RJMI 2021

Recherche en intégration de données
Le cas de ConnectionLens

Nelly Barret – Doctorante depuis janvier 2021

Laboratoire Inria / École Polytechnique

21 février 2021

Présentation personnelle

Qui suis-je?

Lycée (Lyon)

- Baccalauréat Scientifique, spécialité informatique
  - Puissance 4

NIVERSITÉ
E LYON

Licence (Univ. de Lyon)

- Licence Informatique
  - Aide à la recherche immobilière

NIVERSITÉ
E LYON

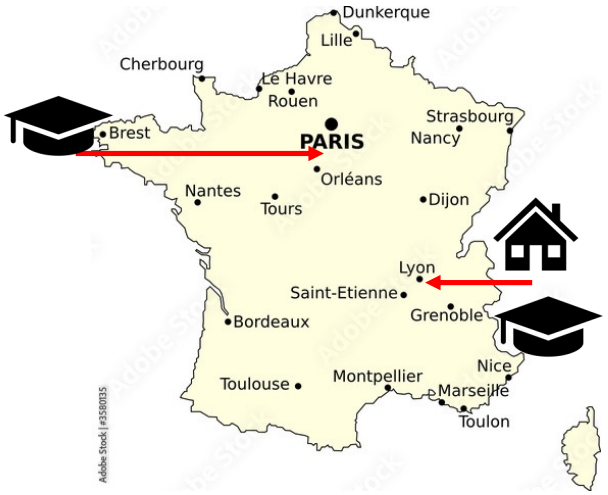
Master (Univ. de Lyon)

- Master Informatique, spécialité IA
  - Appariement et fusion d'entités spatiales
  - Prédiction de l'environnement des quartiers en France

ECHNIQUE
É PARIS-SACLAY

Doctorat (Inria/École Polytechnique)

- Sujet : intégration efficace et expressive de données hétérogènes



Qui suis-je?

2019

- À la recherche du quartier idéal
- Spatial entity matching with GeoAlign

2020

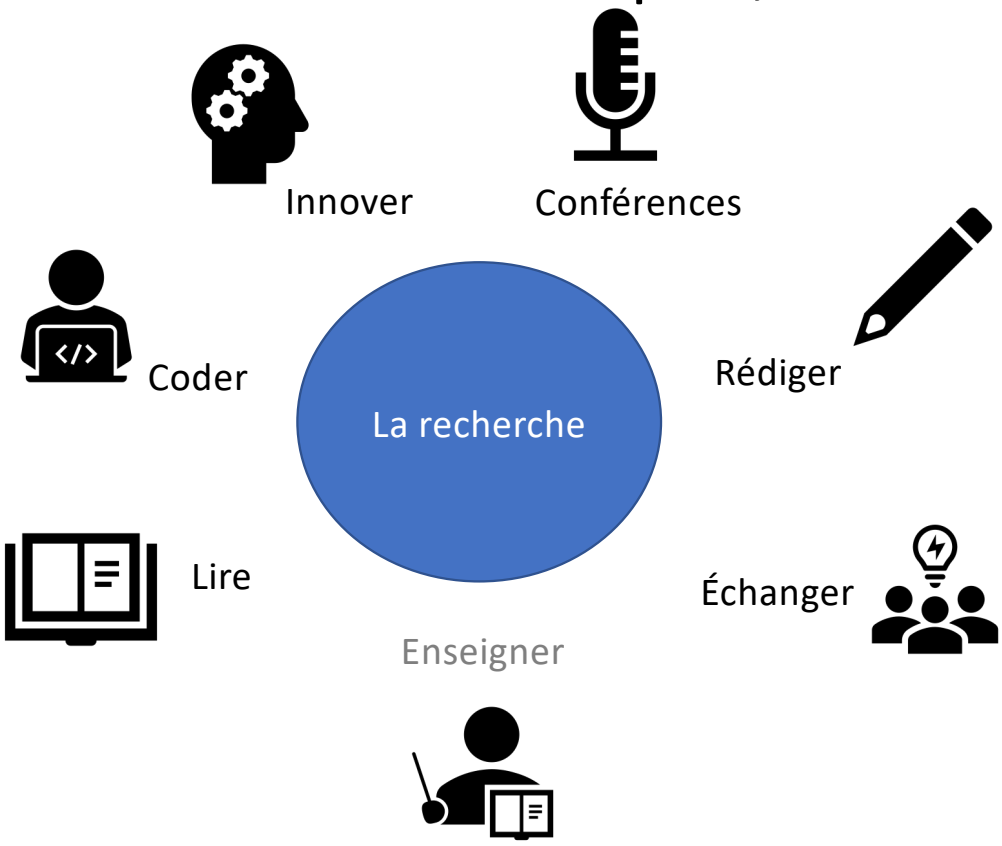
- Predicting the environment of a neighbourhood: a use case for France
- Predihood: an open-source tool for predicting neighbourhoods' information

