

Plongez dans les panama papers avec Neo4j

Benoit Simard (@logisima)



Introduction

Benoit Simard

- Consultant Neo4j
- benoit.simard@neotechnology.com
- @logisima
- Addicte aux graphes
- Formateur



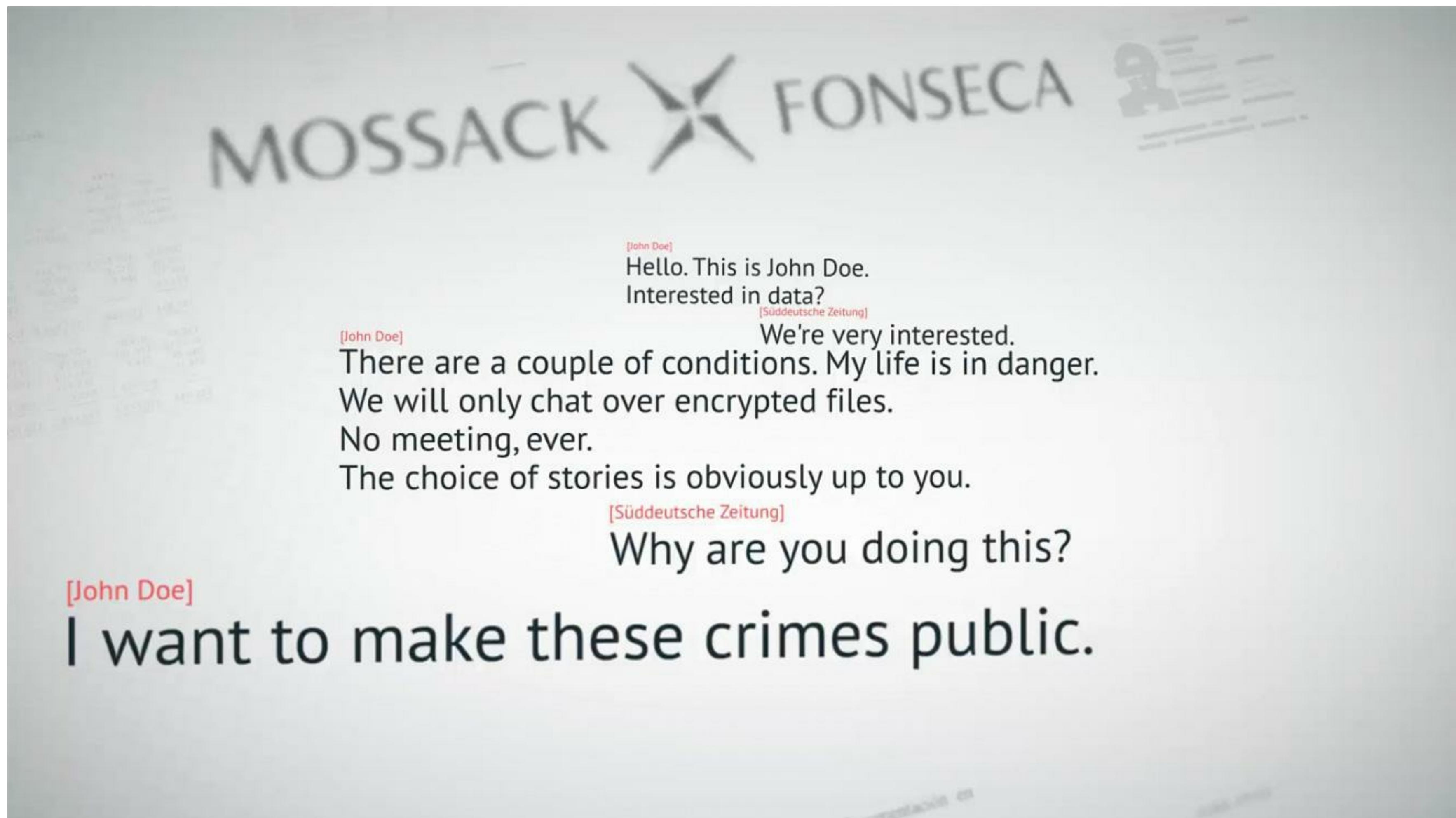
Les sources de données

Les sources de données de cette présentation proviennent des endroits suivants

- Les présentations de l'ICIJ
 - Graph Connect Londre
 - Sud-web
- [Reddit](#)
- <https://panamapapers.icij.org/>

L'histoire

L'origine



La volumétrie

C'est la plus large fuite de données de l'histoire, et de loin !





~190 journalistes dans plus de 65 pays



Equipe de 12 personnes à temps plein (USA, Costa Rica, Venezuela, Germany, France, Spain)
dont 50% travaillent sur l'analyse des données.



