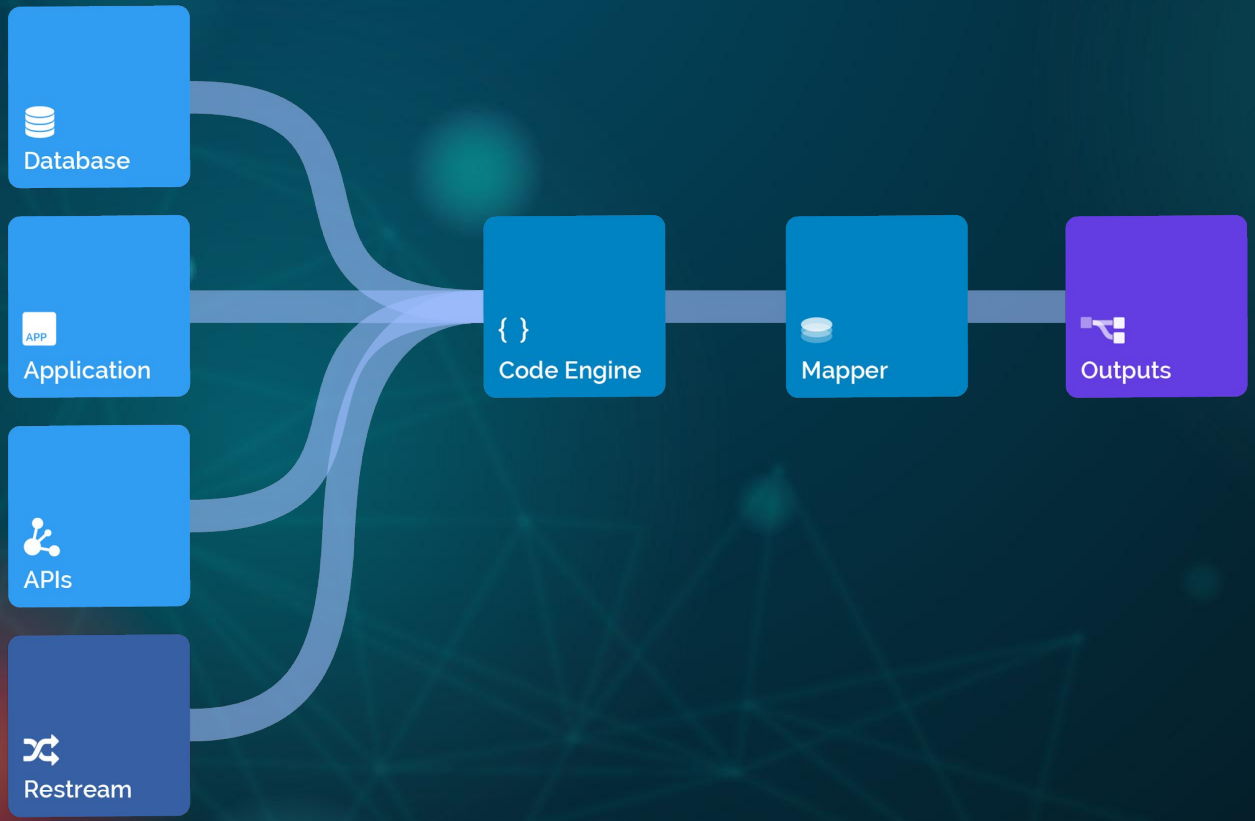
```
>>> q = ('SELECT DISTINCT ?station, ?orbits WHERE { '
...      '?station a <http://dbpedia.org/ontology/SpaceStation> . '
...      '?station <http://dbpedia.org/property/orbits> ?orbits . '
...      'FILTER(?orbits > 50000) } ORDER BY DESC(?orbits)')
>>> result = sparql.query('http://dbpedia.org/sparql', q)
```

```
>>> result.variables
[u'station', u'orbits']
```

