Panama Papers - Investigation et Big Data

Breizh Data Day, 19/10/2017

Présentations

Michel Caradec (mcaradec@altaven.com)

Project Manager, Software/Data Engineer

Agenda

1. L'affaire.
2. Les données.
3. Mise en oeuvre.
4. Méthodes d'investigation.
5. Organisation.

L'affaire

L'affaire

- Mossack Fonseca : cabinet de juristes, basé au Panama.
  - 40 agences dans le monde.
  - Spécialisé dans le service aux sociétés.
  - Aide à la création de sociétés extra-territoriales (offshore).

L'affaire

Rappels :

- La création d'une société offshore n'est pas illégale.
- Doit-être déclarée à l'administration fiscale.
- Illégal si utilisé comme société-écran pour l'évasion fiscale ou le blanchiment d'argent.

L'affaire

- Début 2015 : fuite de données orchestrée par un employé (pseudonyme John Doe) :
  - Divulgation de 214 488 comptes offshore,
  - contrats entre personnes / sociétés,
  - pour plus de 200 pays.

L'affaire

- Données récupérées par le Süddeutsche Zeitung (quotidien allemand).
- Partagées avec l'ICIJ (International Consortium of Investigative Journalists), et aussi la BBC, le Guardian et l'Indian Express.
- Premières publications le 3 avril 2016.
- La plus importante fuite de données financières de l'histoire.

L'affaire

Source : ICIJ

L'affaire

- 109 organismes de presse.
- 370 journalistes.
- 80 pays.

L'affaire

Source : https://panamapapers.icij.org/the_power_players/

Points marquants

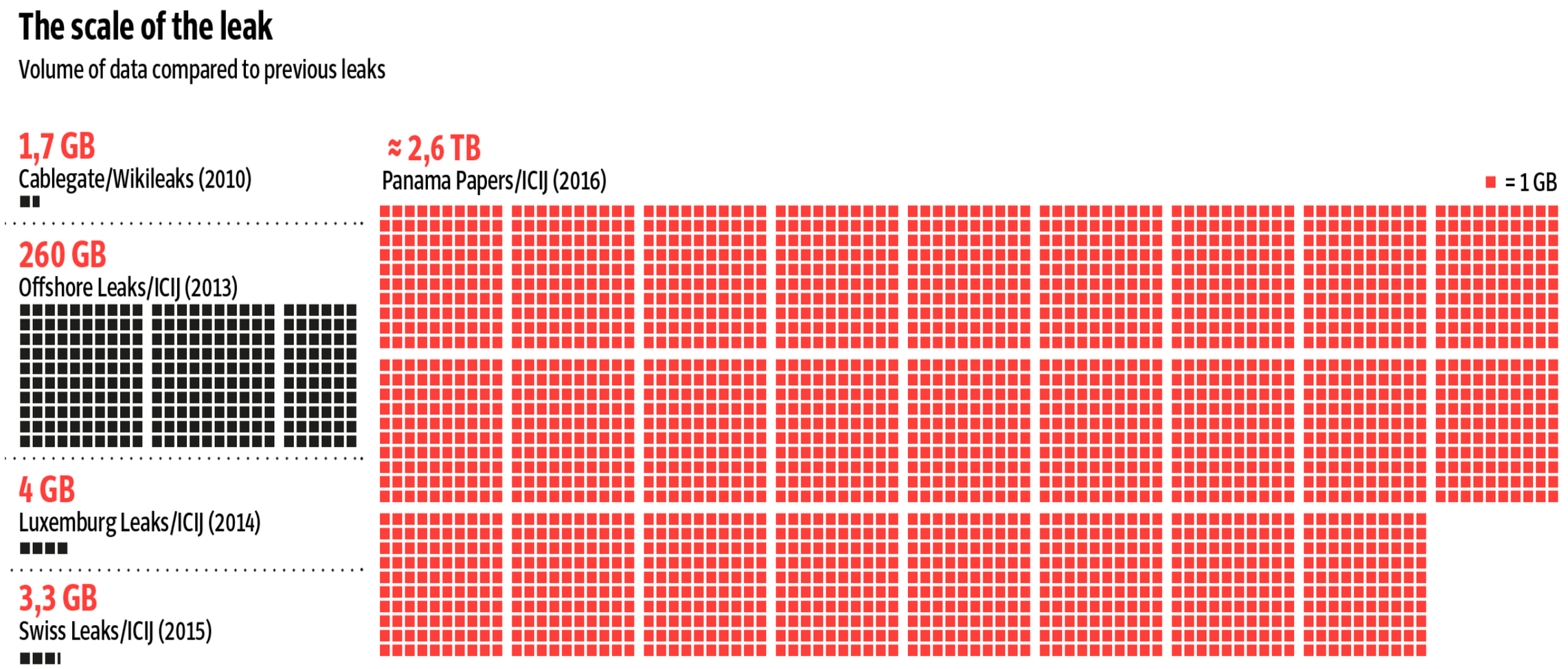
- Peu de sociétés américaines apparaissent...
- Mossack Fonseca hacké 3 fois depuis 2013.
    - Données sur le dark web.
    - Non diffusées, car sporadiques et incomplètes,
    - et pour la difficulté à traiter de telles données?

Les données

Données

- 40 années d'historique (fin 1970-2015).
- 2.6 tera octets de données (= 2 662.4 Go = 2 726 297.6 Mo).
  - = 665 DVDs.
    - pile de 9m30.
    - poids de 10.64 Kg.
  - = 18 175 albums MP3.

