ANT The System Architecture of an Academic Search Engine

José Devezas
joseluisdevezas@gmail.com

INESC TEC & FEUP InfoLab
MAP-i 2016/2017

Universidade do Minho
universidade de aveiro
U.PORTO

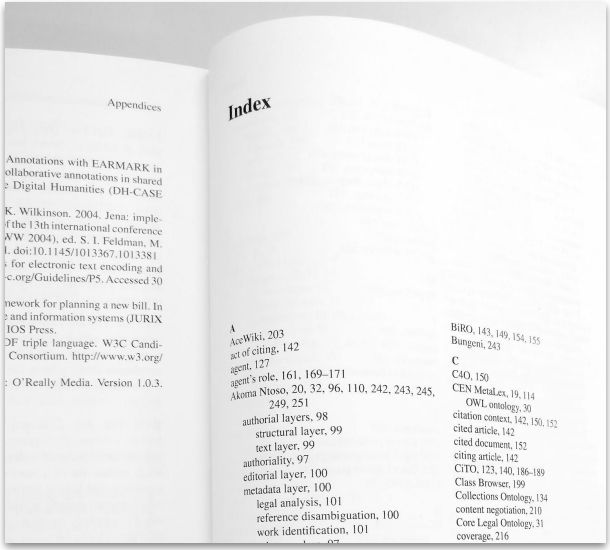
Contents

Introduction

What is entity-oriented search?

Search using keyword-based matching

- Modeled after the back-of-the-book index.
- Finding relevant content involves:
  1. Selecting one or several keywords;
  2. Jumping to the indicated pages;
  3. Reading passages and using knowledge, either internal or external to the book, to assess the relevance.



Appendices

Annotations with EARMARK in
ollaborative annotations in shared
e Digital Humanities (DH-CASE

K. Wilkinson. 2004. Jena: imple-
f the 13th international conference
AW 2004), ed. S. I. Feldman, M.
I. doi:10.1145/1013367.1013381
for electronic text encoding and
c.org/Guidelines/P5. Accessed 30

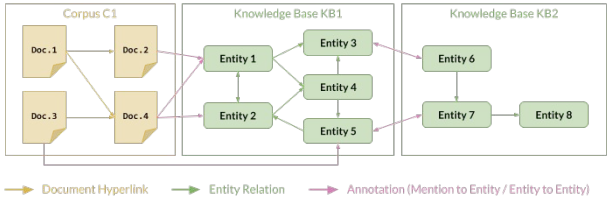
nework for planning a new bill. In
e and information systems (JURIX
IOS Press.
DF triple language. W3C Candi-
Consortium. http://www.w3.org/

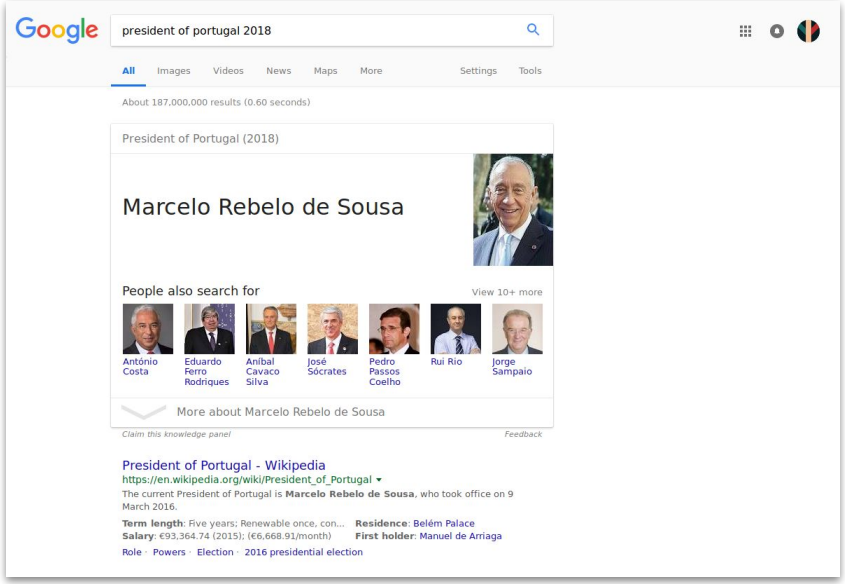
: O'Really Media. Version 1.0.3.