Aplicabilidade

THE DEVELOPER'S CONFERENCE

Fim do ETL



Aplicabilidade

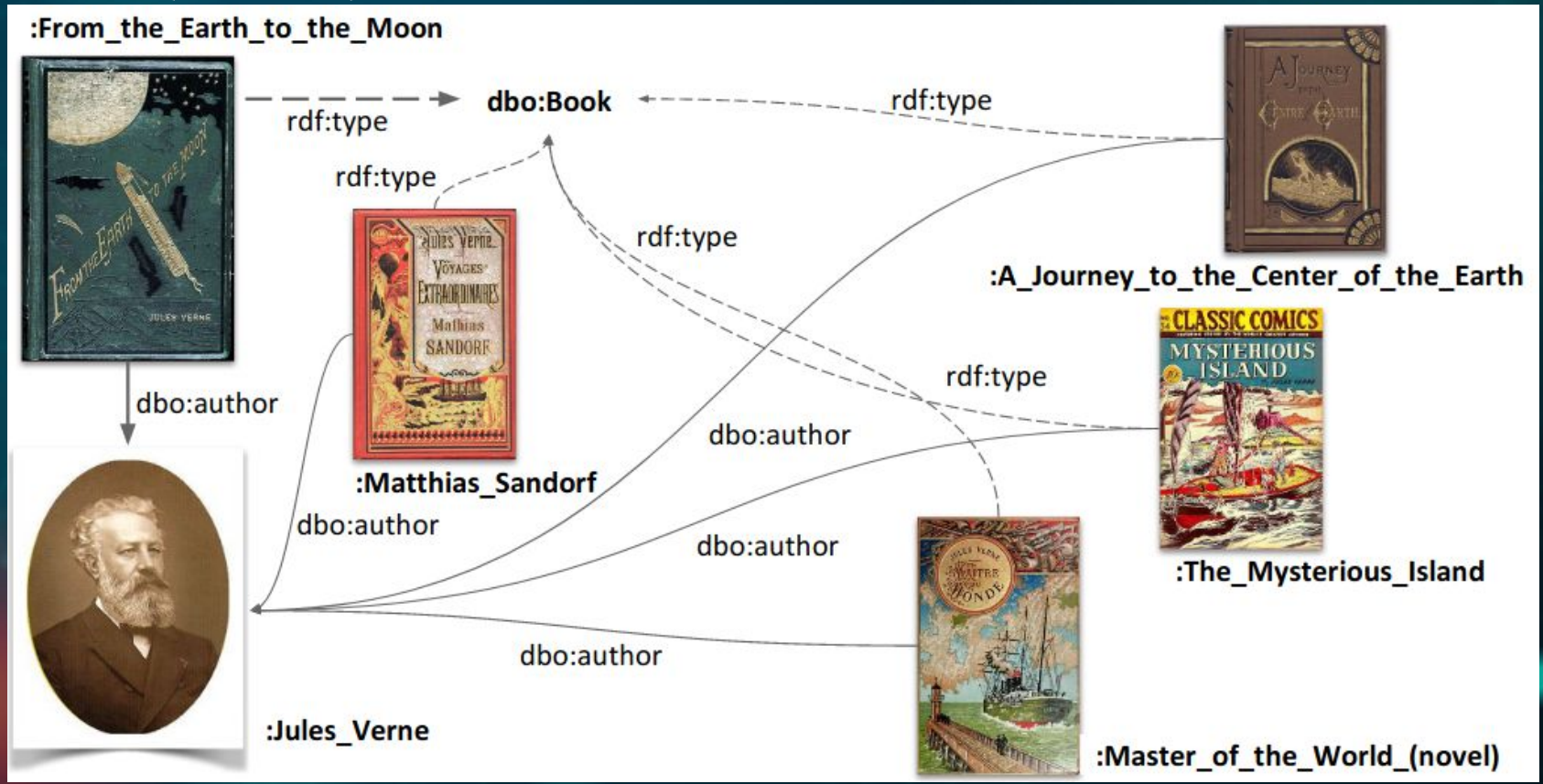
THE DEVELOPER'S CONFERENCE

- Bem vindo a ingestão de dados

Aplicabilidade

THE DEVELOPER'S CONFERENCE

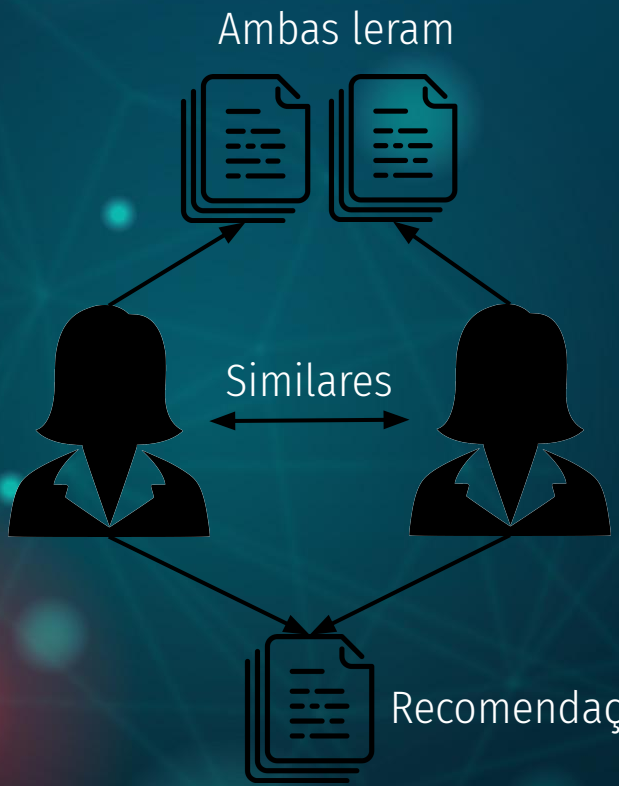
Pesquisa exploratória [HPI, 2021]