Données

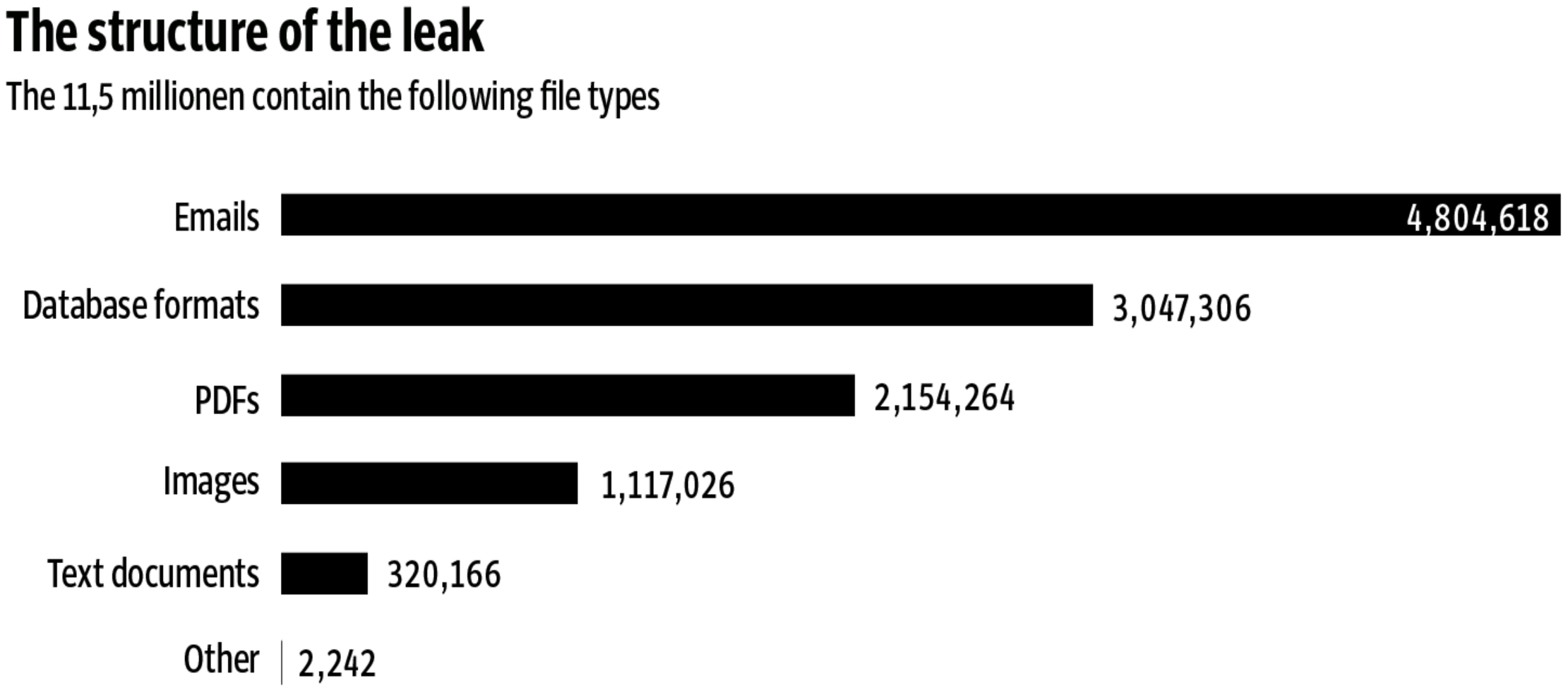


Source : Süddeutsche Zeitung

Données en détail

- 11.5 millions de documents.
- Données hétérogènes, non structurées :
  - .doc, .pdf, .xls, .msg, image.
  - Essentiellement texte (peu de chiffres).
- Données complémentaires en Open Data :
  - Registre du commerce Luxembourgeois (http://www.legilux.public.lu/).
  - Société.com, Verif.com.

Données



Source : Le Monde

Mise en oeuvre

Big Data - Définition

Big Data (méga données) = données ne pouvant être traitées dans temps raisonnable sur une seule machine de par :

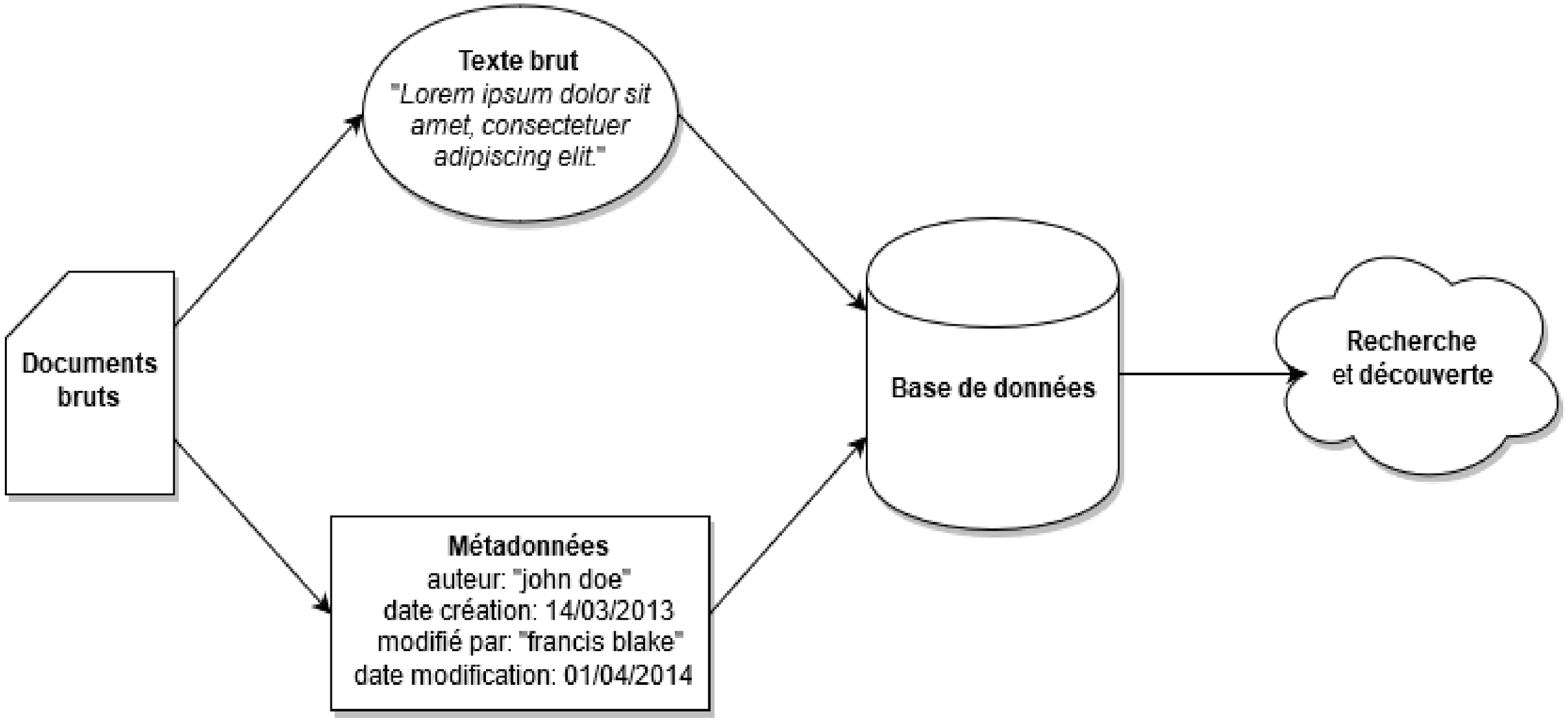
- leur taille (Volumétrie),
- ou leur caractère hétérogène (Variété),
- ou leur vitesse de création/assimilation (Vélocité).