Les outils

Le processus

```
Failed to generate image: Could not find the 'dot' executable in PATH; add it to the PATH or specify its location using the 'graphviz_dot' document attribute
digraph finite_state_machine {
    rankdir=TB;
    node [ fontsize=20, shape = Mrecord];
    edge [ fontsize=20 ];

    "Fichier brut";
    "Texte brut";
    "Métadonnées";
    "Base de données";
    "Explorer / Analyser";

    "Fichier brut" -> "Texte brut";
    "Fichier brut" -> "Métadonnées";
    "Texte brut" -> "Base de données";
    "Métadonnées" -> "Base de données";
    "Base de données" -> "Explorer / Analyser";
}
```

Le cloud à la rescousse

Utilisation de Niux OCR pour récupérer le texte des documents, ainsi que d'autres outils pour établir des références croisées entre des millions de documents sur le nom des clients de Mossack Fonseca.



- 3 millions de fichiers x 10 secondes/fichier ⇒ **1 année**
- 1 année / 35 serveurs ⇒ **1,5 semaine**

Blacklight

Projet open-source de gestion des catalogues de bibliothèque, permettant de réaliser des **requêtes solr**

The screenshot shows the Blacklight search interface. At the top, there's a navigation bar with links for Bookmarks (4), Saved Searches, History, Log Out, mcaruana@icij.org, Users, Forum, and the International Consortium of Investigative Journalists (ICIJ) logo. Below the navigation is a search bar with dropdowns for "Everything" and the query "joaquin loera"~2, followed by a "Search" button. To the right of the search bar is a red box containing the number "400". On the left, there's a sidebar titled "Limit your results" with dropdowns for Data update, Path, Year created or sent, and Type. A red arrow points from the "Lucene syntax queries with" text below to the search bar. Another red arrow points from the "400" box to the search results area. The search results show 1 - 4 of 4 items. Item 1 has fields for Text, Subject, Date, and Creator, with "Joaquin Guzmán Loera" appearing in the Text and Subject fields. A red arrow points to the "Joaquin Guzmán Loera" text in the Subject field. At the bottom right is a red box with the "Solr" logo.

blacklight

Bookmarks (4) Saved Searches History Log Out mcaruana@icij.org Users Forum INTERNATIONAL CONSORTIUM OF INVESTIGATIVE JOURNALISTS ICIJ Start Over

Everything "joaquin loera"~2 Search

You searched for: "joaquin loera"~2 × 400

Sort by Relevance ▾ 10 per page

Limit your results

Data update

Path

Year created or sent

Type

1 - 4 of 4

1. [REDACTED]

Text: [REDACTED] of Joaquin Guzmán Loera [REDACTED]

Subject: [REDACTED] Joaquin Guzmán Loera [REDACTED]

Date: [REDACTED]

Creator: [REDACTED]

