Keyword Search over COVID-19 Data

Yenier T. Izquierdo[1,2], Grettel M. García[2], Melissa Lemos[1,2], Alexandre Novello[1], Bruno Novelli[2], Cleber Damasceno[2], Luiz André P.P. Leme[3], *Marco A. Casanova*[1,2]

[1]Department of Informatics, PUC-Rio, Rio de Janeiro, RJ – Brazil

[2]TecGraf Institute, PUC-Rio, Rio de Janeiro, RJ – Brazil

[3]Institute of Computing, UFF, Niterói, RJ – Brazil

Agenda

- Motivation
- DANKE
- *CovidKeys*
- Conclusions

Motivation

- Examples of COVID-19 data
  - FAPESP / USP / Instituto Fleury, H. Sírio-Libanês and H. Israelita Albert Einstein
    - The COVID-19 Data Sharing / BR initiative
  - Ministry of Health/Brazil
    - NSG (*Notificações de Síndrome Gripal*)
    - SRAG 2020 - Acute Respiratory Syndrome (SARS) and COVID-19 data
  - The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
    - COVID-19 Data Repository
  - Our World in Data – Global Change Data Lab / Oxford Martin School
    - Coronavirus pandemic: daily updated research and data
  - The Google Cloud services
    - The COVID-19 Public Datasets program

Motivation

- COVID-19 data
  - Data is typically available...
    - as downloadable CSV files
    - as datasets that can be queried through a user interface (e.g., Elastic Search)
    - as datasets available through a "sandbox" (e.g., Google Cloud services)
  - Files can be quite large, with hundreds of columns and millions of lines
  - Data can be at different levels of granularity

Motivation

- COVID-19 data
  - Cumbersome to process with the usual desktop spreadsheet tools
  - A perhaps more robust approach:
    - Download and store the data in a standard DBMS
    - Define a query that retrieves the data the user is interested in
    - Export the query results to a data analysis tool

Motivation

- COVID-19 data
  - Cumbersome to process with the usual desktop spreadsheet tools
  - A perhaps more robust approach:
    - Download and store the data in a standard DBMS
    - Define a query that retrieves the data the user is interested in ← **Keyword search**
    - Export the query results to a data analysis tool

Motivation

- Keyword search over databases
  - the user specifies a few terms, called *keywords*
    - A *keyword query* is simply a list of keywords
  - the system must retrieve the data that best match the list of keywords
    - Keyword search has been expanded to relational databases (and RDF datasets)
    - Keyword queries are typically compiled into SQL (or SPARQL)

Agenda

- Motivation
- DANKE
- *CovidKeys*
- Conclusions

DANKE

- What is DANKE?
  - a platform for data and knowledge retrieval

- Main Components
  - Data Ingestion and Indexing Services
  - Keyword Query Processing Service
    - Compiles a keyword query into an SQL (or SPARQL) query
    - Uses the database schema to guide the process
  - …

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA ECONOMIA
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512020001044-0**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 15/05/2020, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

Título: DANKE - Data and Knowledge Retrieval Platform

Data de publicação: 15/05/2020

Data de criação: 15/05/2020

Titular(es): FACULDADES CATÓLICAS

Autor(es): MARCO ANTÔNIO CASANOVA; GRETTEL MONTEAGUDO GARCIA; YENIER TORRES IZQUIERDO; MELISSA LEMOS CAVALIERE; BRUNO ALVES NÓBELLI; CLEBER OLIVEIRA DAMASCENO