2021

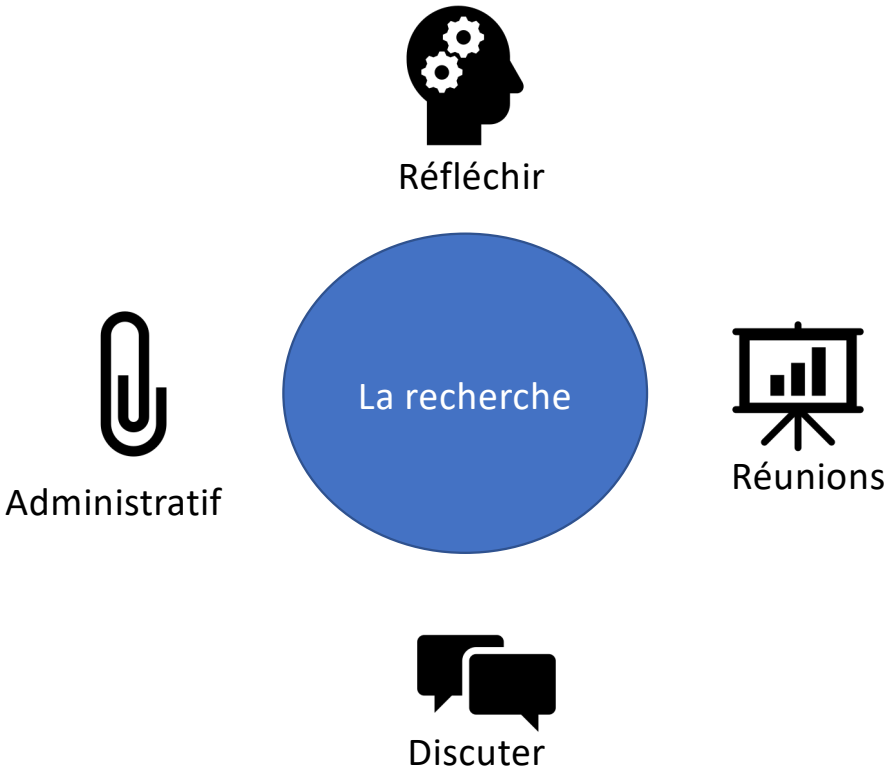
- An environmental Study of French neighbourhoods
- Toward Generic abstractions for data of any model
- Facilitating heterogeneous data understanding

La recherche en informatique

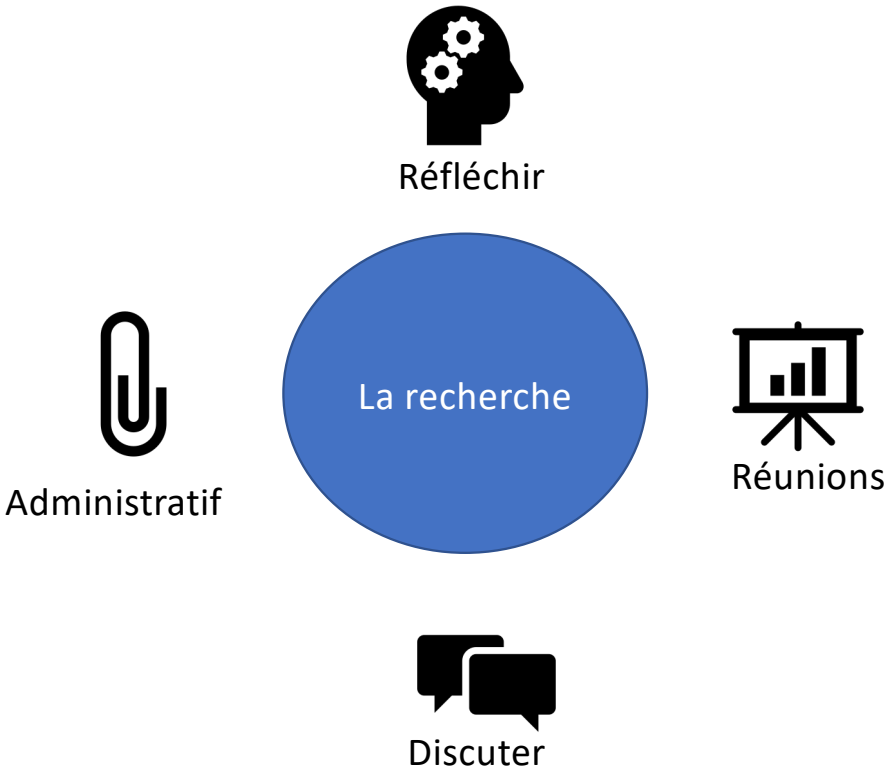
La recherche en informatique, c’est



La recherche en informatique, c’est aussi



La recherche en informatique, c’est aussi



Et bien sûr, prendre du plaisir à faire de la recherche ☺

Présentation de ConnectionLens

Qu’est-ce que ConnectionLens?