Big Data - Définition

Exemples :

1. Base Sirene des entreprises (8 Go) = small data.
2. Données de géolocalisation de téléphones (calcul du meilleur itinéraire) = big data.
3. Analyse des données d'une flotte d'avions (1/2 To par vol) = big data.

Chaîne de traitement de l'information des Panama Papers



Chaîne de traitement de l'information des Panama Papers

1. Acquisition des documents (et autres données).
2. Classification des documents.
3. Nettoyage des données.
4. Intégration / Stockage.
5. Exploitation des données.

1. Acquisition des documents / données

- Nuix : reconnaissance de caractères (OCR).
- Tesseract OCR : moteur OCR Open Source.
  - Provision de 30-40 serveurs temporaires sur Amazon (Cloud Computing).
  - 3 millions documents x 10 secondes par document = 1 année / 35 serveurs = 1.5 semaine.
  - 99% des données scannées et indexées.

1. Acquisition des documents / données

- Python : data scraping, extraction de données structurées.

2. Classification des documents

- Apache Tika
  - Détection automatique de types de documents (Detectors).
  - Extraction de texte (contenu) et de métadonnées (auteur, date, etc.) de multiples types de fichiers.
  - Analyse de contenu, identification de la langue (pour la traduction).
- Piloté avec l'utilitaire ICIJ Extract (projet Open Source).

2. Classification des documents - Analyse textuelle

- Stemming : algorithme Porter-Stemmer (racinisation par suppression de la fin du mot).
  - `fishing, fished, fish, fisher` => `fish`
- Lemmatisation (obtention de la forme canonique d'un mot).
  - `positionnant, positions, positionnées` => `positionner`
- n-Gramme : séquence de n mots (traitement du langage naturel).
  - 3-gramme `transport en commun` : `transport en` => mot suivant = `commun`

2. Classification des documents - Analyse textuelle

Pour aller plus loin :

- "Premiers pas en text-mining avec R" sur Data-Bzh
  - http://data-bzh.fr/text-mining-r-part-1/
  - http://data-bzh.fr/text-mining-r-part-2/
  - http://data-bzh.fr/text-mining-r-part-3/
  - http://data-bzh.fr/text-mining-r-part-4/
  - http://data-bzh.fr/premiers-pas-en-text-mining-avec-r-partie-5/
- "Introduction au text-mining avec R" sur ThinkR
  - https://thinkr.fr/tm-ou-tidytext-introduction-au-text-mining-avec-r/
  - https://thinkr.fr/text-mining-n-gramme-avec-r/
  - https://thinkr.fr/text-mining-et-topic-modeling-avec-r/
- NLPolitics (http://www.nlpolitics.com/).

3. Nettoyage des données

- OpenRefine (http://openrefine.org) : harmonisation, segmentation (regroupement automatique).
    - `France`, `FR`, `FRA` => `France`

4. Intégration / Stockage

- Talend : alimentation des bases de données.

4. Intégration / Stockage

- Apache Solr : moteur de recherche.

- Redis : base de données mémoire (cache = rapidité).