Linguagem: JAVA; OUTROS

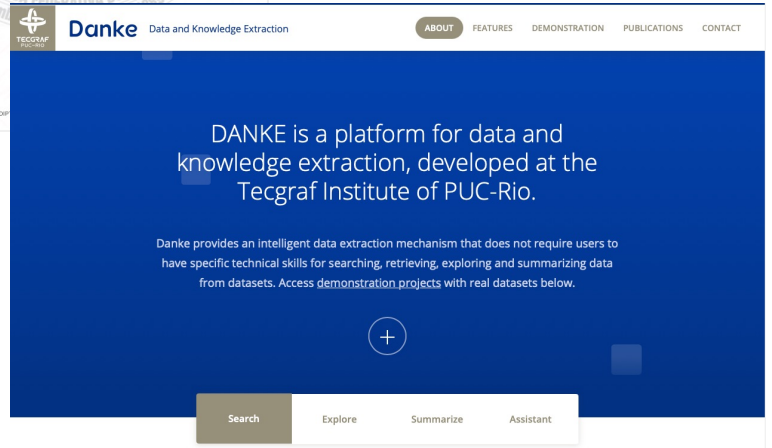
Campo de aplicação: IF-07

Tipo de programa: GI-01; GI-08; TC-01

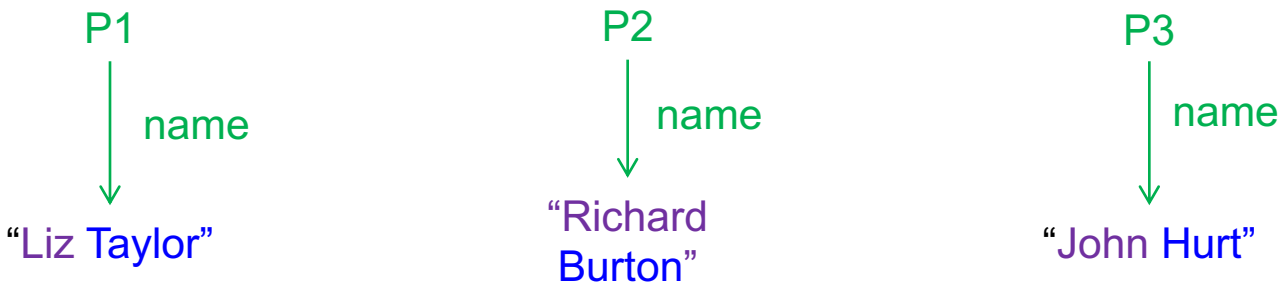
Algoritmo hash: SHA-256

Resumo digital hash: 4e6e5c4cc38fcc4c4f46f42fb29071cb39af1d17e70be530db02be81178c82bc34

