

# Raisonner avec la provenance sur les données du Web

---

**Antoine Amarilli**

Concours Inria CR2

12 mai 2016

**2013–2016 :** Thèse à **Télécom ParisTech** avec **Pierre Senellart** :

- *Tirer parti de la structure des données incertaines*
- Soutenue le **14 mars 2016**

**2012–2013 :** **Pré-doc :**

- 3 mois à **Tel Aviv** avec Tova Milo
- 5 mois à **Oxford** avec Michael Benedikt

**2009–2013 :** **École normale supérieure** de Paris, master **MPRI**

- Vainqueur des **concours de programmation**  
Google Hash Code (2015) et Prologin (2008)

## Bases de données :

- **ICDT'14** (prédop à Tel Aviv)
- **PODS'16** (thèse)

## Logique et automates :

- **ICALP'15** (thèse)
- **LICS'15** (thèse)

## Intelligence artificielle :

- **IJCAI'15** (pré-doc à Oxford)
- **IJCAI'16** (avec Oxford)
- **7** autres publications internationales avec comité de lecture
- **1** brevet avec Google New York (stage de M1)

# Résumé des travaux antérieurs

---

Interrogation de données relationnelles incertaines

# Vue d'ensemble : Données relationnelles incertaines

Évaluer une *requête logique* sur une *base de données relationnelle*

**Problème** : On ne dispose pas toujours des données **exactes** :

- Données créées par des méthodes **faillibles** et **non-exhaustives**
- Données annotées par des techniques **d'apprentissage**
- Données **bruitées** ou **périmées**

→ Gérer les données relationnelles **avec leur incertitude**

**Problème** : Tâches souvent **complexes** voire **indécidables**

- **1.** données **incomplètes**
- **2.** données **probabilistes**

# Données incomplètes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont  
pendant mes congés?*

# Données incomplètes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés?*



# Données incomplètes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés?*



## Résultat :

| jour |
|------|
| 18   |

# Données incomplètes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés?*



# Données incomplètes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés?*

## Règles logiques :

- Je ne reviens pas pour **un seul jour**



# Données incomplètes

## Données :

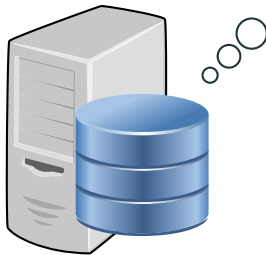
| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés ?*

## Règles logiques :

- Je ne reviens pas pour **un seul jour**



| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

# Données incomplètes

## Données :

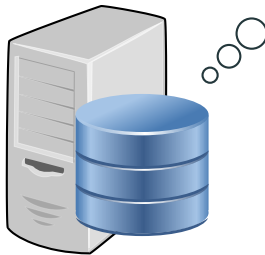
| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés ?*

## Règles logiques :

- Je ne reviens pas pour **un seul jour**



| jour      | type          |
|-----------|---------------|
| 9         | congés        |
| 10        | réunion       |
| 11        | congés        |
| 18        | congés        |
| 18        | réunion       |
| <b>10</b> | <b>congés</b> |

# Données incomplètes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés?*

## Règles logiques :

- Je ne reviens pas pour **un seul jour**



| jour      | type          |
|-----------|---------------|
| 9         | congés        |
| 10        | réunion       |
| 11        | congés        |
| 18        | congés        |
| 18        | réunion       |
| <b>10</b> | <b>congés</b> |

## Résultat :

| jour |
|------|
| 18   |

# Données incomplètes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés ?*

## Règles logiques :

- Je ne reviens pas pour **un seul jour**



| jour      | type          |
|-----------|---------------|
| 9         | congés        |
| 10        | réunion       |
| 11        | congés        |
| 18        | congés        |
| 18        | réunion       |
| <b>10</b> | <b>congés</b> |

## Résultat :

| jour      |
|-----------|
| 18        |
| <b>10</b> |

# Résumé : raisonner sur les données incomplètes

→ Problème fondamental en **intelligence artificielle** :

Quelles réponses à la **requête** de l'utilisateur sont vraies dans toutes les complétions des **données** qui satisfont des **règles logiques**?