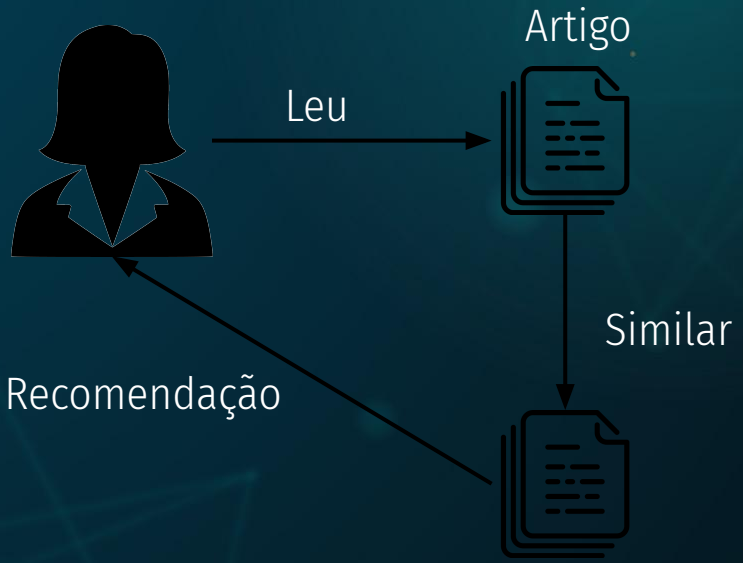


Aplicabilidade

THE DEVELOPER'S CONFERENCE

- Sistema de recomendações





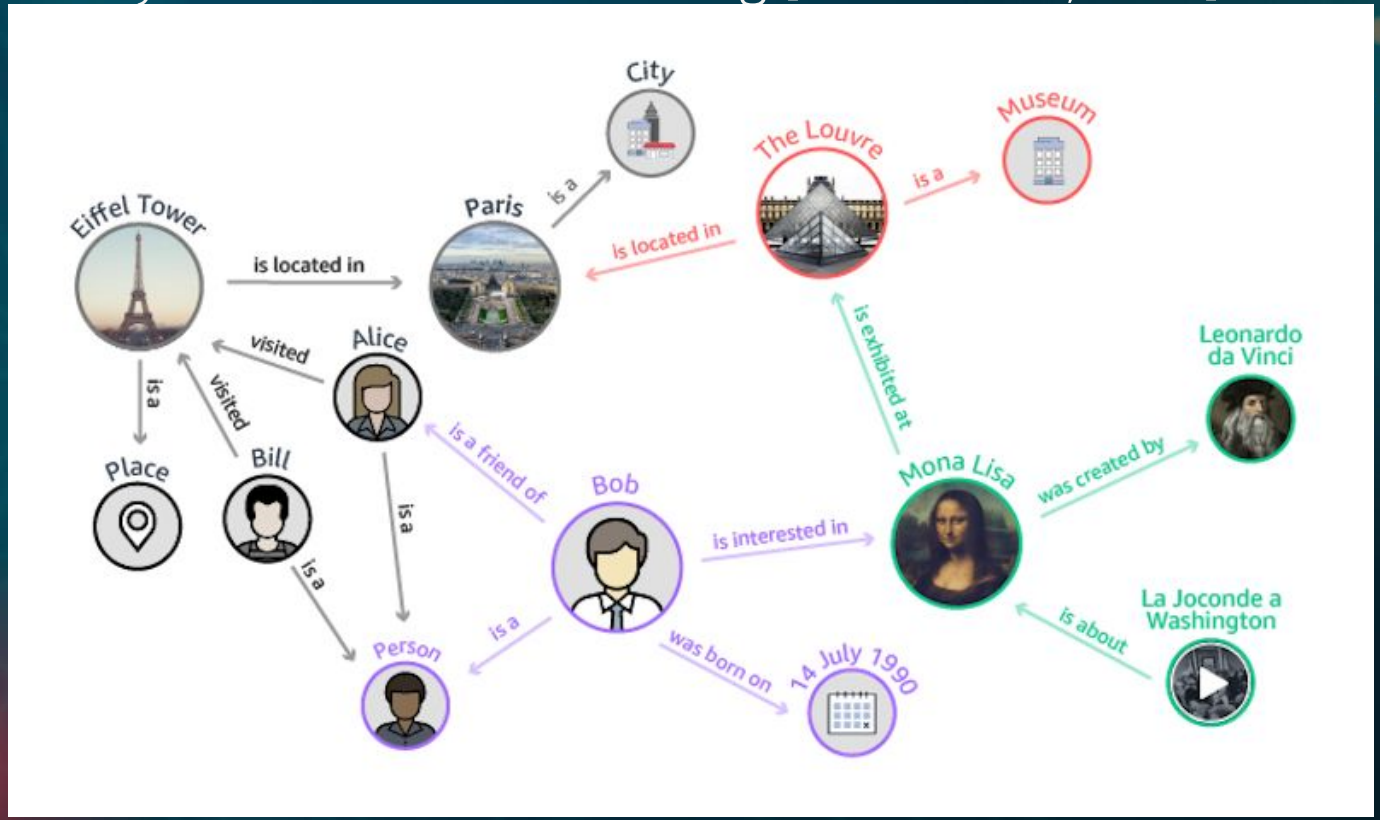
Aplicabilidade

THE DEVELOPER'S CONFERENCE

- Detecção de fraude [FRAUD, 2021]

Aplicabilidade

- Analytics e Machine Learning [NODE2VEC, 2021]



Considerações Finais

04

Considerações Finais

- Grafo de conhecimento é forma eficiente de manter dados semânticos.
- Existem diversos bancos de dados orientados a grafos ou triplestores.
- Triple stores disponibiliza API para a recuperação das triplas.
- Fim dos pipelines de ETL
- Dados para Analytics, Sistema de recomendação, ML, IA...

REFERÊNCIAS

[ANHØJ, 2003] ANHØJ, Jacob. Generic design of web-based clinical databases. Journal of Medical Internet Research, v. 5, n. 4, p. e27, 2003.
[FRAUD, 2021] https://neo4j.com/blog/financial-services-neo4j-fraud-detection/
[HPI, 2021] https://open.hpi.de/courses/knowledgegraphs2020
[NODE2VEC, 2021] https://www.kaggle.com/ferdzso/knowledge-graph-analysis-with-node2vec
[HORRIDGE, 2009] HORRIDGE, Matthew et al. A practical guide to building owl ontologies using protégé 4 and co-ode tools edition 1. 2. The university of Manchester, v. 107, 2009.