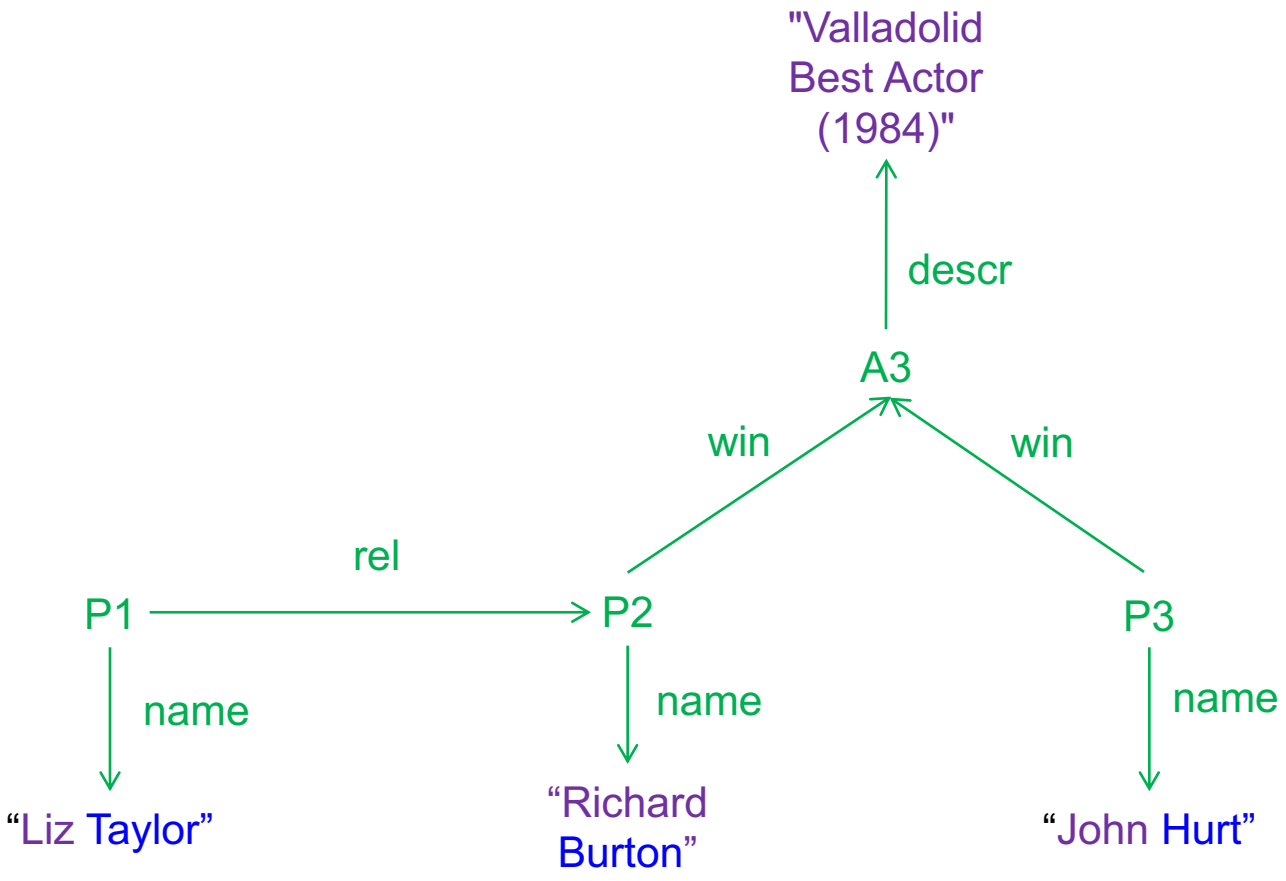
Expedido em: 16/06/2020



Query: Taylor Burton Hurt

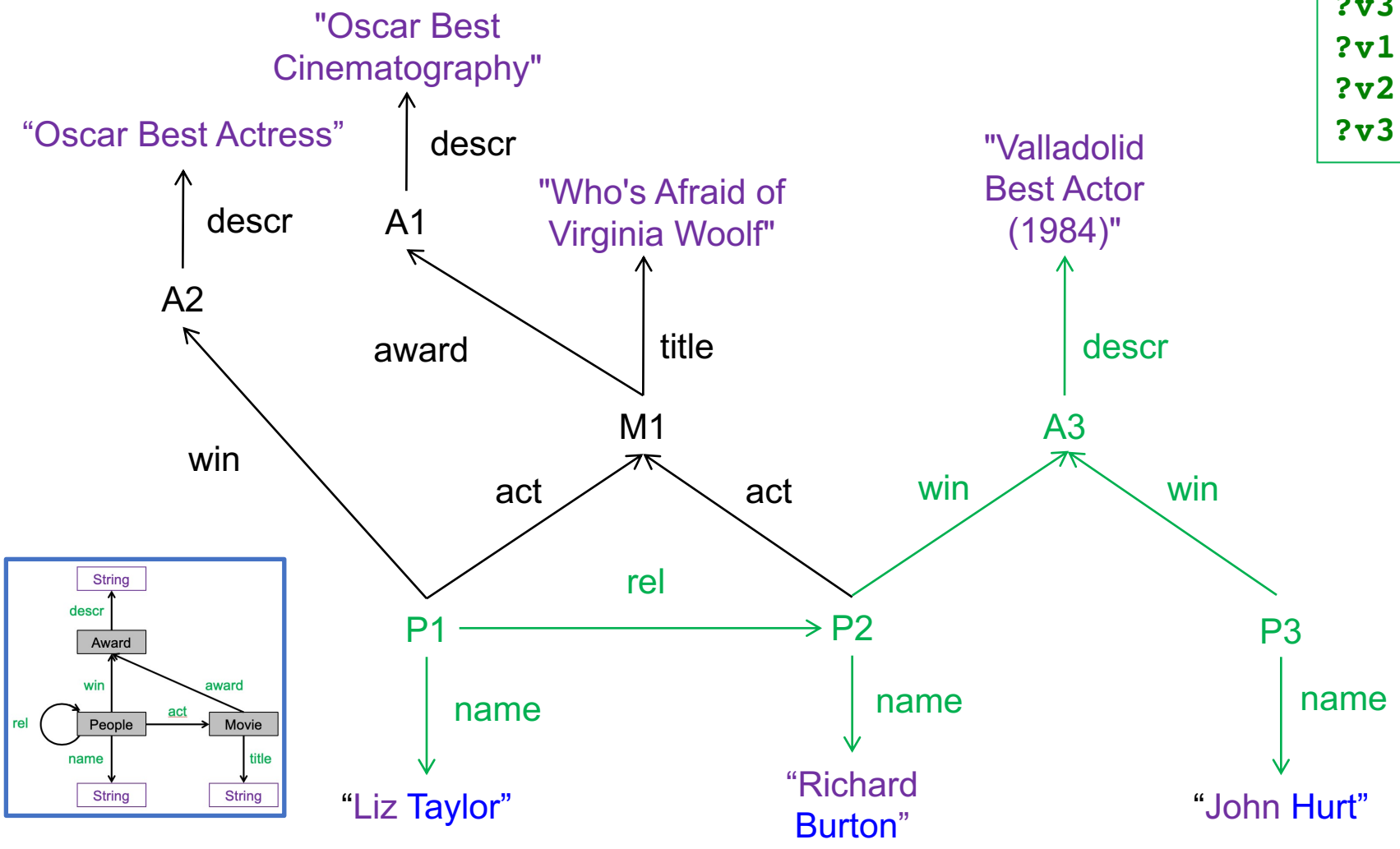


Query: Taylor Burton Hurt



Query: Taylor Burton Hurt

```
?v1 name "Taylor"
?v2 name "Burton"
?v3 name "Hurt"
?v1 rel  ?v2
?v2 win  ?v4
?v3 win  ?v4
```



Taylor Burton Hurt

Dame
Elizabeth Taylor
DBE

Colorized film still of Taylor, late 1950s

Born	Elizabeth Rosemond Taylor February 27, 1932 London, England, U.K.
Died	March 23, 2011 (aged 79) Los Angeles, California, U.S.
Resting place	Forest Lawn Memorial Park, Glendale, California, U.S.
Nationality	British · American
Occupation	Actress · businesswoman · humanitarian
Years active	1942–2003
Title	Dame (2000)
Spouse(s)	Conrad Hilton Jr. (m. 1950; div. 1951) Michael Wilding (m. 1952; div. 1957) Mike Todd (m. 1957; died 1958) Eddie Fisher (m. 1959; div. 1964) Richard Burton (m. 1964, div. 1975; m. 1975, div. 1976) John Warner (m. 1976; div. 1982) Larry Fortensky (m. 1991; div. 1996)
Children	4
Parent(s)	Francis Lenn Taylor Sara Sothern