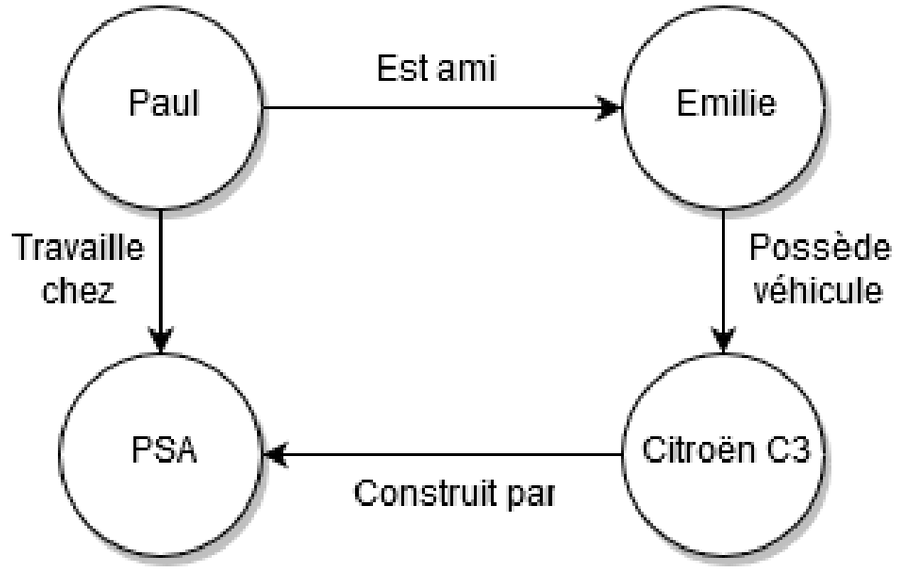
4. Intégration / Stockage

- Neo4j : base de données orientée graphes.

Bases de données Graphes

- Données stockées selon le principe des graphes.
  - Entités = noeuds (nodes, vertices) reliées par des verbes = arcs (edges, links).

```
A est ami avec B.
B possède un véhicule C.
C est construit par D.
A travaille chez D.
```

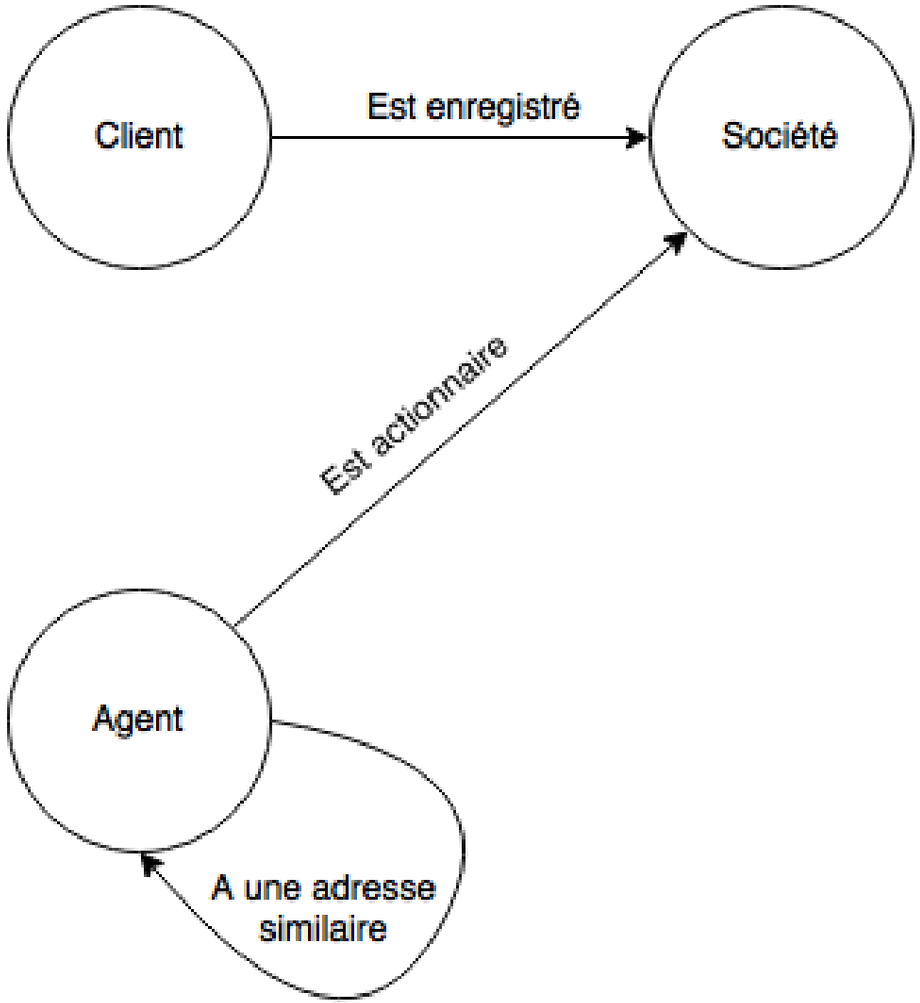


Bases de données Graphes

Requêtes multi-niveaux :

- Mailing : `Liste des employés de PSA ayant des amis avec un véhicule d'une marque différente de PSA`.
- Détection de fraude : `titulaires de comptes ayant des contacts en commun`.