**Approches existantes** : Langages de règles décidables en IA :

- Uniquement sur des **graphes de données**
- Ne considèrent pas spécifiquement les complétions **finies**

# Résumé : raisonner sur les données incomplètes

→ Problème fondamental en **intelligence artificielle** :

Quelles réponses à la **requête** de l'utilisateur sont vraies dans toutes les complétions des **données** qui satisfont des **règles logiques** ?

**Approches existantes** : Langages de règles décidables en IA :

- Uniquement sur des **graphes de données**
- Ne considèrent pas spécifiquement les complétions **finies**

→ J'ai transposé ces résultats aux **bases de données** :

- Étendre aux **hypergraphes** [Amarilli, Benedikt, **IJCAI'15**]
- Restreindre aux complétions **finies** [Amarilli, Benedikt, **LICS'15**]

# Données probabilistes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont  
pendant mes congés?*

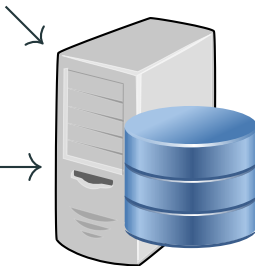
# Données probabilistes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont  
pendant mes congés?*



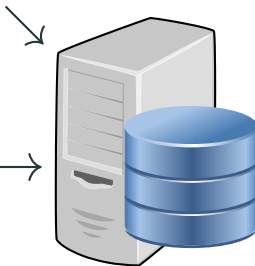
# Données probabilistes

## Données :

| jour | type    |
|------|---------|
| 9    | congés  |
| 10   | réunion |
| 11   | congés  |
| 18   | congés  |
| 18   | réunion |

## Requête logique :

*Quelles réunions sont pendant mes congés?*



## Résultat :

| jour |
|------|
| 18   |

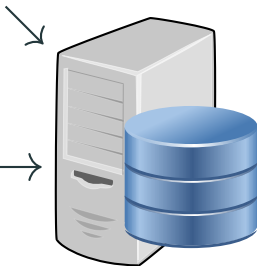
# Données probabilistes

## Données :

| jour | type    |     |
|------|---------|-----|
| 9    | congés  | 95% |
| 10   | réunion | 20% |
| 11   | congés  | 30% |
| 18   | congés  | 80% |
| 18   | réunion | 90% |

## Requête logique :

*Quelles réunions sont pendant mes congés?*



## Résultat :

jour

18

# Données probabilistes

## Données :

| jour | type    |     |
|------|---------|-----|
| 9    | congés  | 95% |
| 10   | réunion | 20% |
| 11   | congés  | 30% |
| 18   | congés  | 80% |
| 18   | réunion | 90% |

## Requête logique :

*Quelles réunions sont pendant mes congés?*



$$90\% \times 80\% =$$

## Résultat :

| jour |
|------|
| 18   |

# Données probabilistes

## Données :

| jour | type    |     |
|------|---------|-----|
| 9    | congés  | 95% |
| 10   | réunion | 20% |
| 11   | congés  | 30% |
| 18   | congés  | 80% |
| 18   | réunion | 90% |

## Requête logique :

*Quelles réunions sont pendant mes congés?*



$$90\% \times 80\% \\ = 72\%$$

## Résultat :

| jour |
|------|
| 18   |

# Données probabilistes

## Données :

| jour | type    |     |
|------|---------|-----|
| 9    | congés  | 95% |
| 10   | réunion | 20% |
| 11   | congés  | 30% |
| 18   | congés  | 80% |
| 18   | réunion | 90% |

## Requête logique :

*Quelles réunions sont pendant mes congés?*



$$90\% \times 80\% = 72\%$$

## Résultat :

| jour |     |
|------|-----|
| 18   | 72% |

# Données probabilistes : travaux existants

→ Problème d'évaluation de requêtes sur données **probabilistes** :

Quelle est la **probabilité totale** de chaque réponse quand les faits sont présents ou absents **indépendamment** avec la **probabilité** indiquée ?