Awards	Full list
Website	elizabethtaylor.com

Richard Burton
CBE

Burton in *The Robe* (1953)

Born	Richard Walter Jenkins Jr. 10 November 1925 Pontrhydyfen, Glamorgan, Wales
Died	5 August 1984 (aged 58) Céligny, Geneva, Switzerland
Occupation	Actor
Years active	1943–1984
Spouse(s)	Sybil Williams (m. 1949; div. 1963) Elizabeth Taylor (m. 1964; div. 1974) (m. 1975; div. 1976) Suzy Miller (m. 1976; div. 1982) Sally Hay (m. 1983)
Children	3, including Kate Burton

...

1984	*Nineteen Eighty-Four*	Valladolid International Film Festival Award for Best Actor (shared with John Hurt)
	Ellis Island	Nominated—Primetime Emmy Award for Outstanding Supporting Actor in a Miniseries or a Movie

Sir
John Hurt
CBE

Hurt in January 2015

Born	22 January 1940 Chesterfield, Derbyshire, England
Died	25 January 2017 (aged 77) Cromer, Norfolk, England
Occupation	Actor
Years active	1961–2017
Spouse(s)	Annette Robertson (m. 1962; div. 1964) Donna Peacock (m. 1984; div. 1990) Joan Dalton (m. 1990; div. 1996) Anwen Rees-Myers (m. 2005)
Children	2