Qu’est-ce que ConnectionLens?

intégration efficace de données hétérogènes

Qu’est-ce que ConnectionLens?

intégration efficace de **données** hétérogènes

N’importe quelle
source de données:
- Base de données
- Fichier Excel
- Fichier texte
- Fichier JSON
- …

Qu’est-ce que ConnectionLens?

intégration efficace de **données hétérogènes**

N’importe quelle source de données:
- Base de données
- Fichier Excel
- Fichier texte
- Fichier JSON
- …

Des données qui :
- Proviennent de différents endroits
- Sont de différents formats
- Sont représentées différemment

Qu’est-ce que ConnectionLens?

intégration efficace de données hétérogènes

Mettre dans la même boîte dans un format commun

N’importe quelle source de données:
- Base de données
- Fichier Excel
- Fichier texte
- Fichier JSON
- …

Des données qui :
- Proviennent de différents endroits
- Sont de différents formats
- Sont représentées différemment

Qu’est-ce que ConnectionLens?

intégration **efficace** de données hétérogènes

Mettre dans la même boîte dans un format commun

Des données parfois très grandes (Méga, Giga)

N’importe quelle source de données:
- Base de données
- Fichier Excel
- Fichier texte
- Fichier JSON
- …

Des données qui :
- Proviennent de différents endroits
- Sont de différents formats
- Sont représentées différemment

Qu’est-ce que ConnectionLens?

intégration efficace de données hétérogènes

=

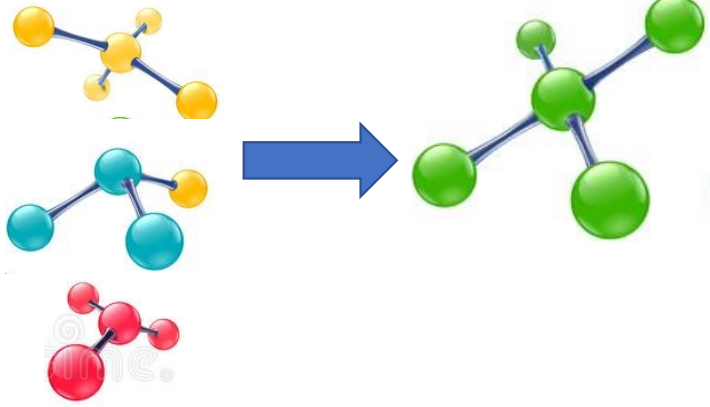
Mettre n’importe quelles données (quelque soit la provenance et le format), parfois très grandes, dans la même boîte et dans un format commun.

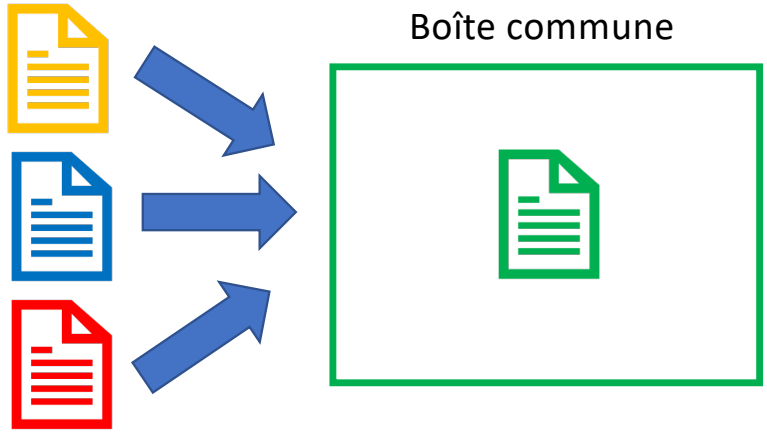
Comment fonctionne ConnectionLens ?

But : mettre les données dans la boîte commune qui contient un format générique, on appelle cela de l'intégration de données

Comment ?

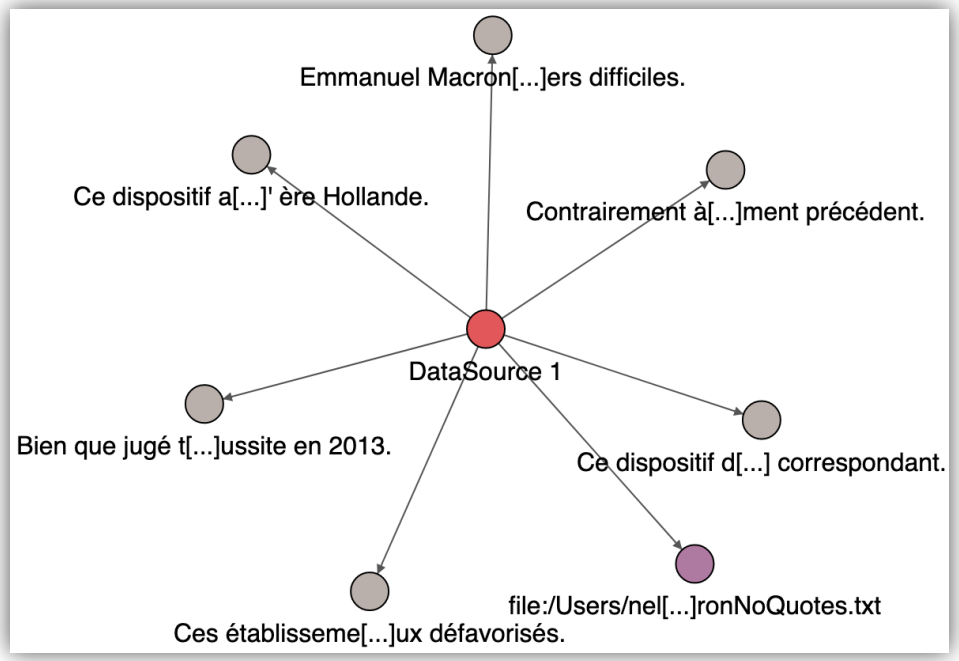
Transformer les données pour qu'elles s'adaptent au format de la boîte commune