Index

Search using semantic matching

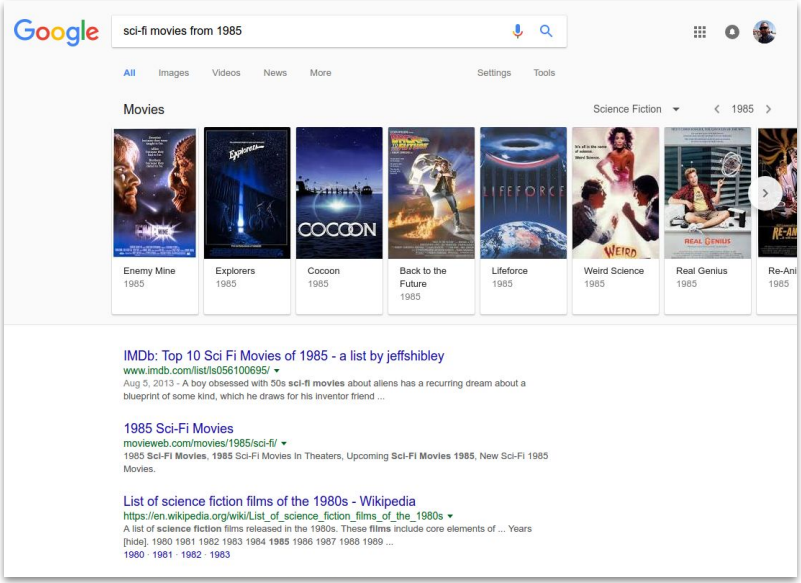
- Closer to the user's information need.
- Requires interpretation of query meaning and document semantics.
- And the combination of unstructured and structured data from corpora and knowledge bases.

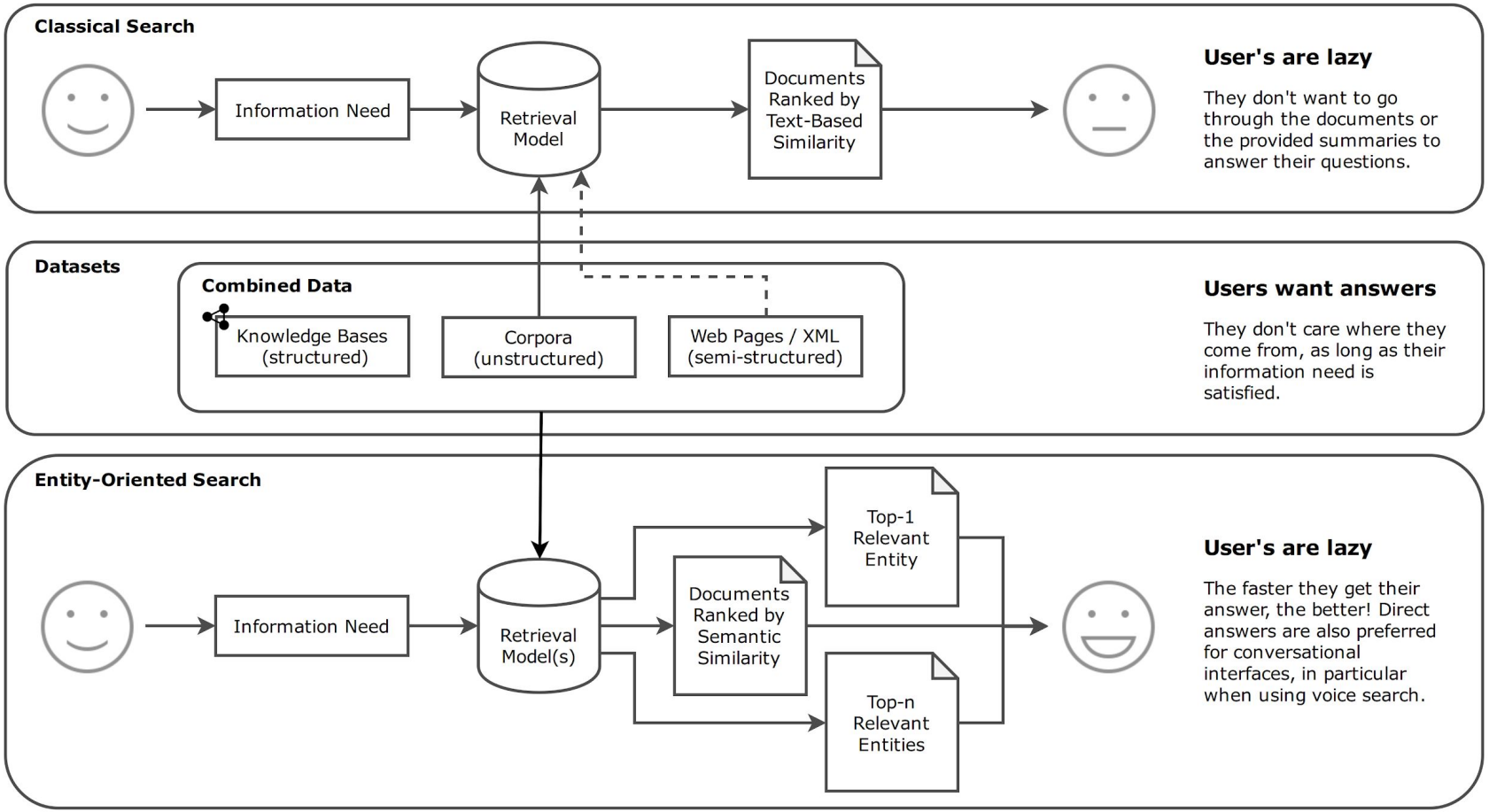




Search using semantic matching

- It becomes possible to, more adequately, answer queries like:
  - `[ president of portugal 2018 ]`
  - `[ sci-fi movies from 1985 ]`
- In addition to ranking documents containing the keywords;
- An entity or list of entities is directly provided as the answer.