# Données probabilistes : travaux existants

→ Problème d'évaluation de requêtes sur données **probabilistes** :

Quelle est la **probabilité totale** de chaque réponse quand les faits sont présents ou absents **indépendamment** avec la **probabilité** indiquée ?

**Approche existante** (intensionnelle) :

| Calendrier |      |         |
|------------|------|---------|
| id         | jour | type    |
| $t_1$      | 9    | congés  |
| $t_2$      | 10   | réunion |
| $t_3$      | 11   | congés  |
| $t_4$      | 18   | congés  |
| $t_5$      | 18   | réunion |

**Requête** conjonctive

*Y a-t-il une réunion pendant mes congés ?*

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$

$\wedge \text{Calendrier}(t', d, \text{"réunion"})$

# Données probabilistes : travaux existants

→ Problème d'évaluation de requêtes sur données **probabilistes** :

Quelle est la **probabilité totale** de chaque réponse quand les faits sont présents ou absents **indépendamment** avec la **probabilité** indiquée ?

**Approche existante** (intensionnelle) :

| Calendrier |      |         |
|------------|------|---------|
| id         | jour | type    |
| $t_1$      | 9    | congés  |
| $t_2$      | 10   | réunion |
| $t_3$      | 11   | congés  |
| $t_4$      | 18   | congés  |
| $t_5$      | 18   | réunion |

**Requête conjonctive**

Y a-t-il une réunion pendant mes congés ?

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

**PTIME**

Formule de  
**provenance**

$t_4 \wedge t_5$

# Données probabilistes : travaux existants

→ Problème d'évaluation de requêtes sur données **probabilistes** :

Quelle est la **probabilité totale** de chaque réponse quand les faits sont présents ou absents **indépendamment** avec la **probabilité** indiquée ?

**Approche existante** (intensionnelle) :

| Calendrier |      |         |     |  |
|------------|------|---------|-----|--|
| id         | jour | type    |     |  |
| $t_1$      | 9    | congés  | 95% |  |
| $t_2$      | 10   | réunion | 20% |  |
| $t_3$      | 11   | congés  | 30% |  |
| $t_4$      | 18   | congés  | 80% |  |
| $t_5$      | 18   | réunion | 90% |  |

**Requête conjonctive**

Y a-t-il une réunion pendant mes congés ?

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

**PTIME**

Formule de  
**provenance**

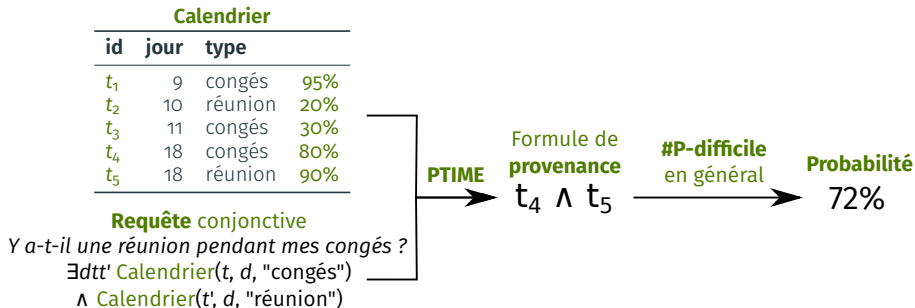
$t_4 \wedge t_5$

# Données probabilistes : travaux existants

→ Problème d'évaluation de requêtes sur données **probabilistes** :

Quelle est la **probabilité totale** de chaque réponse quand les faits sont présents ou absents **indépendamment** avec la **probabilité** indiquée ?