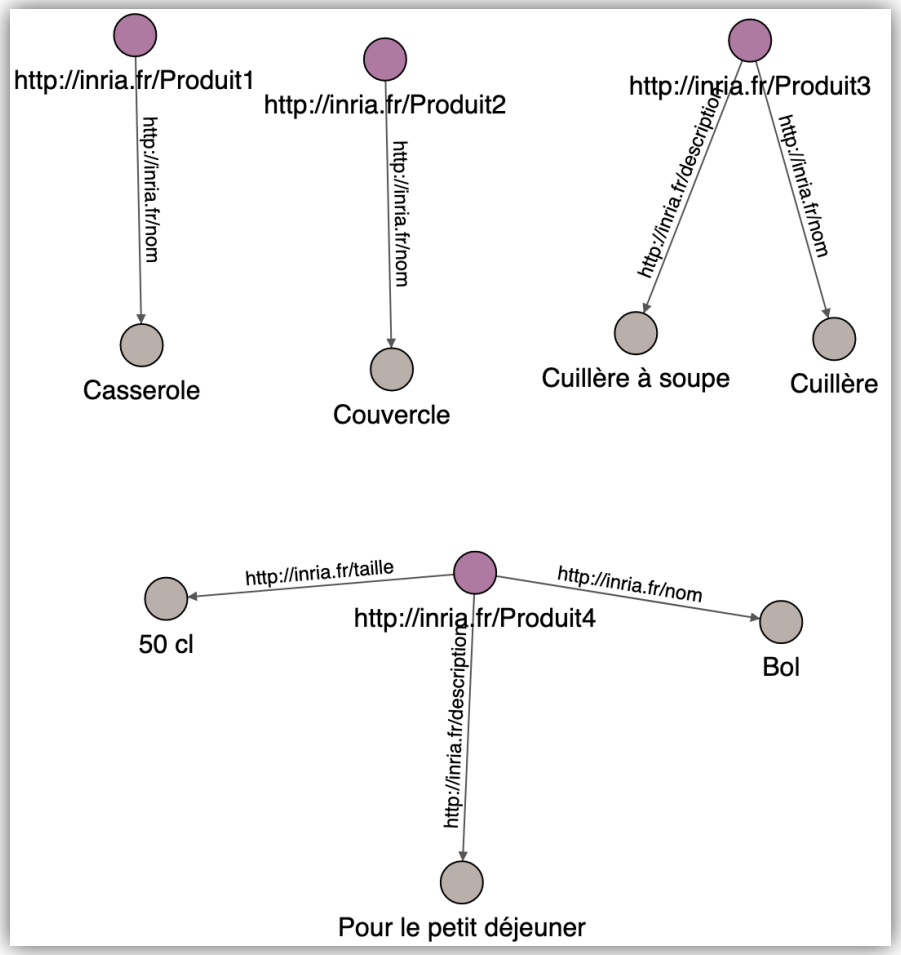
À quoi ressemble le format générique ?