Classical Search
Information Need
Retrieval Model
Documents Ranked by Text-Based Similarity
User's are lazy
They don't want to go through the documents or the provided summaries to answer their questions.
Datasets
Combined Data
Knowledge Bases (structured)
Corpora (unstructured)
Web Pages / XML (semi-structured)
Users want answers
They don't care where they come from, as long as their information need is satisfied.
Entity-Oriented Search
Information Need
Retrieval Model(s)
Documents Ranked by Semantic Similarity
Top-1 Relevant Entity
Top-n Relevant Entities
User's are lazy
The faster they get their answer, the better! Direct answers are also preferred for conversational interfaces, in particular when using voice search.

Requirements of an EOS engine

- Integrate text, entities and their relations.
  - Where to integrate?
    - Index level;
    - Ranking level.
  - What are our technological options?
    - Inverted index;
      e.g., Apache Lucene.
    - Triple store;
      e.g., OpenLink Virtuoso.
    - Hybrid technology?
      Higher-risk: still being researched.
- Match query and documents/entities using all available information for ranking.
  - Query segmentation and semantic tagging;
    NER in query / linking entities to the knowledge base.
  - Document annotation.
    NER in documents / linking entities to the knowledge base.

ANT

Search engine architecture.

An academic EOS engine

Depending on the query, results can be:

- Documents*
  - Retrieved using semantic information (entities and their relations).
- Entities
  - A specific one, a list, or both.
  - Retrieved by name, type, or another description.
  - Representing attributes or relations.

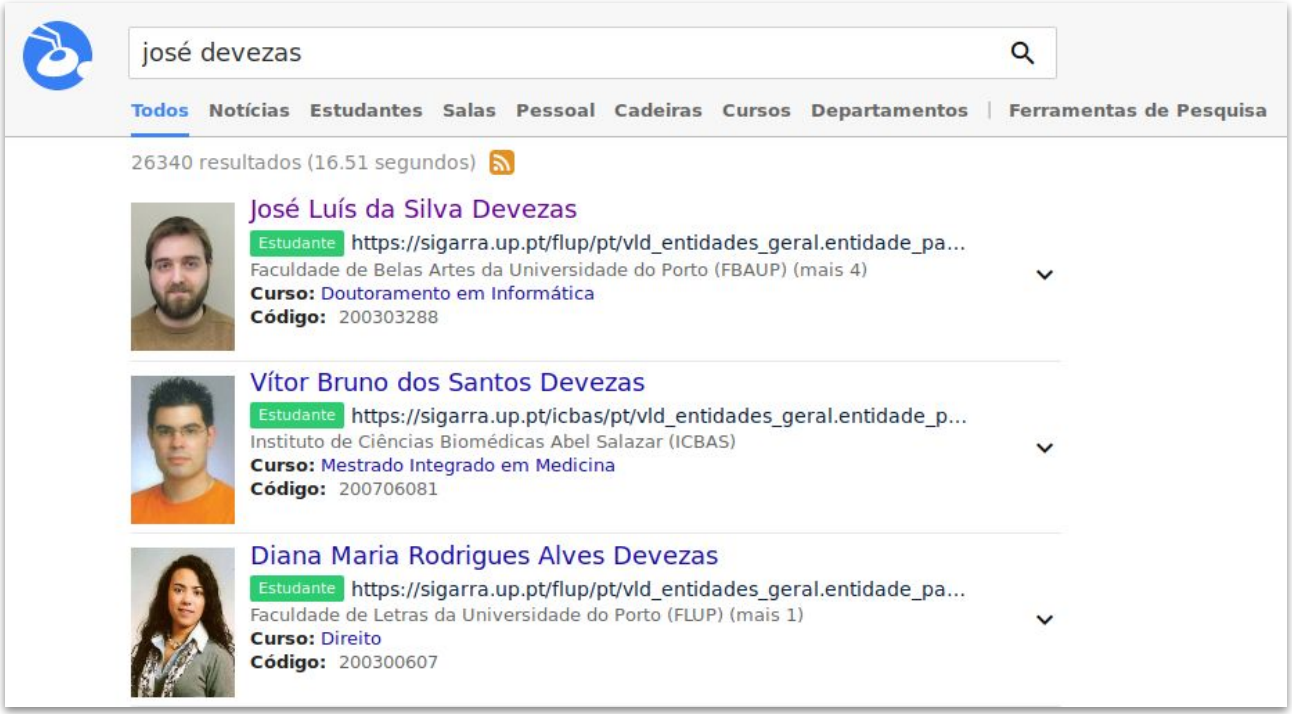
Entities can be:

- Students
- Staff
- Departments
- Rooms
- Curricular Units
- Courses
- News*

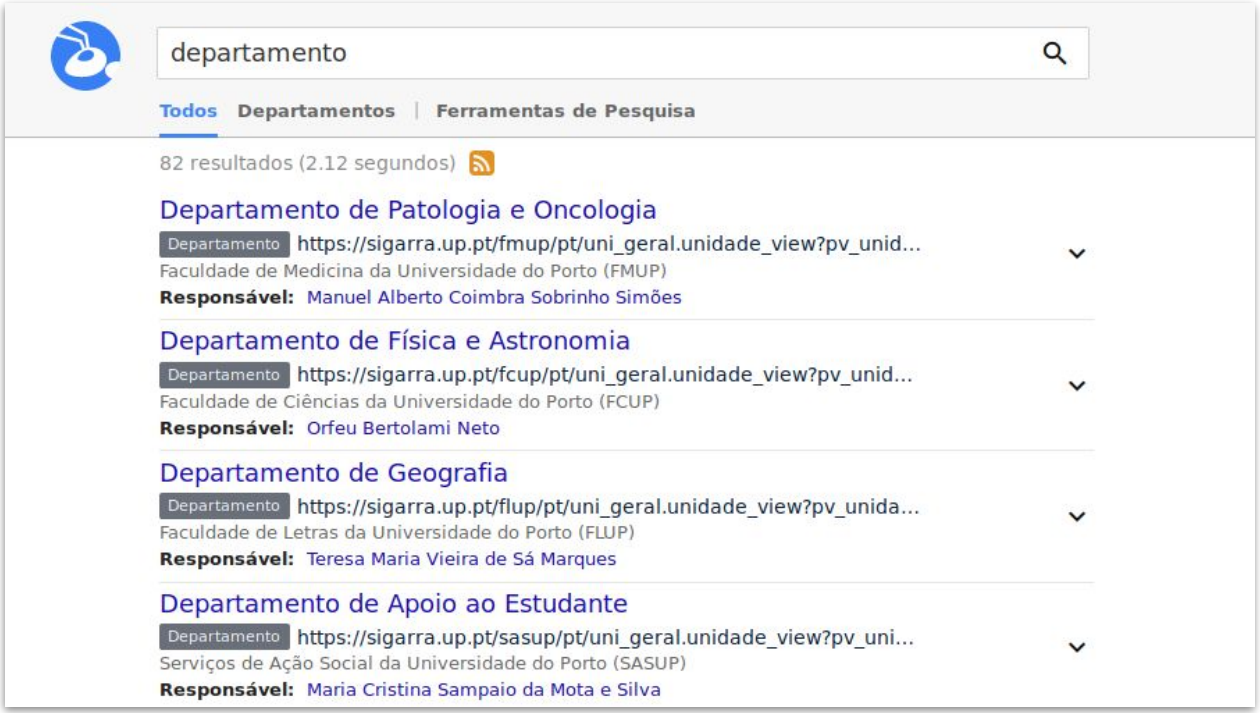
ANT

Ad hoc search of eNtities and Text.