Bookmark

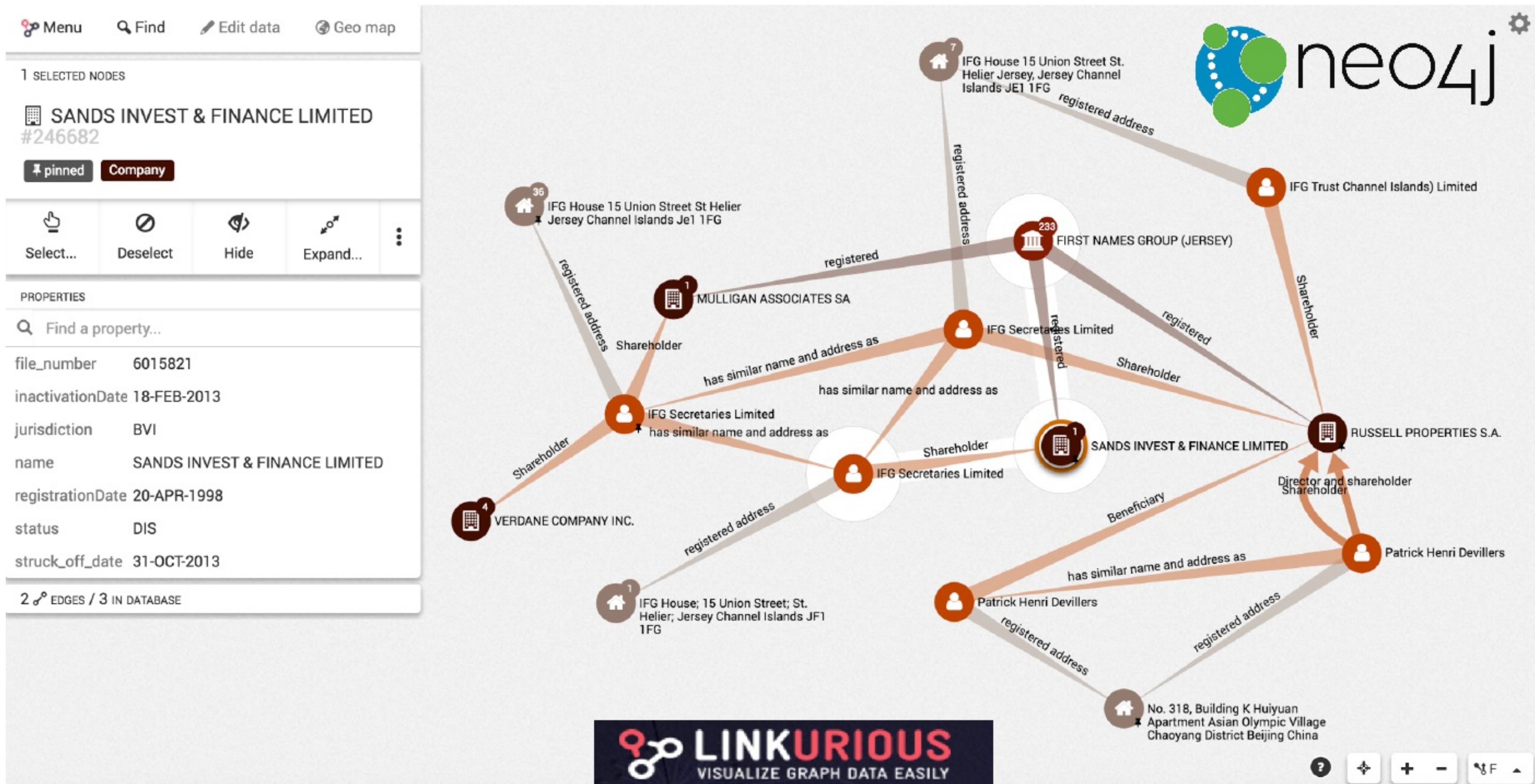
Lucene syntax queries with

Solr

La stack technique

- **Extraction des données non structurées** : Nuix OCR, ICIJ Extract (open source, Java: <https://github.com/ICIJ/extract>), leverages Apache Tika, Tesseract OCR and JBIG2-ImageIO.
- **Extraction des données structurées** : un peu de Python
- **Base de données** : Apache Solr (open source, Java), Redis (open source, C), Neo4j (open source, Java)
- **Application** : Blacklight (open source, Rails), Linkurious (closed source, JS)