Discours d’ E. Macron retranscrit par un journaliste dans ConnectionLens



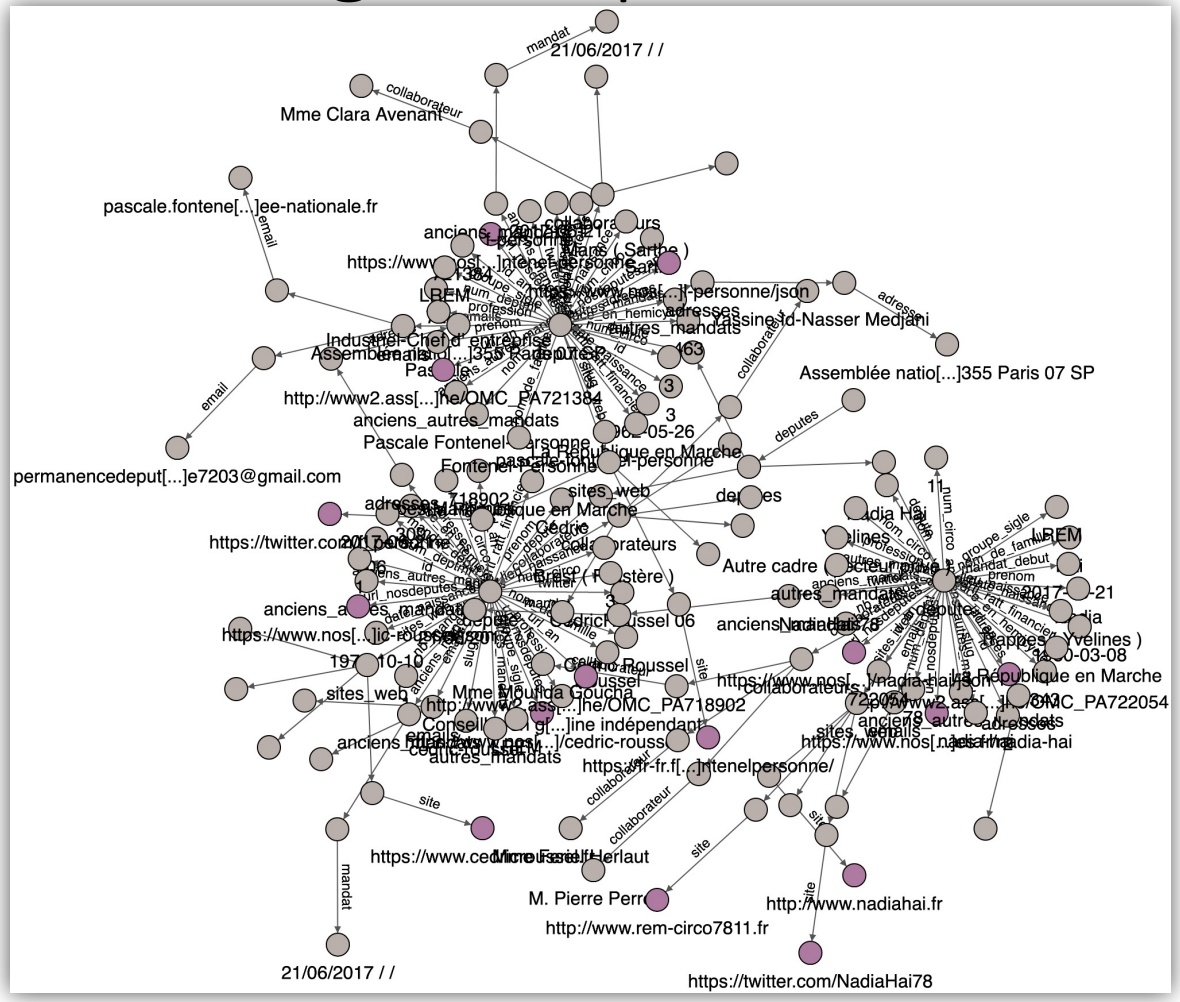
À quoi ressemble le format générique ?

Articles de cuisine avec des noms et des descriptions



À quoi ressemble le format générique ?

Document avec les informations de 3 députés français



Exemple : les Balkany

Villas à Marrakech, fonds « occultes »... : les époux Balkany jugés lundi

Soupçonnés d'avoir dissimulé 13 millions d'euros d'avoirs au fisc, les édiles de Levallois-Perret comparaissent pour fraude fiscale et blanchiment.

Source AFP

Affaire Balkany : l'ex-maire de Levallois incarcéré ce lundi, son épouse toujours à l'hôpital

Convoqué ce lundi matin à la gendarmerie de Vexin-sur-Epte (Eure), l'ancien maire de Levallois-Perret s'est vu signifier son incarcération, après la révocation de son placement sous bracelet électronique. Patrick Balkany va passer sa première nuit à la prison de Fleury-Mérogis (Essonne).

Enquête

Troisième villa saisie pour les Balkany

Après les propriétés de Giverny et de Saint-Martin, la justice a demandé la saisie de la villa «Dar Gyucy», à Marrakech, acquise en 2009 via de tortueux montages fiscaux.

Exemple : les Balkany

Public officials transparency high authority (CSV)

Name	Owner	Location	Type
Dar Gyucy	P. Balkany	Marrakech	Real Estate
Moulin Cossy	I. Balkany	Giverny	Real Estate

National Directory of Elected Officials (JSON)

```
[{
   name: "Levallois-Perret",
   mayor: "P. Balkany",
   city-council: [
     {name: "I. Balkany"},
      …
   ]
}, …]
```

dbpedia.org (RDF)

```
{
dbr:Marrakech
  dbr:name       "Marrakech"
  rdf:type       dbo:City ;
  dbo:country    dbr:Morrocco .
dbr:Morocco
  dbr:name       "Morocco"
  rdf:type       dbo:Country
  dbo:locatedIn  dbr:Africa .
dbr:CentralAfricanRepublic
  dbr:name       "Central African Republic"
  dbo:locatedIn  dbr:Africa .
}
```