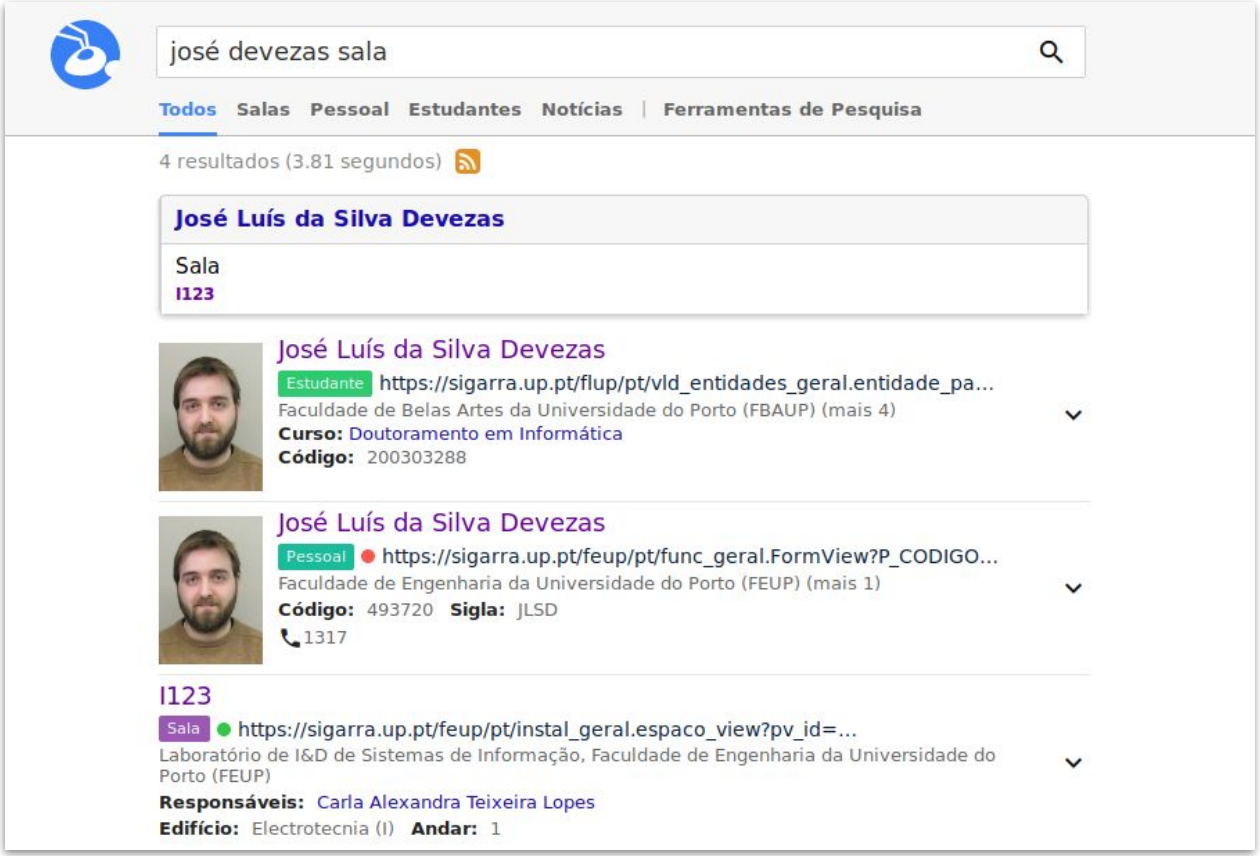
- Designed to support the five query categories, as defined by Pound et al. (2010):
  - Entity;
  - Type;
  - Attribute;
  - Relation;
  - Keyword.
- Based on two Lucene indexes:
  - Query analysis index;
  - Entity index.
- And a Virtuoso quad store:
  - Useful for attribute and relation queries.



Entity query. *The intention of the query is to find an entity.*



Type query. *The intention of the query is to find entities of a given type or class.*



Attribute query. *The intention of the query is to find values for a given attribute of a particular entity or type.*

ANT Pesquisa Orientada a Entidades Início Sobre Colaboração Contactos

josé devezas tiago devezas Pesquisar ?

6 resultados obtidos em 2,88 segundos

José Luís da Silva Devezas
Faculdade de Belas Artes da Universidade do Porto (FBAUP), Faculdade de Letras da Universidade do Porto (FLUP), Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Economia da Universidade do Porto (FEP), Faculdade de Ciências da Universidade do Porto (FCUP)
Estudante
score = 0.002238

Tiago Nuno Mesquita Folgado leitão Devezas
Faculdade de Engenharia da Universidade do Porto (FEUP) Domingo, 07 Outubro 2018, 00h32
Pessoal
score = 0.001971

José Luís da Silva Devezas
Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Ciências da Universidade do Porto (FCUP) Domingo, 02 Setembro 2018, 00h26
Pessoal
score = 0.001106

Tiago Nuno Mesquita Folgado Leitão Devezas
Faculdade de Belas Artes da Universidade do Porto (FBAUP), Faculdade de Letras da Universidade do Porto (FLUP), Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Economia da Universidade do Porto (FEP), Faculdade de Ciências da Universidade do Porto (FCUP)
Estudante
score = 0.001789

I123
Faculdade de Engenharia da Universidade do Porto (FEUP)
Sala
score = 0.000431

Projetos no laboratório SAPO/U.Porto
Faculdade de Engenharia da Universidade do Porto (FEUP) Sexta-feira, 28 Junho 2013, 00h00
Notícia
score = 0.0

<< Anterior Página 1 de 1 Seguinte >>

Entidades

Tiago Nuno Mesquita Folgado leitão Devezas

José Luís da Silva Devezas

Ligações

Faculdade: Faculdade de Engenharia, Faculdade de Engenharia da Universidade do Porto

Sala: I123

Projetor	Não
Edifício	Electrotecnia (I)
Mapa	https://sigarra.up.pt/feup/pt/instal_geral2.get_mapa?pv_id=77486