Linkurious

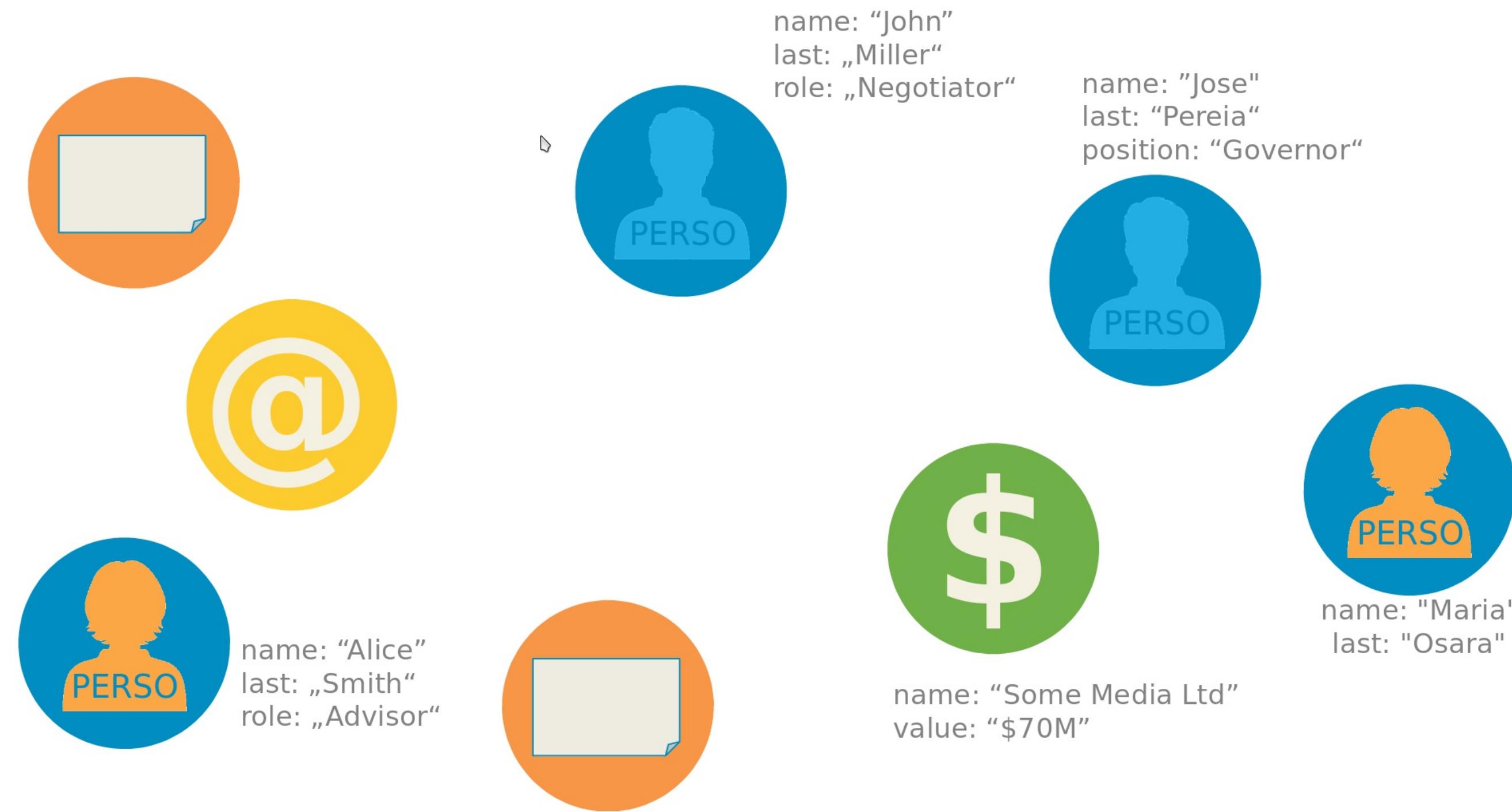


Comment lier les données ?