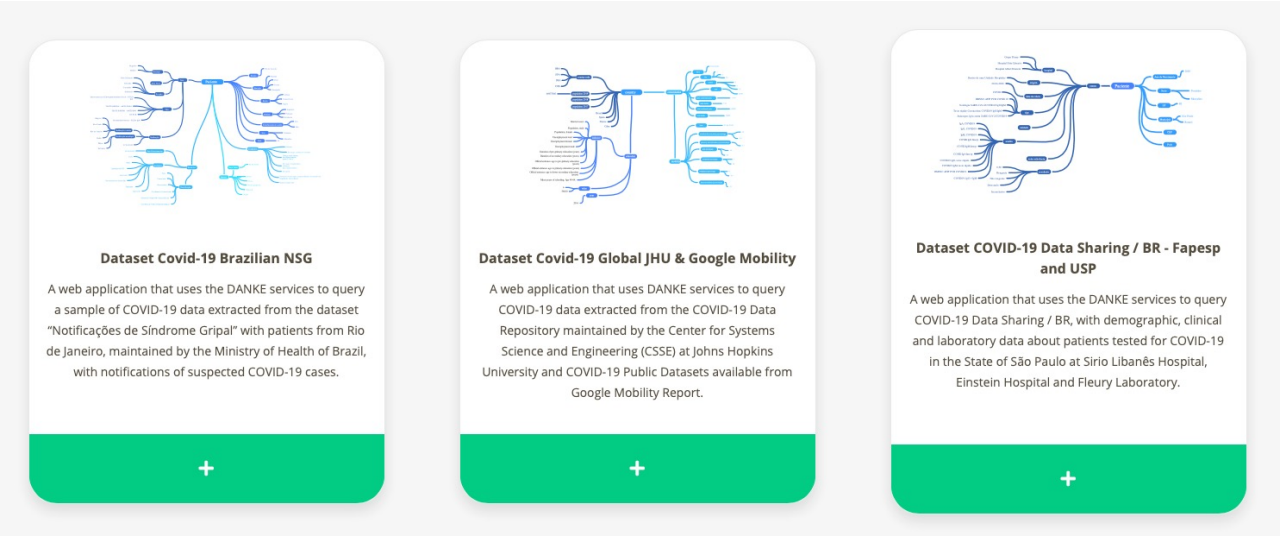
13

Agenda

- Motivation
- DANKE
- *CovidKeys*
- Conclusions

Experiments with *CovidKeys*

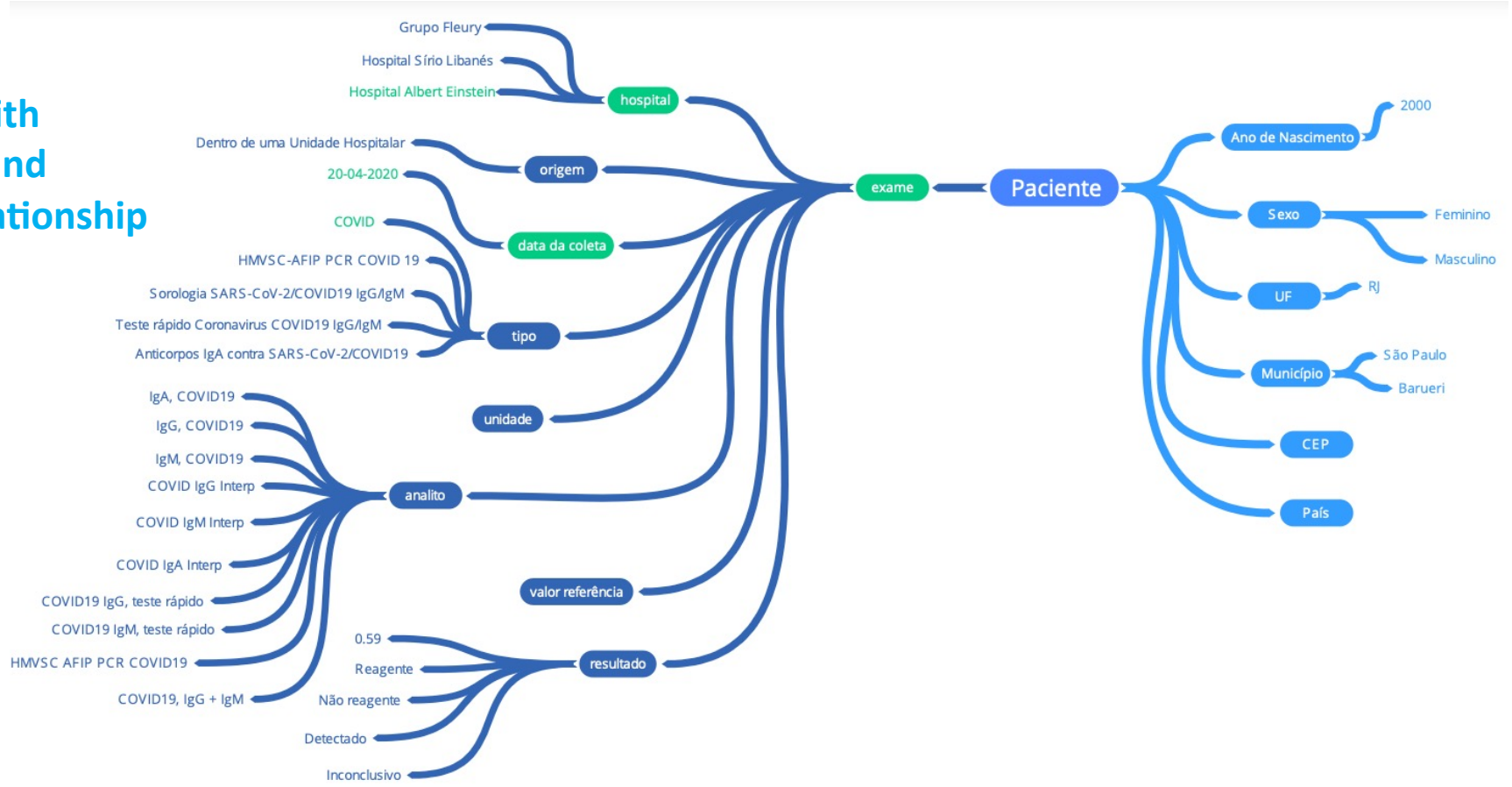
- *CovidKeyS*
  - Web application that offers keyword search over COVID-19 data
  - Built on top of DANKE
  - Three proof-of-concept scenarios
    - Notificações de Síndrome Gripal - NSG