[KARMA, 2021] - https://usc-isi-i2.github.io/karma/
[OPEN, 2021] - https://openrefine.org/
[OWL, 2002] - https://www.w3.org/TR/2002/WD-owl-absyn-20020729/
[RDF, 1997] - https://www.w3.org/TR/WD-rdf-syntax-971002/
[RDFS, 1999] - https://www.w3.org/TR/1999/PR-rdf-schema-19990303/
[RML, 2021] - https://rml.io/specs/rml/
[R2RML, 2021]- https://www.w3.org/TR/r2rml/
[SPARQLP, 2021] - https://pypi.org/project/sparql-client/
[SPARQLR, 2021] - https://www.r-bloggers.com/2013/01/sparql-with-r-in-less-than-5-minutes/
[STAR, 2021] - https://en.wikipedia.org/wiki/Star_schema
[TOWARDS, 2021] - https://towardsdatascience.com/data-analysis-in-mysql-operators-joins-and-more-in-relational-databases-26c0a968e61e

Amazon Neptune - https://aws.amazon.com/pt/neptune/
DBPedia - https://dbpedia.org/sparql/
GraphDB - https://graphdb.ontotext.com/
Neo4J - https://neo4j.com
OrientDB - https://orientdb.org/
OpenLink Virtuoso - https://virtuoso.openlinksw.com/

THANKS!

Do you have any questions?

Rafael-rocha |at| ufmg |dot| br

https://www.linkedin.com/in/rafael-rocha-12974226/