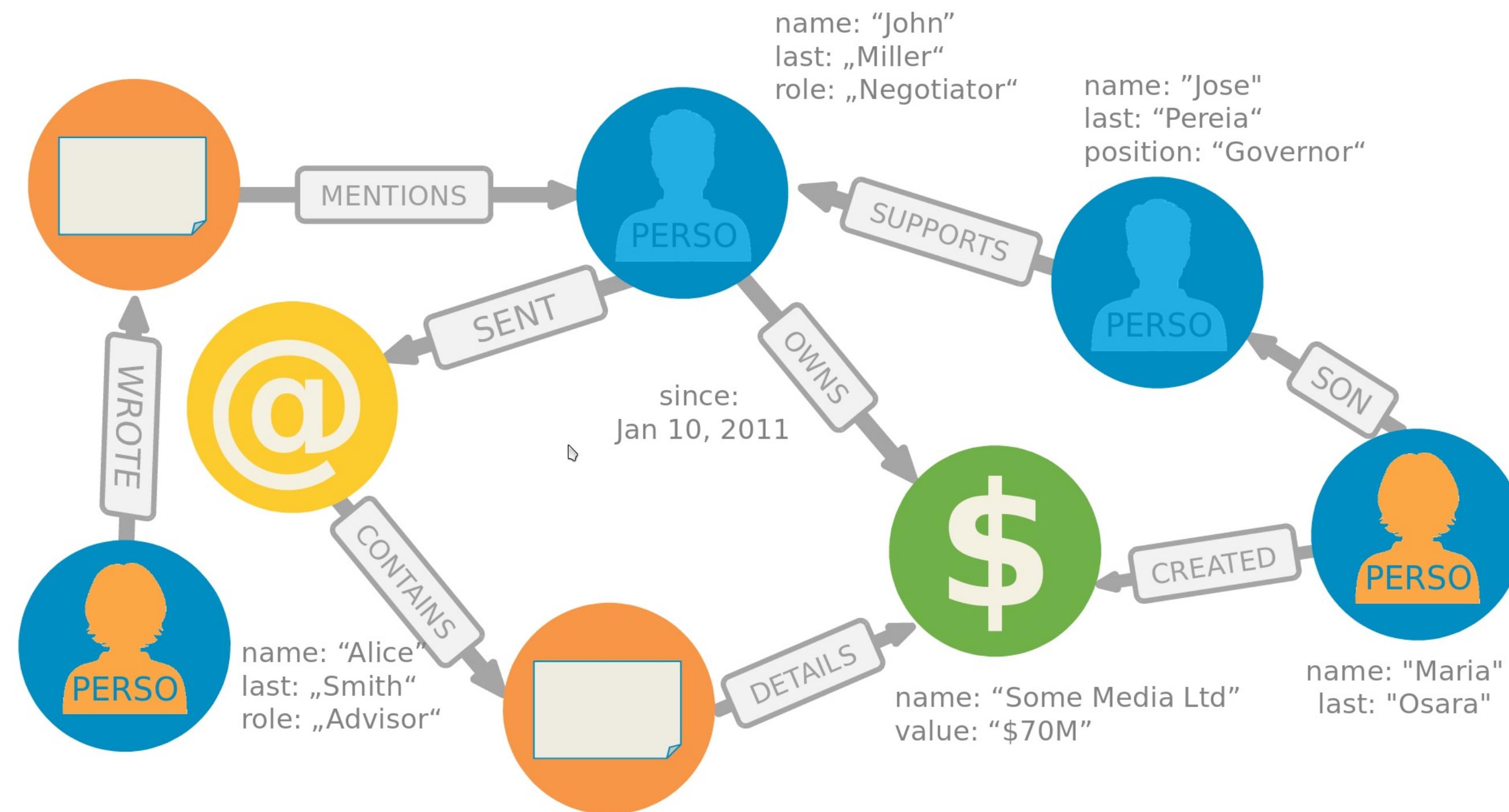
Des documents déconnectées



Des données connectées ?