    - Google COVID-19 Public Datasets
      - Google Mobility
      - John Hopkins University
      - World Bank Global
    - **FAPESP COVID-19 Data Sharing/BR**



Experiments with *CovidKeys*

- Scenario 3 – FAPESP COVID-19 Data Sharing/BR Dataset (July 2020)

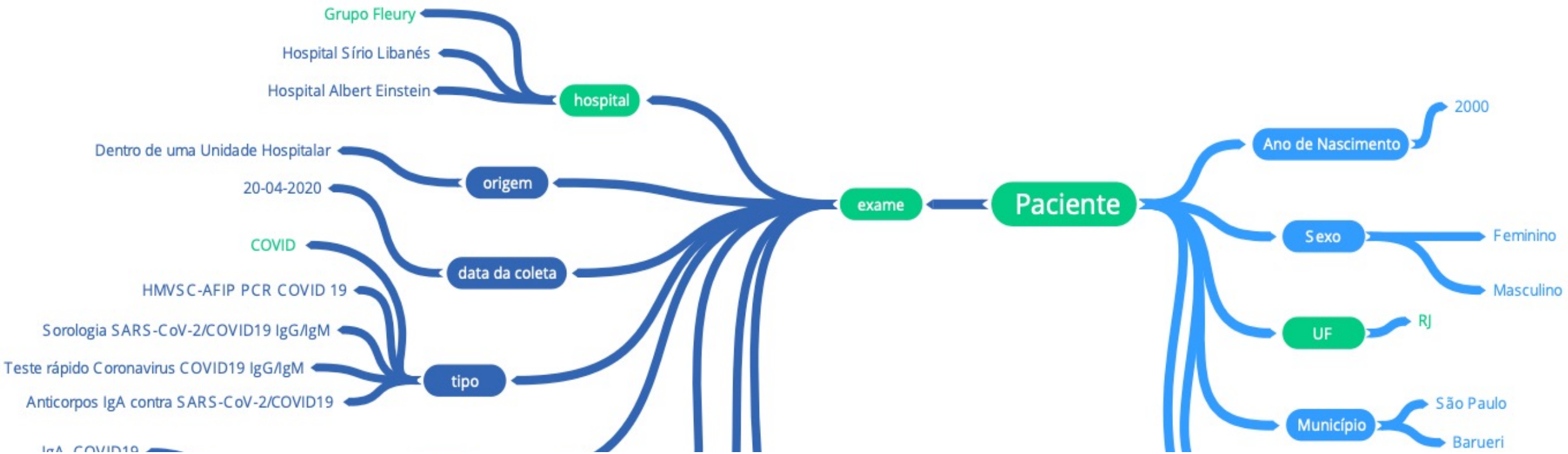
Database with two tables and a 1-to-N relationship