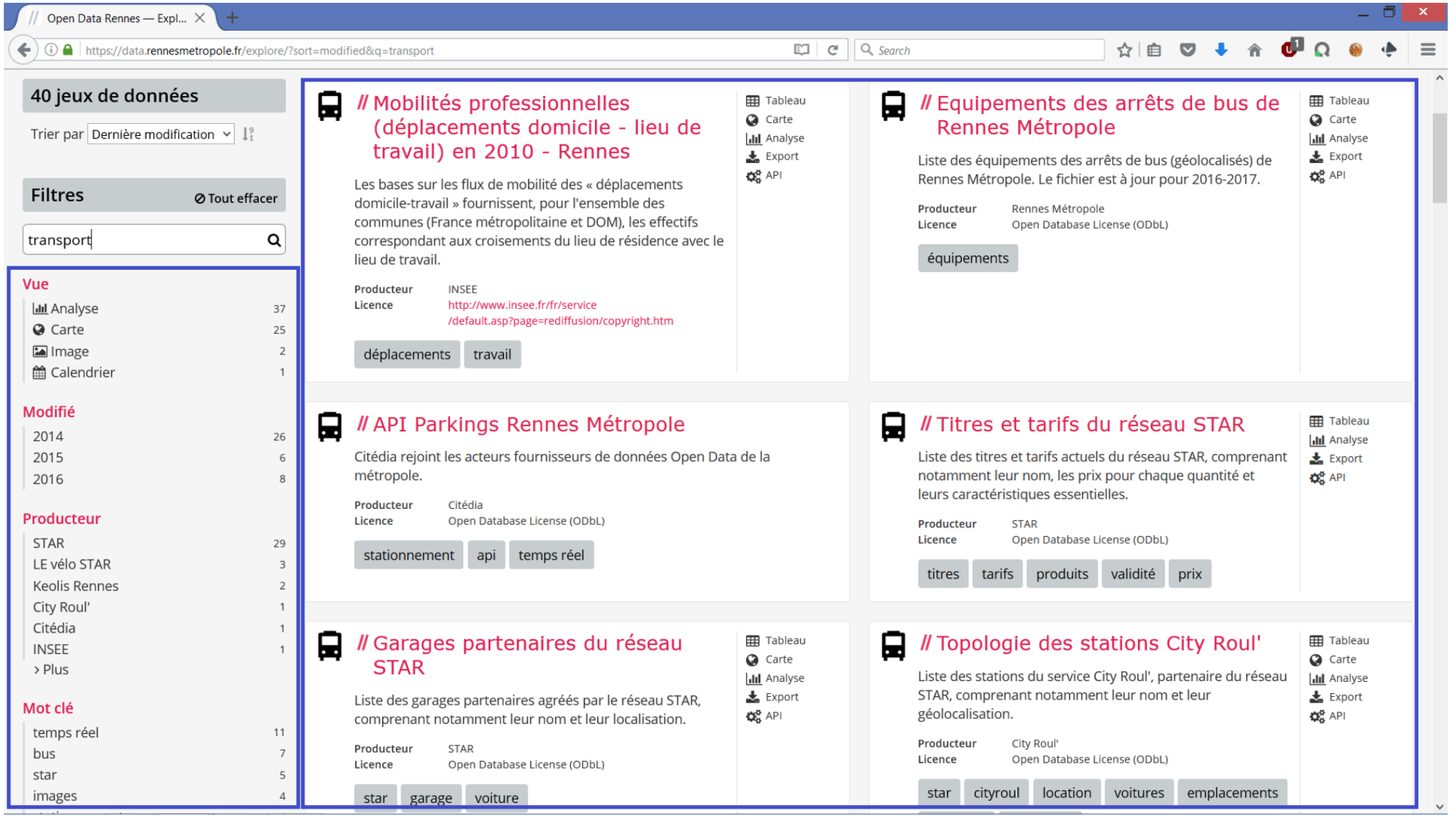
Modèle de données des Panama Papers



5. Exploitation des données

- Blacklight : frontal pour Solr.
  - Recherche par facettes (découverte de données).
  - Outils initialement prévu pour les libraires.