C'est une histoire de contexte



(graph)-[:ARE]→(everywhere)

Le monde est un graph, tout est connecté !

- les personnes, les lieux, les évènements
- Les entreprises, le marché, les clients
- les pays, l'histoire, les politiques
- technologie, les réseaux, les machines, les utilisateurs
- les applications, le code, les dépendances, l'architecture, le déploiement
- ...



Stocker & requêter

Le graphe



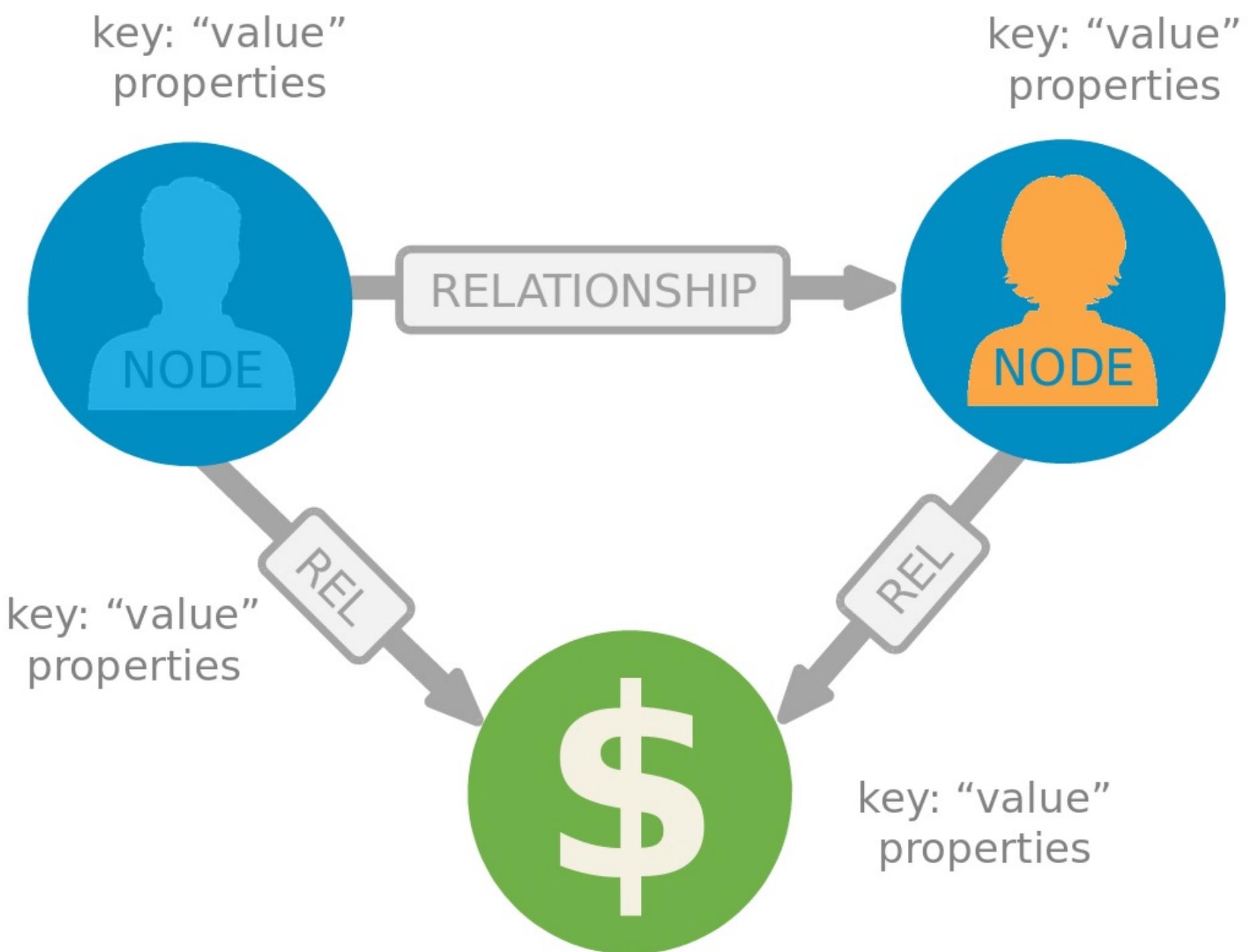
Modélisation en graphe

Les noeuds

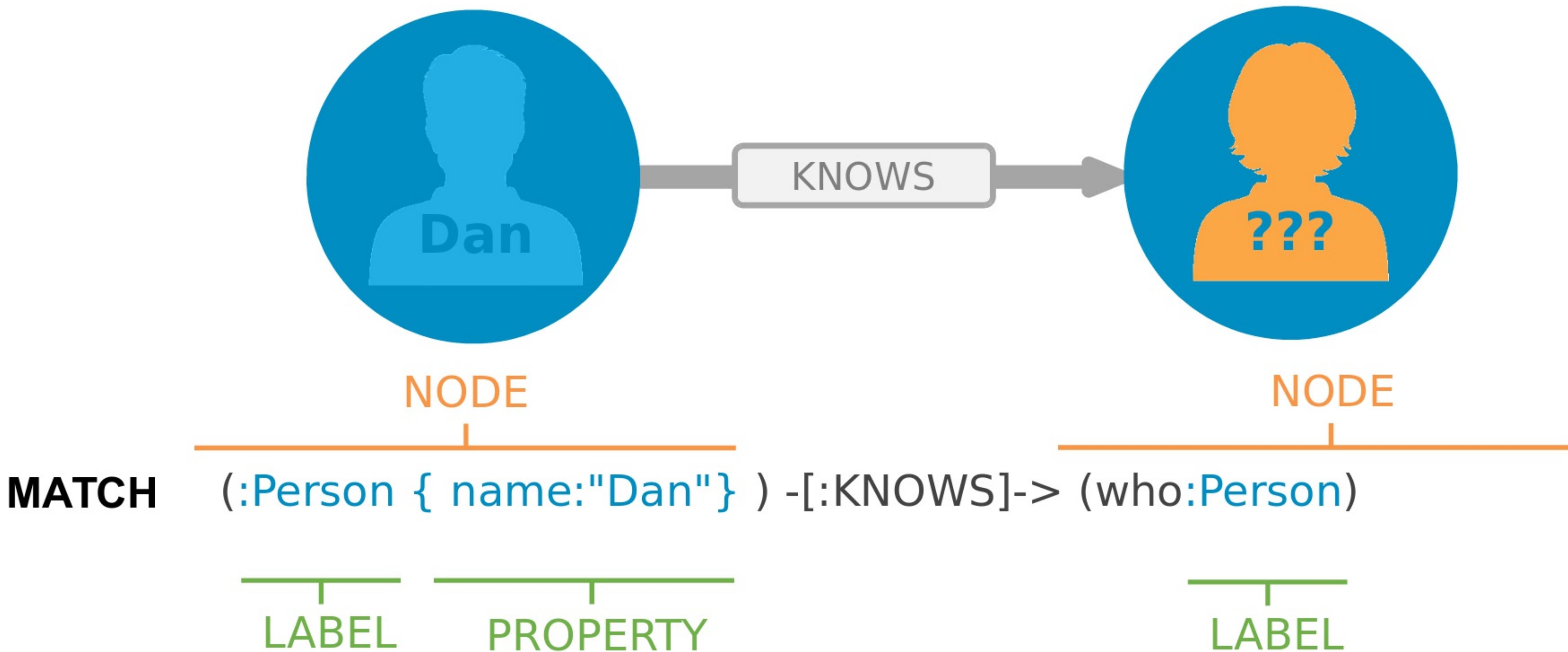
- Les entités du graphe
- Peuvent avoir des propriétés
- Peuvent avoir des labels (étiquettes)

Les relations

- Relient des noeuds avec un type et une direction
- Peuvent avoir des propriétés