Quais **pacientes** do estado do **RJ** que fizeram **exame** de **covid** no **Hospital Fleury** ?

paciente UF RJ exame covid Hospital Fleury

Combine keywords from the dataset mindmap to query the database



Danke

Covid Fapesp

paciente UF RJ exame covid Hospital Fleury

Showing 1 to 25 of 81524 entries

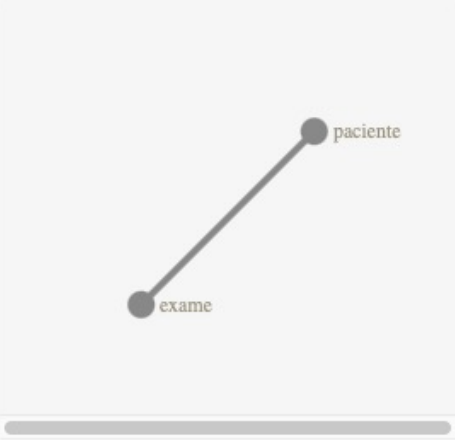
« 1 2 3 4 5 » Showing: 25

Paciente	Uf	Exame	Analito	Hospital
9261A7310492C9A0021EEAC99CEE252F	RJ	COVID19, ANTICORPOS IgG, soro - Paciente: 9261A7310492C9A0021EEAC99CEE252F	Covid 19, Anticorpos IgG, Quimioluminescência	Grupo Fleury
44F129F3A64FA2337F08800345D76D15	RJ	NOVO CORONAVÍRUS 2019 (SARS-CoV-2), DETECÇÃO POR PCR - Paciente: 44F129F3A64FA2337F08800345D76D15	Covid 19, Material	Grupo Fleury
A18C6EBE647F76611F498C9AD85A28F0	RJ	COVID19, ANTICORPOS IgM, soro - Paciente: A18C6EBE647F76611F498C9AD85A28F0	Covid 19, Anticorpos IgM, Quimiolumin.-Índice	Grupo Fleury
46897E23116801F3FF17F2318048EC6B	RJ	COVID19, ANTICORPOS IgM, soro - Paciente: 46897E23116801F3FF17F2318048EC6B	Covid 19, Anticorpos IgM, Quimiolumin.-Índice	Grupo Fleury
5DC7BD249DA3503D31FF58351FB2584A	RJ	COVID19, ANTICORPOS IgG, soro - Paciente: 5DC7BD249DA3503D31FF58351FB2584A	Covid 19, Anticorpos IgG, Quimiolumin.-Índice	Grupo Fleury
2483150F785C46FFC2D7AB865D74D662	RJ	NOVO CORONAVÍRUS 2019 (SARS-CoV-2), DETECÇÃO POR PCR - Paciente: 2483150F785C46FFC2D7AB865D74D662	Covid 19, Detecção por PCR	Grupo Fleury
B100E5CB6B72C748DB045FAC08E1DB55	RJ	COVID19, ANTICORPOS IgM, soro - Paciente: B100E5CB6B72C748DB045FAC08E1DB55	Covid 19, Anticorpos IgM, Quimioluminescência	Grupo Fleury
CC905F33C111519B5F11AEF2478901BB	RJ	COVID19, ANTICORPOS IgM, soro - Paciente: CC905F33C111519B5F11AEF2478901BB	Covid 19, Anticorpos IgM, Quimioluminescência	Grupo Fleury
B9E07C4A134F76E03CD439CE48F6AC42	RJ	COVID19, ANTICORPOS IgG, soro - Paciente: B9E07C4A134F76E03CD439CE48F6AC42	Covid 19, Anticorpos IgG, Quimioluminescência	Grupo Fleury