Recherche par facettes

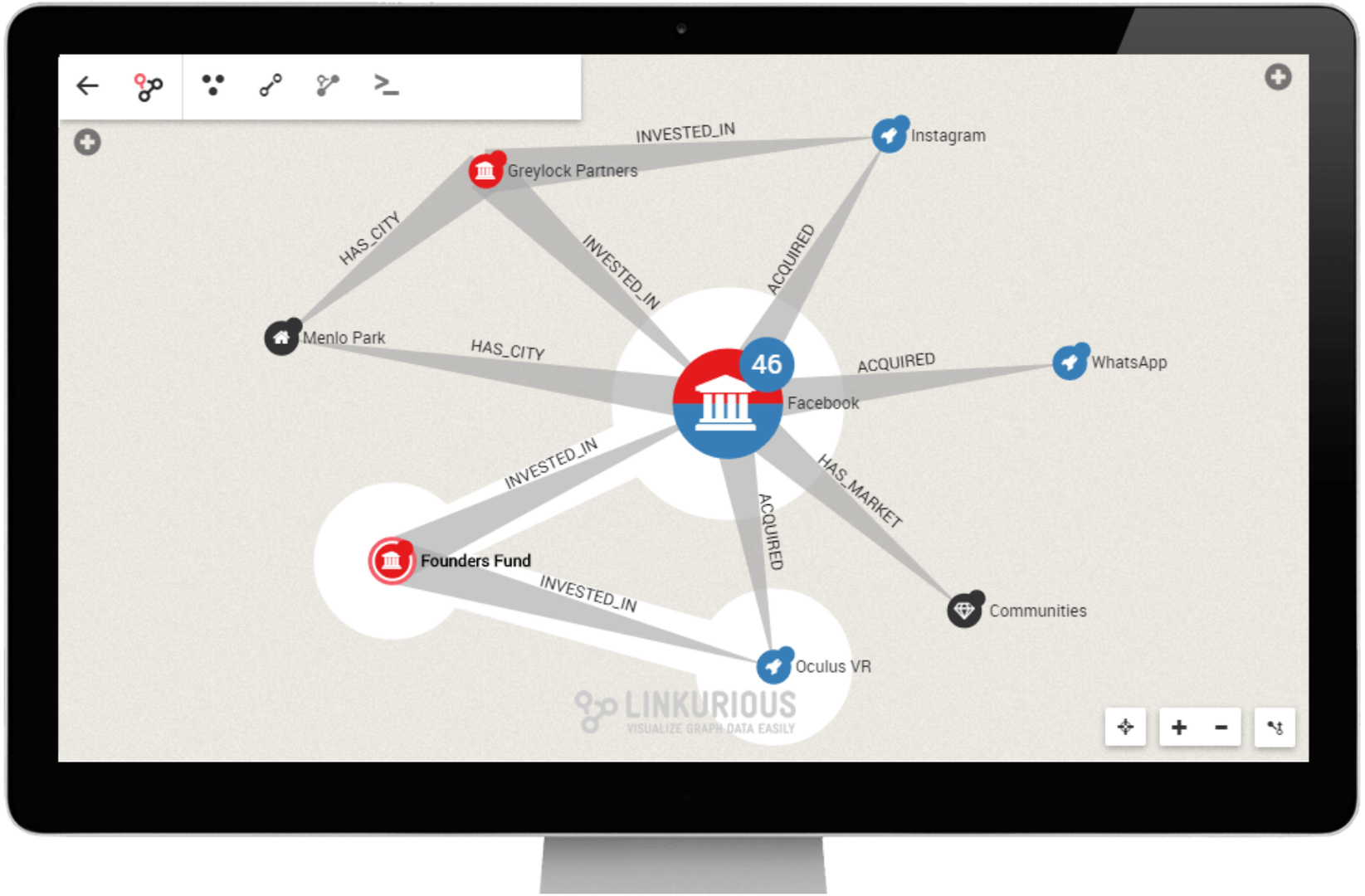


Source : https://data.rennesmetropole.fr/

5. Exploitation des données

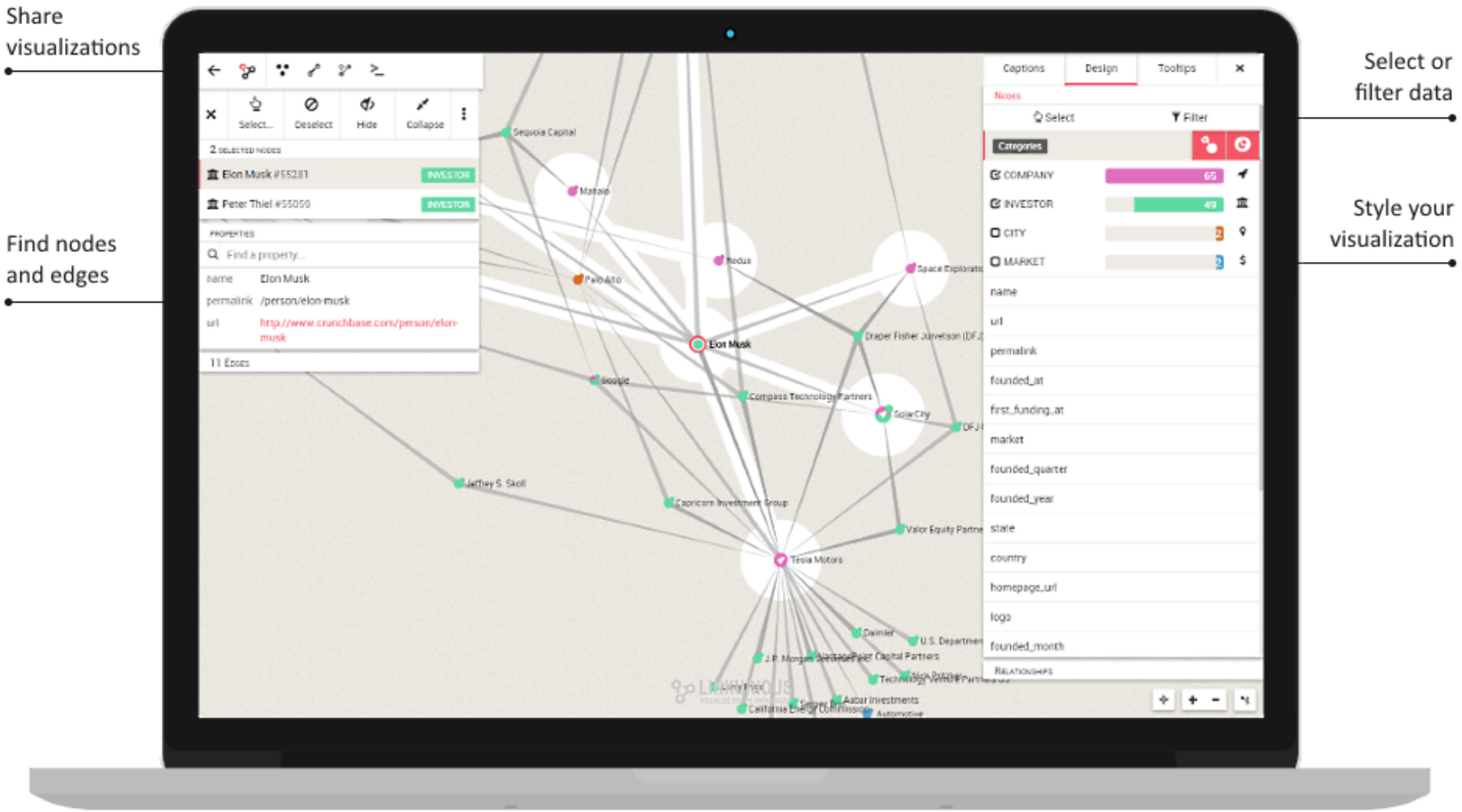
Linkurious : exploration de données (recherche intuitive), graph dataviz, analytics, analyses collaboratives.