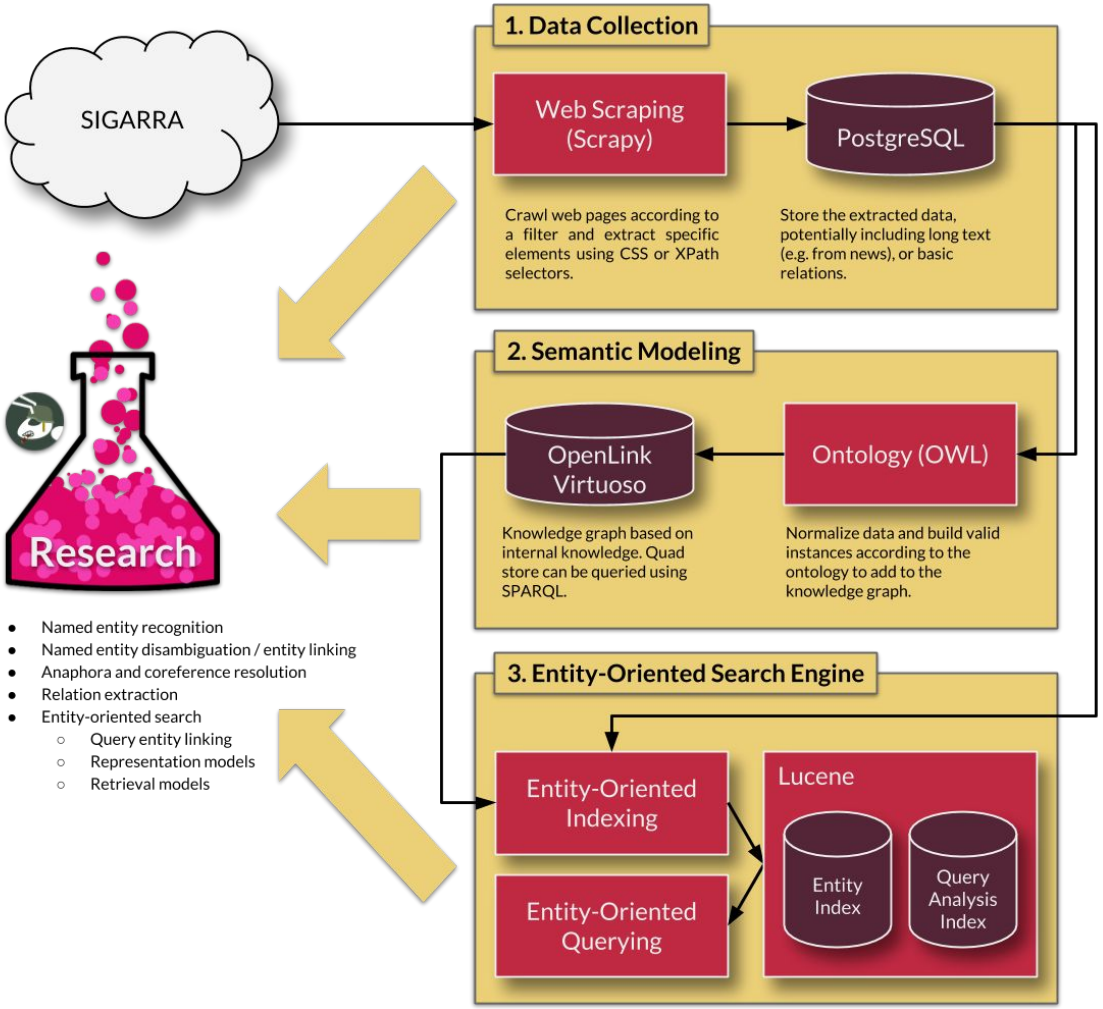
Ver mais

Relation query. *The intention of the query is to find how two or more entities or types are related.*

Keyword query. *Anything that doesn't fit the previous four categories.*

Search engine architecture

ANT components, from data collection to search.


SIGARRA
1. Data Collection
Web Scraping (Scrapy)
PostgreSQL
Crawl web pages according to a filter and extract specific elements using CSS or XPath selectors.
Store the extracted data, potentially including long text (e.g. from news), or basic relations.
2. Semantic Modeling
OpenLink Virtuoso
Ontology (OWL)
Knowledge graph based on internal knowledge. Quad store can be queried using SPARQL.
Normalize data and build valid instances according to the ontology to add to the knowledge graph.
3. Entity-Oriented Search Engine
Entity-Oriented Indexing
Entity-Oriented Querying
Lucene
Entity Index
Query Analysis Index
Research
Named entity recognition
Named entity disambiguation / entity linking
Anaphora and coreference resolution
Relation extraction
Entity-oriented search
Query entity linking
Representation models
Retrieval models

1. Data Collection

Scrapy

Specify:
- Allowed domains (up.pt)
- List of start URLs (all SIGARRAs)
- LinkExtractor rules

Crawl:
- Launch crawlers from start URLs
- Fetch and parse web pages
- Extract URLs from anchors
- URLs matching rules will be scraped
- Follow uncrawled up.pt URLs

Scrape:
- Manually define XPath and CSS selectors for elements to be stored in the model.