Data overview

Showing data about:

- Paciente
  - Uf
- Exame
  - Analito
  - Hospital



Result Analysis

Danke

Covid Fapesp

paciente UF RJ exame covid Hospital Fleury

← Back to search result

Result Analysis

- Paciente
  - Uf
- Exame
  - Hospital
  - Analito

Matched Keywords

covid

Filter Value

CONTAINS (covid)

Least Frequent Value

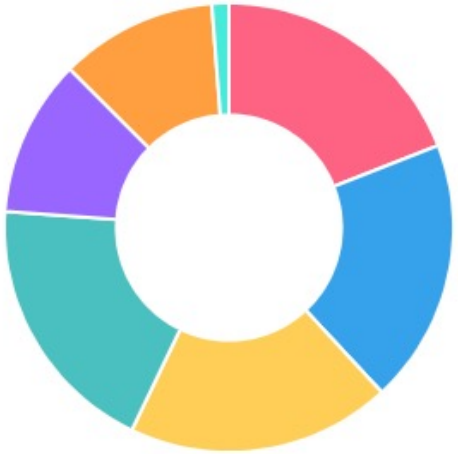
COVID 19,ANTIC IGM/IGG,TESTE RÁPIDO-INTERPRET

Most Frequent Value

COVID 19, ANTICORPOS IGM, QUIMIOLUMIN.-ÍNDICE

Count Distinct

16



19.045%	COVID 19, ANTICORPOS IGM, QUIMIOLUMIN.-ÍNDICE	15526
19.042%	COVID 19, ANTICORPOS IGG, QUIMIOLUMIN.-ÍNDICE	15524
19.036%	COVID 19, ANTICORPOS IGM, QUIMIOLUMINESCÊNCIA	15519
19.034%	COVID 19, ANTICORPOS IGG, QUIMIOLUMINESCÊNCIA	15517
11.323%	COVID 19, MATERIAL	9231

Value	Total
COVID 19, ANTICORPOS IGM, QUIMIOLUMIN.-ÍNDICE	15526
COVID 19, ANTICORPOS IGG, QUIMIOLUMIN.-ÍNDICE	15524
COVID 19, ANTICORPOS IGM, QUIMIOLUMINESCÊNCIA	15519
COVID 19, ANTICORPOS IGG, QUIMIOLUMINESCÊNCIA	15517
COVID 19, MATERIAL	9231
COVID 19, DETECÇÃO POR PCR	9213
COVID 19, ANTICORPOS IGG, ELISA	290
COVID 19, ANTICORPOS IGG, ELISA - ÍNDICE	290
COVID 19, ANTICORPOS IGA, ELISA	99
COVID 19, ANTICORPOS IGA, ELISA - ÍNDICE	99
COVID 19, ANTIC IGA/IGG, ELISA - OBSERVAÇÃO	70
COVID 19, ANTIC IGA/IGG, ELISA-INTERPRETAÇÃO	70
COVID 19,ANTIC IGM/IGG,TESTE RÁPIDO-INTERPRET	19
COVID 19, ANTICORPOS IGG, TESTE RÁPIDO	19

Agenda

- Motivation
- DANKE
  - Architecture
  - Keyword Query Processing
- Experiments with *CovidKeys*
- Conclusions

Conclusions

- Summary
  - DANKE - a platform for data and knowledge retrieval
  - *CovidKeyS*
    - uses DANKE to implement keyword queries over three COVID-19 data scenarios
      - The NSG (Notificações de Síndrome Gripal) dataset
      - Global Data (from the Google COVID-19 Public Datasets)
      - **The FAPESP COVID-19 Data Sharing / BR initiative**

Conclusions

- What else?
  - Extend DANKE to...
    - Operate over federated databases
    - Process change sets
  - Extend keyword query processing to support...
    - Spatial operators (e.g., *near this point*)
    - Temporal operators (e.g., *same date*)
    - Aggregation operations
  - ... beyond keyword query
    - Natural Language Interfaces for Databases – NLIDB

John Snow's map of cholera victims
https://en.wikipedia.org/wiki/John_Snow

Thank You

References for this presentation...

`casanova puc-rio`