Linkurious



Source : Linkurious

Linkurious

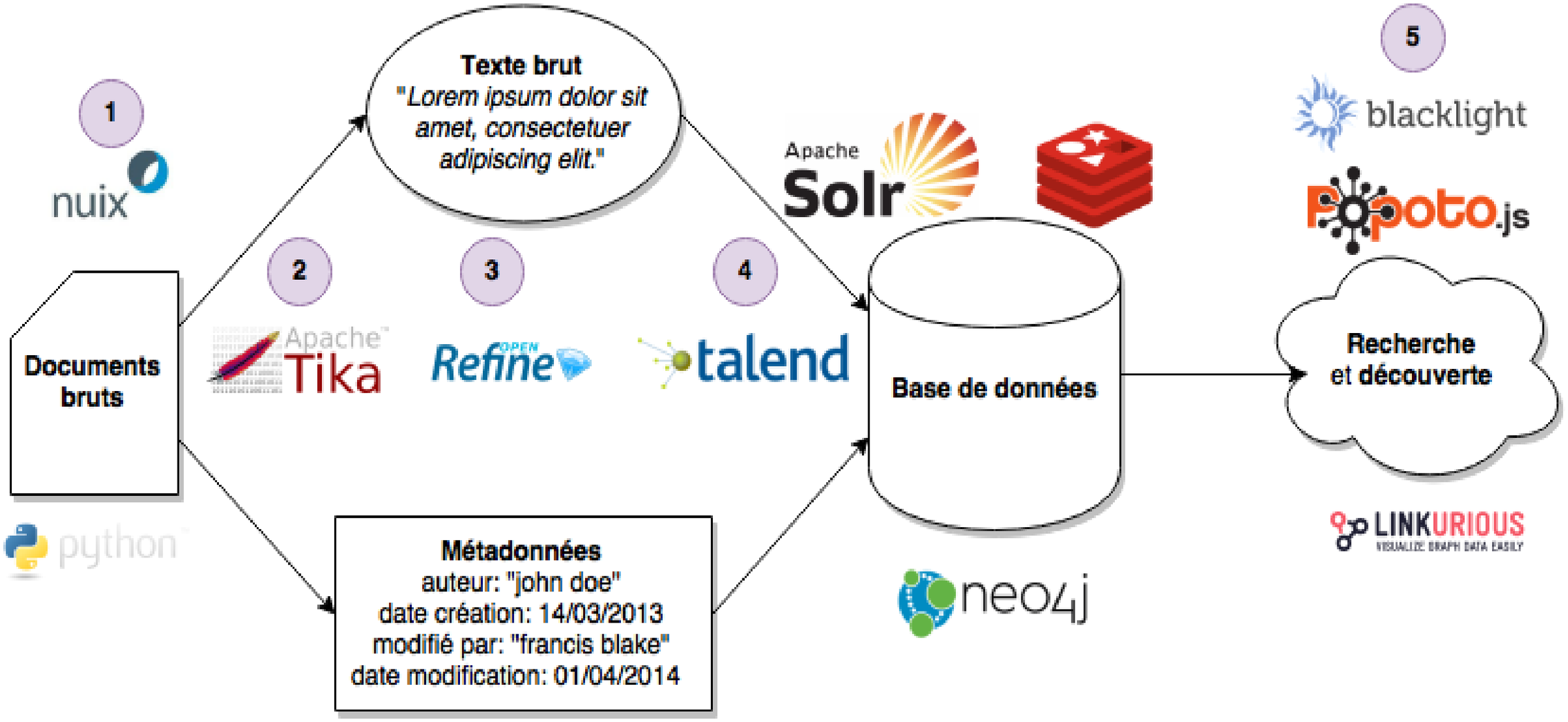


Source : Linkurious

5. Exploitation des données

Popoto.js : interface web de recherche orientée graphe.

Chaîne de traitement de l'information des Panama Papers



Méthodes d'investigation

Méthodes d'investigation

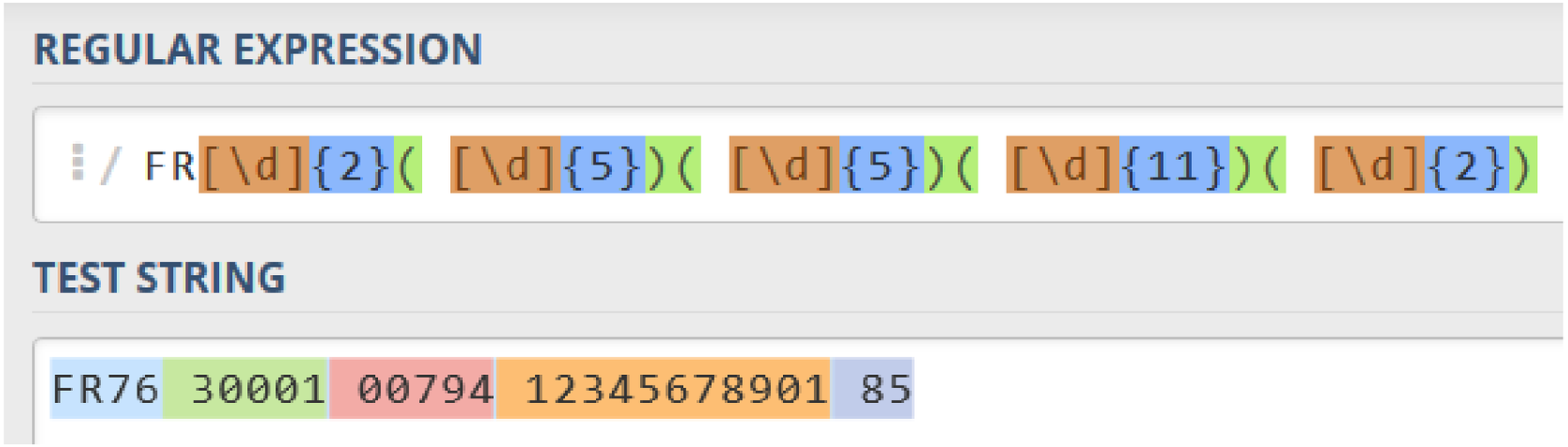
- Investigations à partir de listes :
  - `chefs d'états, ministres, parlementaires, criminels internationaux, athlètes professionnels, Forbes, Challenges`.
- Recherche par personnes liées :
  - `conjoints, passeports, "seconds couteaux"`.
- Processus itératif (résultat => nouvelle recherche).