Spiders → Pipelines → Models → Tables: One per entity + auxiliary tables → PostgreSQL

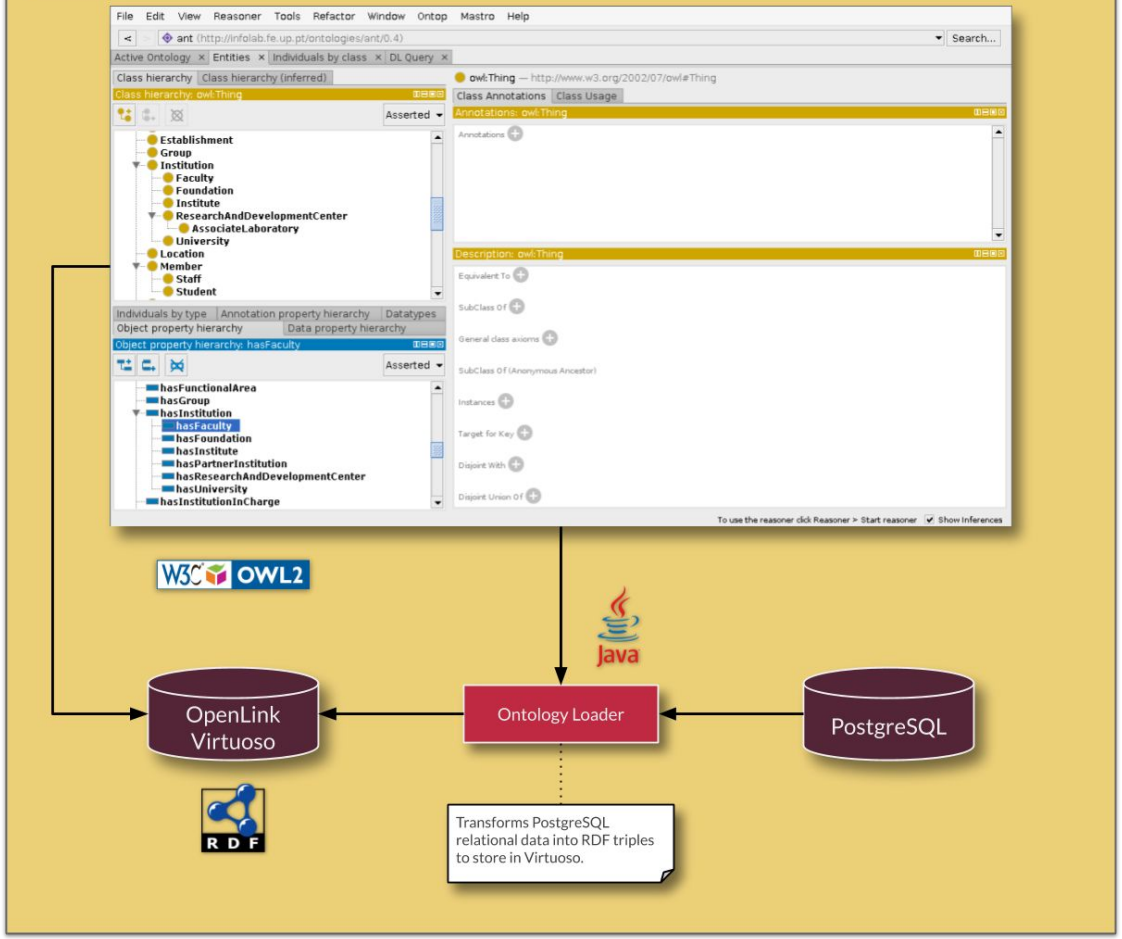
Materialized view: lucene_documents → PostgreSQL