Tout est pattern



RETURN who I TO U4J

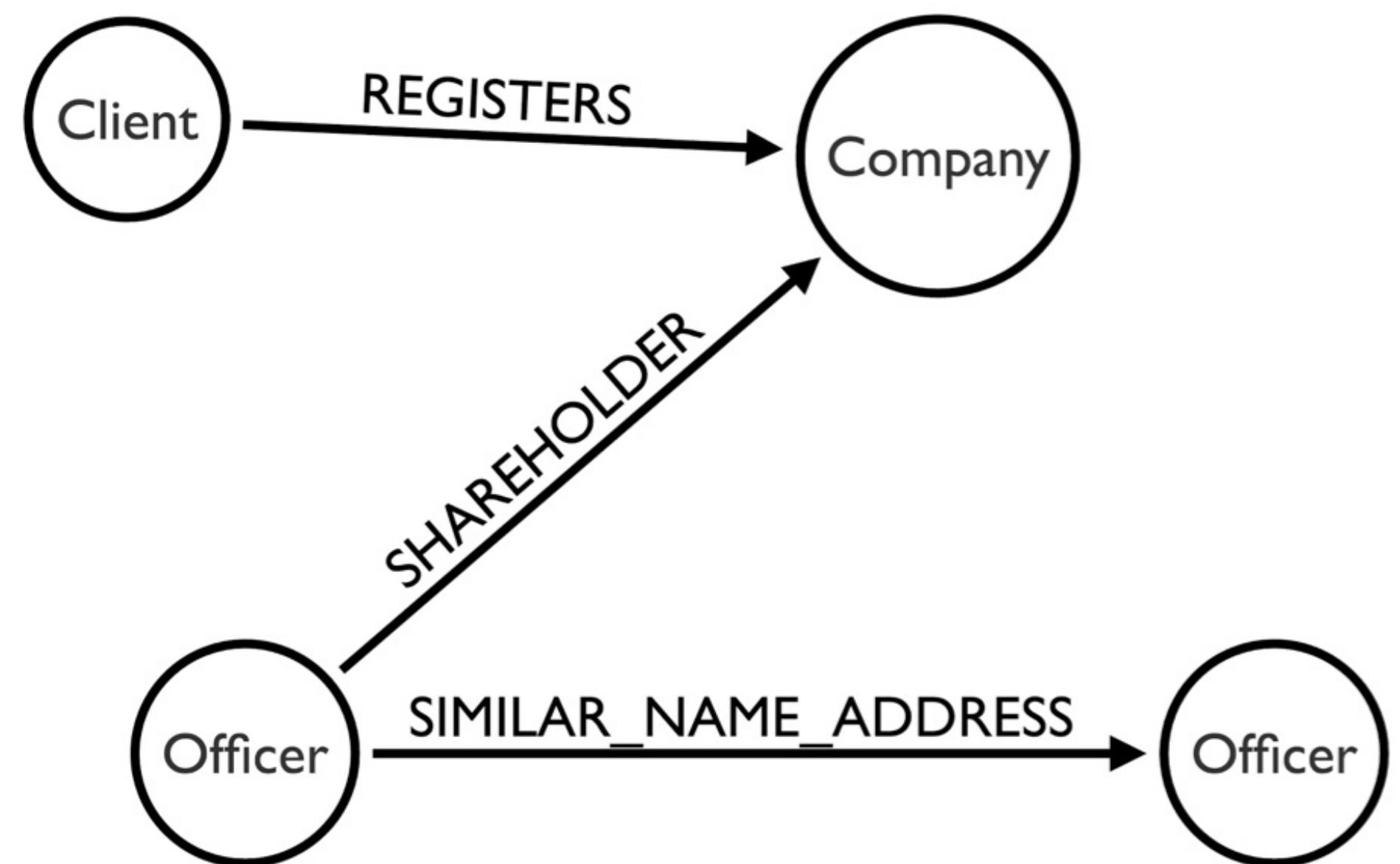
Les étapes à l'analyse des données

1. Obtenir les documents
2. Classifier les documents (Scan, OCR, métadonnées)
3. Etablir une modélisation avec les entités, leurs relations et les propriétés
4. Développer des analyseurs, des parseurs, des règles permettant de récupérer les noms des entités
5. Parser les documents et stocker les données trouvées ainsi que les méta-données
6. Déduire les relations entre les entités (grâce au contexte)
7. Calculer les similitudes, trouver les relations transitivités, triangulaire
8. Analyser (cypher) et explorer (Linkurious)

Les Panama Papers

Le modèle

- Modèle simpliste (4 entités et 5 relations)
- On ne connaît que le modèle publié
- Il manque : les documents, les metadata, les relations familiales
- Les connections aux données publiques (opendata)
- Contient des doublons
- Les informations de relation sont stockées sur les entités



Exemple : Président Azerbaijan - Ilham Aliyev

Panama Papers The Power Players 



Ilham Aliyev
President of Azerbaijan (2003-present)
Relatives in the data: *Prime Minister Ilham Aliyev's wife, children and sister*