Méthodes d'investigation

- Batch Search : recherche par lots (non interactive).
  - `Documents contenant le nom d'une des 500 familles les plus riches de France`.
  - Résultat sous la forme d'un fichier.
- Recherche par mots-clés et facettes (catégorisation du résultat).
  - `search: "(président" OR "ministre") AND "France"`
- Fuzzy Search : recherche approximative.
  - `search: "Jean Dupont"` => Dupond Jean, Jean Edouard Michel Dupont.

Méthodes d'investigation

- Recherche par expressions régulières :
  - IBAN = `FR[\d]{2}( [\d]{5})( [\d]{5})( [\d]{11})( [\d]{2})`



Méthodes d'investigation

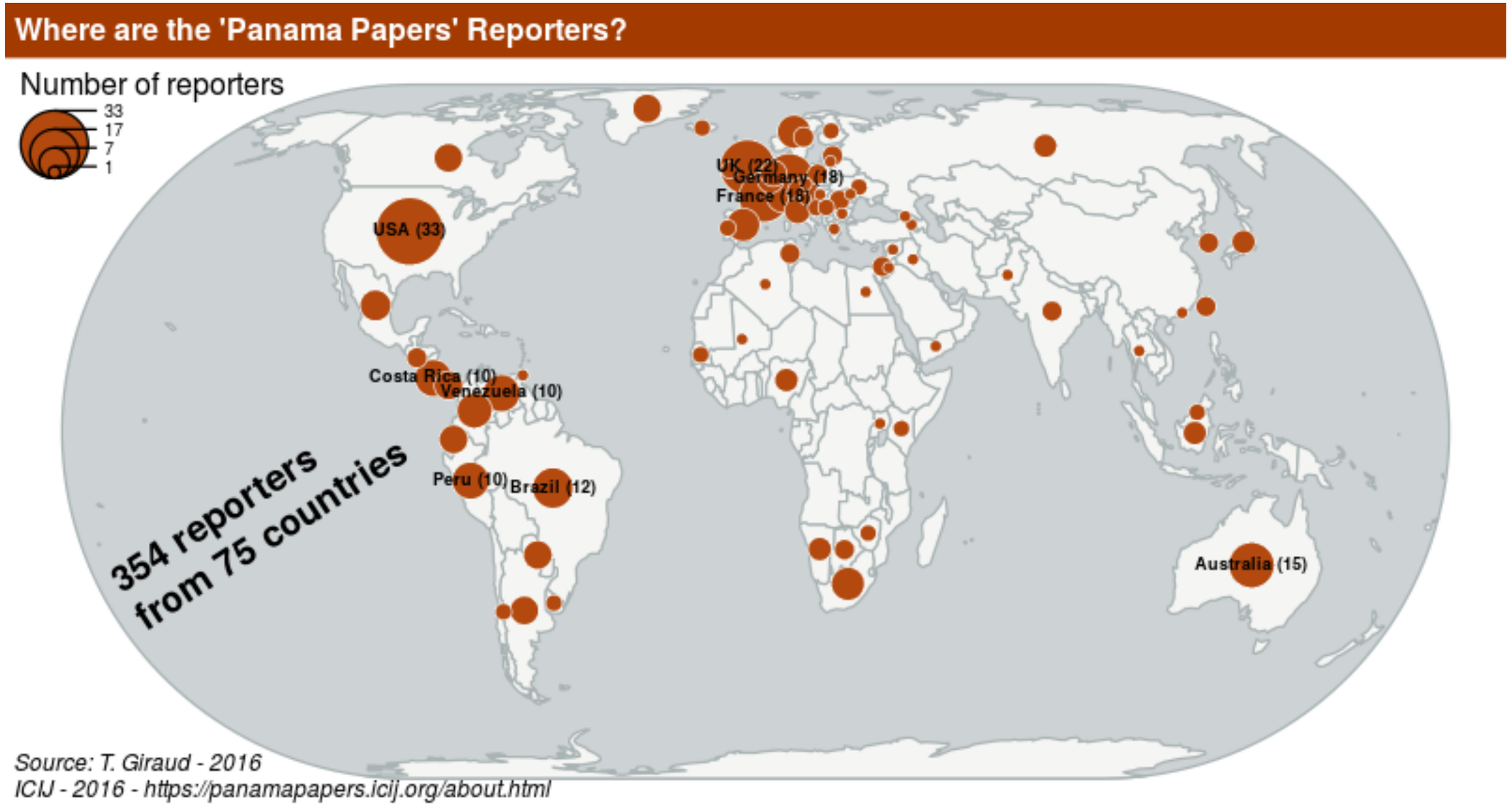
- Social Engineering (appel des personnes impliquées).
- Vérification croisée par crowd-sourcing (3 validations).
- Notes de synthèses rédigées par l'ICIJ (partager le même niveau d'information).

Organisation

Organisation

- Nom de code "Prometheus".
- 109 organismes de presse, 370 journalistes, 80 pays.
- 12 mois d'investigation.
- ICIJ = 12 personnes, 50% de l'équipe = Data & Research Unit.
- Langue d'échange : anglais et espagnol.

Organisation



Organisation

- Chiffrement du poste de travail :
  - Ordinateurs portables dédiés et cryptés (VeraCrypt).
  - Mail via PGP.
  - Messagerie mobile sécurisée Signal.
  - Authentification multi-facteur (Google Authenticator).
- Global i-Hub : forum interne à la façon d'un réseau social (développement maison basé sur Oxwall).

En France

- Le Monde (membre de l'ICIJ).
- Equipe de 12 personnes, dont 5 à plein temps.

Conclusion

Conclusion

- Données massives = besoin des technologies (big data) et infrastructures (cloud computing) appropriées.
- Complexité de l'information = investigation en mode collaboratif.
- Dream Team (expertise) :
    i. Technique (développement, déploiement).
    ii. Métier (finance, droit).
    iii. Statistique.
    iv. Journalisme.

Merci de votre attention.

Questions & Réponses