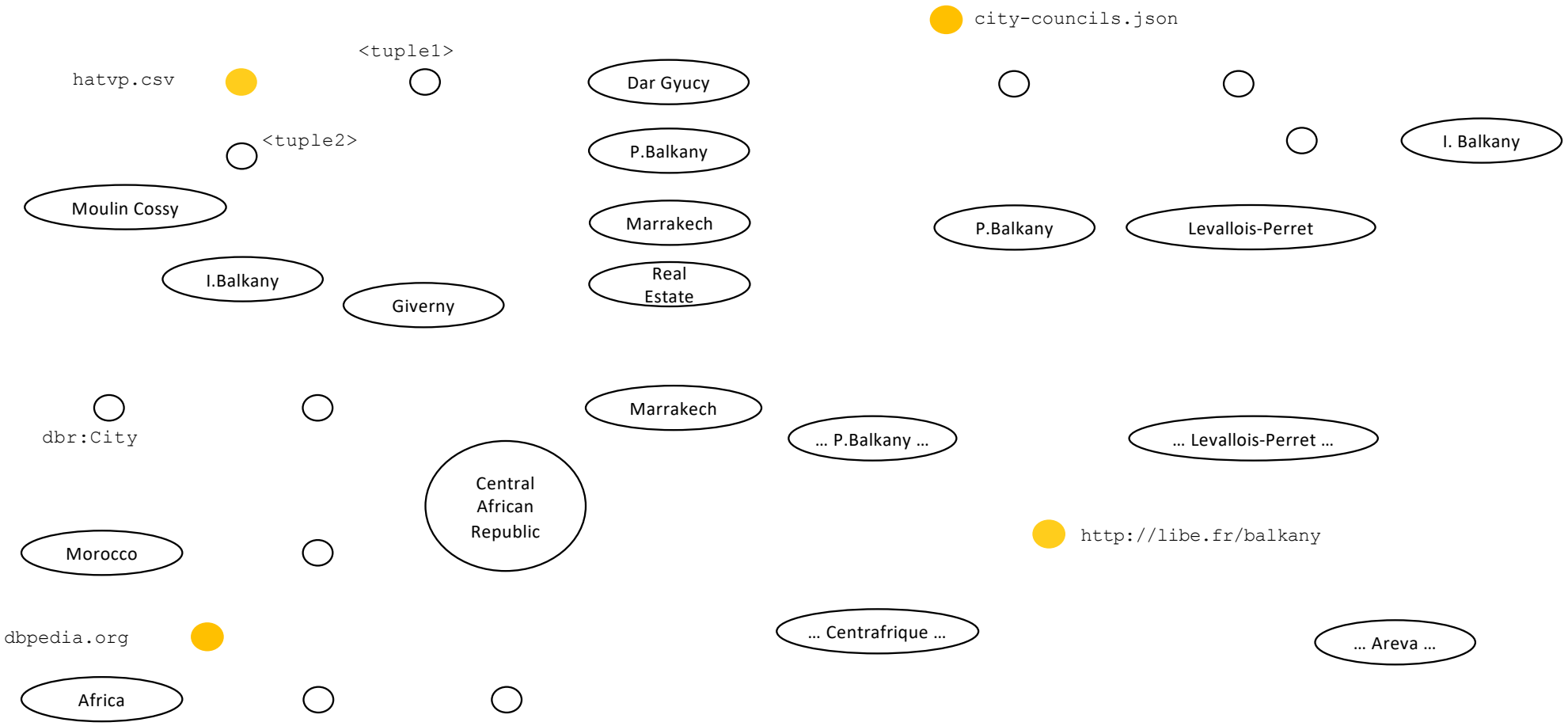
Libération – Nov. 13, 2014 (Text)

Balkany mineur de fonds

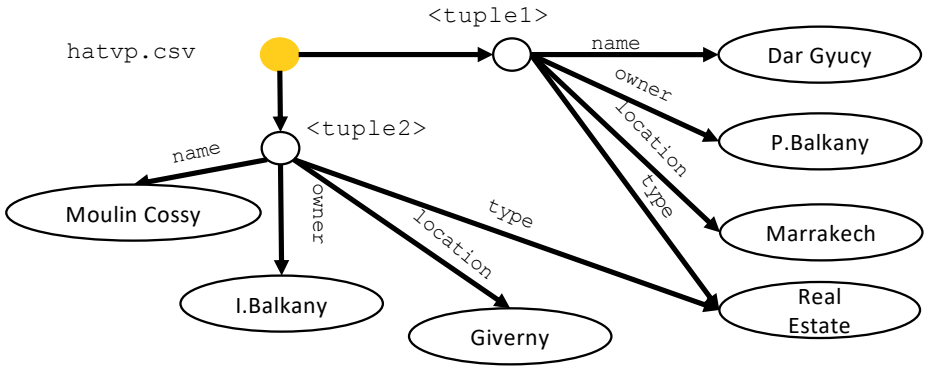
L'élu de Levallois-Perret est soupçonné d'avoir touché 5 millions de dollars de commission en 2009 grâce à son rôle d'intermédiaire entre Areva et la Centrafrique dans le dossier Uramin. […]