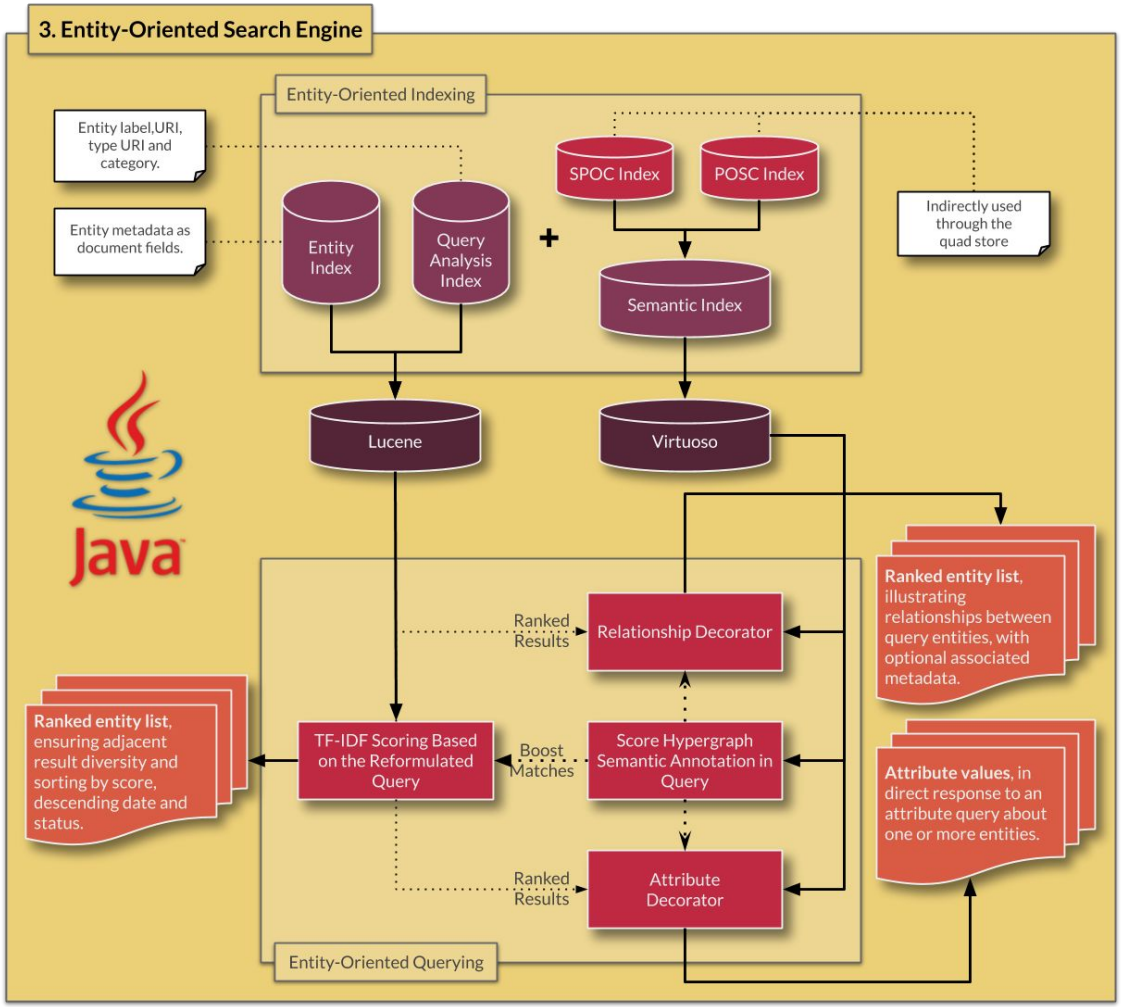
Transform an HTTP **response** into an **item** through predefined CSS and XPath selectors.

Transform an **item** from a spider into an **object** from the a model, and deal with storage operations.

e.g., object-relational mapping (ORM)

2. Semantic Modeling




3. Entity-Oriented Search Engine
Entity-Oriented Indexing
Entity label,URI, type URI and category.
Entity metadata as document fields.
Entity Index
Query Analysis Index
+
SPOC Index
POSC Index
Semantic Index
Indirectly used through the quad store
Lucene
Virtuoso
Java
Ranked Results
Relationship Decorator
Ranked entity list, illustrating relationships between query entities, with optional associated metadata.
Ranked entity list, ensuring adjacent result diversity and sorting by score, descending date and status.
TF-IDF Scoring Based on the Reformulated Query
Boost Matches
Score Hypergraph Semantic Annotation in Query
Attribute values, in direct response to an attribute query about one or more entities.
Ranked Results
Attribute Decorator
Entity-Oriented Querying

REST API

- ANT provides access to search-related services via a REST API.
- We use the OpenAPI 2.0 format (Swagger) to document the API.
  - http://ant.fe.up.pt/api/
  - https://swagger.io/specification/
- Which makes it possible to easily provide a console for API exploration.
  - http://ant.fe.up.pt/api-console/

- Supported services are classified into six categories:
  - Analytics
  - Autocomplete
  - Decorators*
  - JavaScript
  - Log
  - Search*

* Critical services.

Conclusions

Final remarks and related projects.

Final remarks

- The ANT search engine is serving the local academic community, despite infrastructure and human resource limitations (it's a prototype).
- At the same time, it collects implicit relevance feedback, based on result clicks for issued queries.
- ANT is also a platform of collaboration for multiple areas of research:
  - Web Design
    Collaboration with MM for the development of the front-end.
  - User Experience
    MM dissertation in entity-oriented search interfaces.
  - Information Extraction
    MIEIC dissertation in named entity recognition for portuguese web text.

Related projects

Army ANT

- Serving the research needs in the area of entity-oriented search.
- Supporting the study of innovative ideas in search, providing tools for exploration and evaluation.

PhD thesis

- “Graph-Based Entity-Oriented Search”
  - Joint representation of text, entities and their relations.
  - Generalization of entity-oriented search tasks.
  - Improvement of search effectiveness?
- Exploration of random walks in graphs and hypergraphs.

Thank you!

https://ant.fe.up.pt

José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program Human Capital (POCH), supported by the European Social Fund and by national funds from MCTES.