Related countries
Azerbaijan

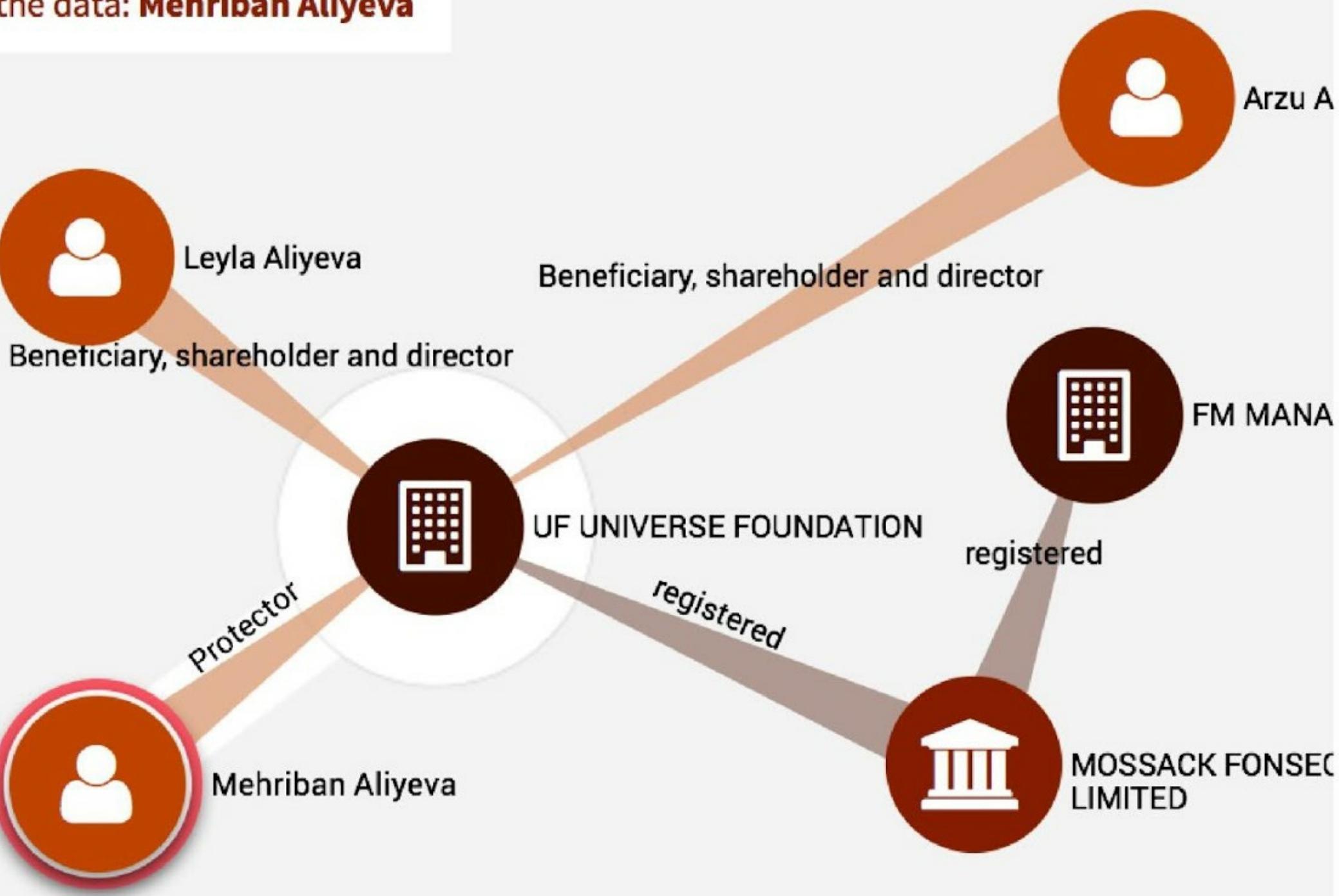
Mehriban Aliyeva, Wife



The family of Azerbaijan President Ilham Aliyev leads a charmed, glamorous life, thanks in part to financial interests in almost every sector of the economy. His wife, Mehriban, comes from the privileged and powerful Pashayev family that

Explore the data: Mehriban Aliyeva



```
graph TD; MA((Mehriban Aliyeva)) --- UF((UF UNIVERSE FOUNDATION)); UF --- A((Arzu A)); UF --- FM((FM MANA)); UF --- MF((MOSSACK FONSECA LIMITED)); Leyla((Leyla Aliyeva)) --- UF
```

Leyla Aliyeva
Beneficiary, shareholder and director

Arzu A
Beneficiary, shareholder and director

FM MANA
registered

MOSSACK FONSECA LIMITED
registered

UF UNIVERSE FOUNDATION
Protector

Mehriban Aliyeva

Pour aller plus loin



Suivez le guide

A taper dans le browser Neo4j

```
1 :play http://guides.neo4j.com/graphgist/panama_papers.html
```

Chargez les données !

<https://offshoreleaks.icij.org/pages/database>

```
1 ./neo4j-community-3.0.1/bin/neo4j-import --into graph.db \
2   --nodes:Address ${data_dir}/Addresses.csv \
3   --nodes:Entity ${data_dir}/Entities.csv \
4   --nodes:Intermediary ${data_dir}/Intermediaries.csv \
5   --nodes:Officer ${data_dir}/Officers.csv \
6   --relationships ${data_dir}/all_edges_header.csv,${data_dir}/all_edges.csv \
7   --ignore-empty-strings true \
8   --skip-duplicate-nodes true \
9   --skip-bad-relationships true \
10  --bad-tolerance 1500 \
11  --multiline-fields=true
```

En savoir plus

- **Neo4j Blog**
 - <http://neo4j.com/blog/panama-papers/>
 - <http://neo4j.com/blog/analyzing-panama-papers-neo4j/>
- **ICIJ**
 - <https://panamapapers.icij.org/>
 - https://panamapapers.icij.org/the_power_players/
 - <https://panamapapers.icij.org/graphs/>
- **SZ**
 - <http://panamapapers.sueddeutsche.de/en/>
- **Guardian**
 - <http://www.theguardian.com/news/series/panama-papers>

Merci



Des questions ?

- **Twitter:** Suivez les comptes @neojFr & @neo4j
- **Google group :** Avec les groupes Neo4jFr & Neo4j
- **Stackoverflow :** avec les tags neo4j & cypher
- **Slack :** <http://neo4j-users-slack-invite.herokuapp.com/>