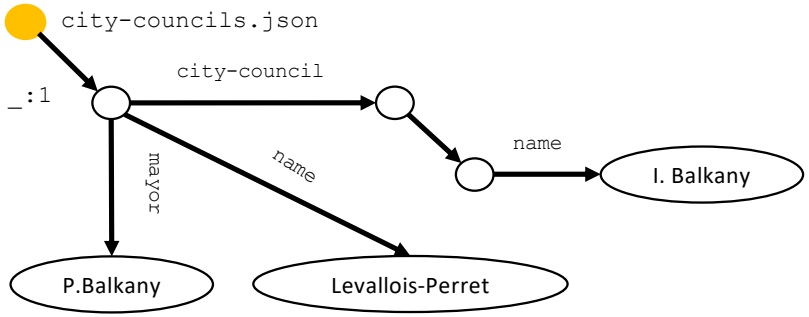
Exemple : les Balkany

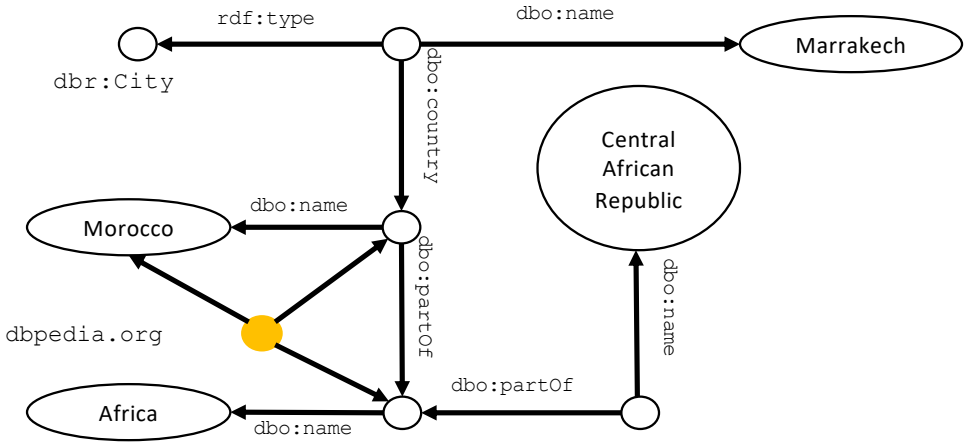
Exemple : les Balkany

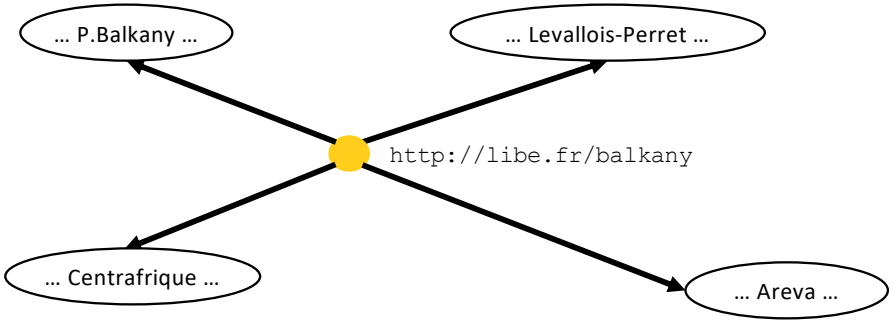


Exemple : les Balkany

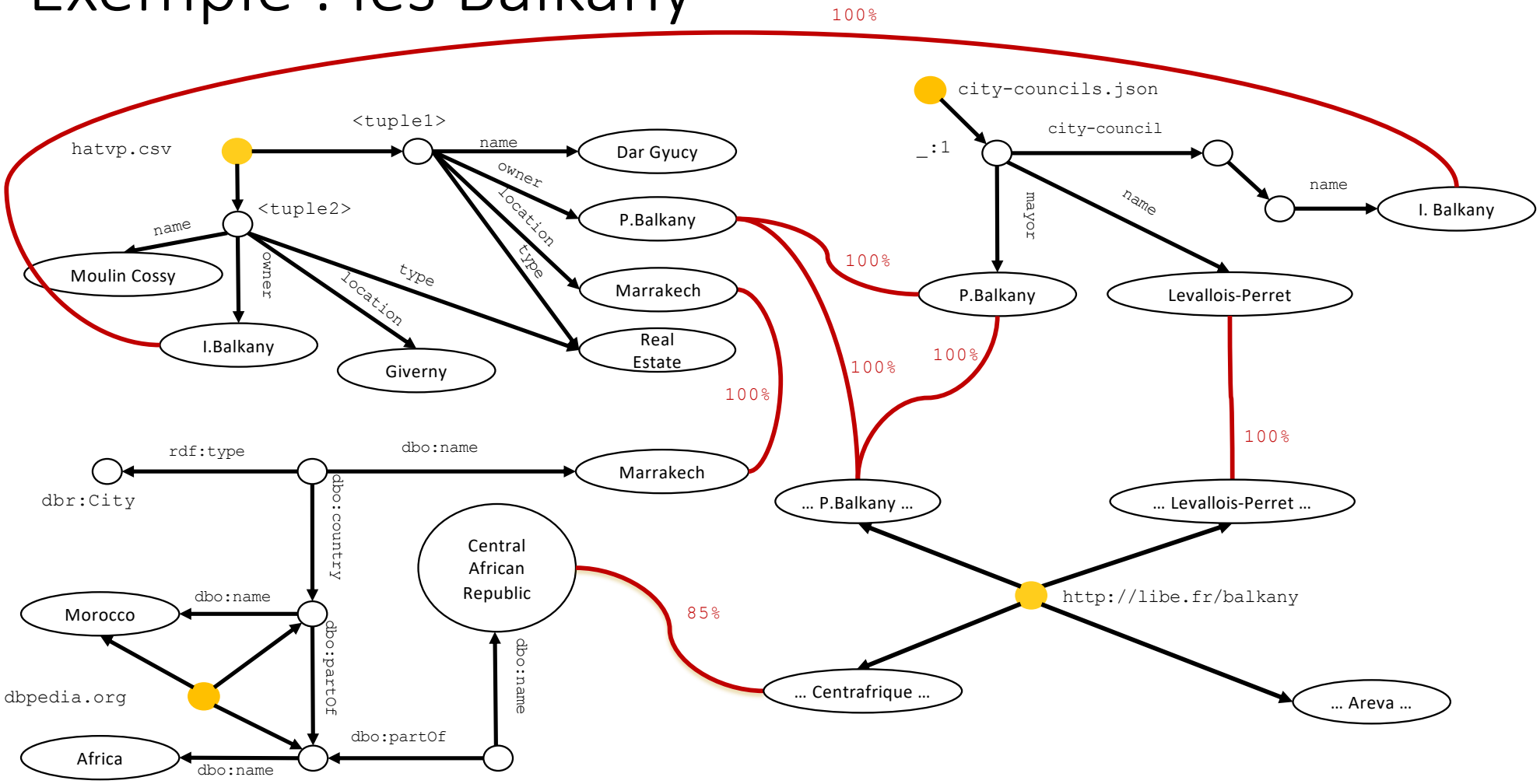








Exemple : les Balkany



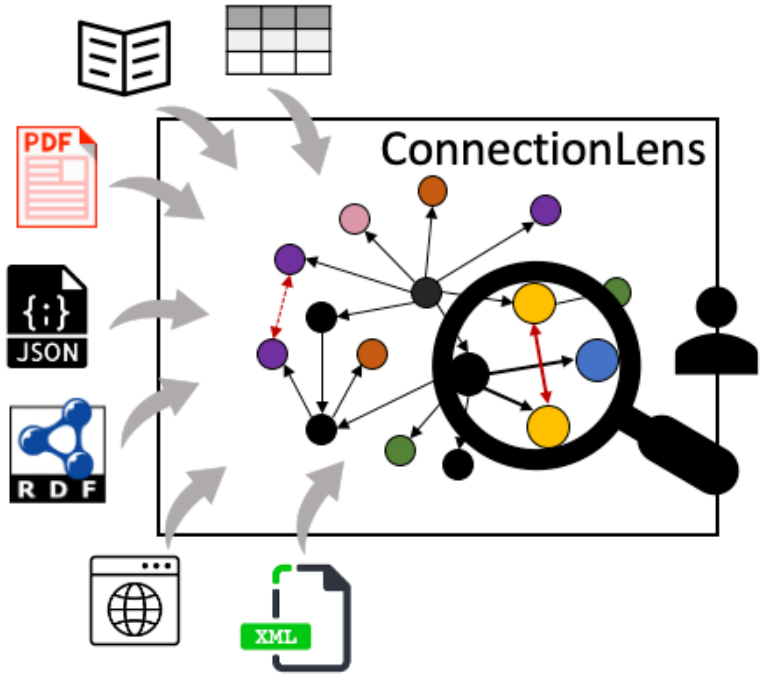
Applications de ConnectionLens

- Faire avancer la science :
  - Graph integration of structured, semistructured and unstructured data
- Trouver des conflits d'intérêts :
  - Discovering Conflicts of Interest across heterogeneous data sources
- Créer des descriptions de n'importe quelle donnée :
  - Toward Generic Abstractions for Data of Any Model
- Aide au « fact-checking » :
  - From Data to the Press: Data Management for Journalism and Fact-Checking
- Et bien d'autres…

Pour finir

- Prendre du plaisir à faire de la recherche

- ConnectionLens est un outil pour
  - Intégrer des données hétérogènes ensemble
  - De multiples applications : conflits d'intérêts, description automatique de données, fact-checking, journalisme…



Merci pour votre attention !

Des questions ?