**Approche existante** (intensionnelle) :



# Données probabilistes : résultat de dichotomie

→ J'ai montré comment exploiter la **structure des données** :

**Théorème** [Amarilli, Bourhis, Senellart, **ICALP'15**]

L'évaluation de requêtes **MSO** est faisable en temps **linéaire** sur des données probabilistes de **largeur d'arbre bornée**

→ En un sens, ce résultat ne peut pas être **amélioré** (dichotomie) :

**Théorème** [Amarilli, Bourhis, Senellart, **PODS'16**]

L'évaluation probabiliste de certaines requêtes **FO** est **#P-difficile** (sous conditions) sur **n'importe quelle** famille de graphes de **largeur d'arbre non bornée**

# Données probabilistes : preuve de la borne supérieure

Généraliser les résultats de [Courcelle, 1990] à la provenance

## Données de largeur bornée

Calendrier

| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |
| $t_4$ | 18   | congés  |
| $t_5$ | 18   | réunion |

### Requête MSO

*Y a-t-il une réunion  
pendant mes congés ?*

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

# Données probabilistes : preuve de la borne supérieure

Généraliser les résultats de [Courcelle, 1990] à la provenance

## Données de largeur bornée

Calendrier

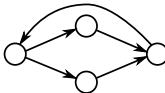
| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |
| $t_4$ | 18   | congés  |
| $t_5$ | 18   | réunion |

### Requête MSO

Y a-t-il une réunion  
pendant mes congés ?

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

### Automate d'arbres



# Données probabilistes : preuve de la borne supérieure

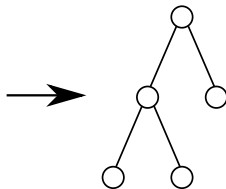
Généraliser les résultats de [Courcelle, 1990] à la provenance

**Données** de largeur bornée

**Calendrier**

| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |
| $t_4$ | 18   | congés  |
| $t_5$ | 18   | réunion |

**Encodage** en arbre

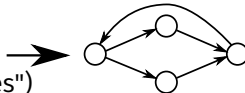


**Requête** MSO

*Y a-t-il une réunion pendant mes congés ?*

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

**Automate** d'arbres



# Données probabilistes : preuve de la borne supérieure

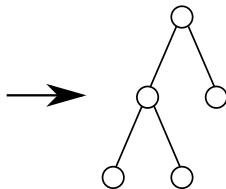
Généraliser les résultats de [Courcelle, 1990] à la provenance

**Données** de largeur bornée

**Calendrier**

| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |
| $t_4$ | 18   | congés  |
| $t_5$ | 18   | réunion |

**Encodage** en arbre



**linéaire**

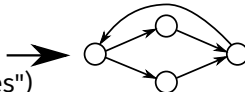
Réponse  
TRUE

**Requête MSO**

Y a-t-il une réunion  
pendant mes congés ?

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

**Automate** d'arbres



# Données probabilistes : preuve de la borne supérieure

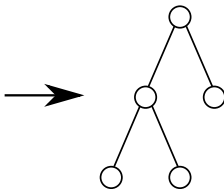
Généraliser les résultats de [Courcelle, 1990] à la provenance

**Données** de largeur bornée

**Calendrier**

| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |
| $t_4$ | 18   | congés  |
| $t_5$ | 18   | réunion |

**Encodage** en arbre

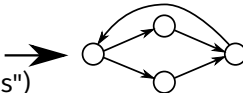


**Requête** MSO

*Y a-t-il une réunion  
pendant mes congés ?*

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

**Automate** d'arbres



# Données probabilistes : preuve de la borne supérieure

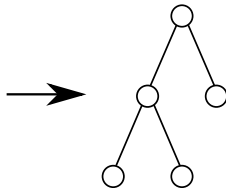
Généraliser les résultats de [Courcelle, 1990] à la provenance

**Données** de largeur bornée

**Calendrier**

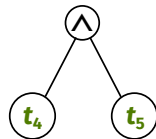
| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |
| $t_4$ | 18   | congés  |
| $t_5$ | 18   | réunion |

**Encodage** en arbre



**linéaire**

**Circuit de provenance**

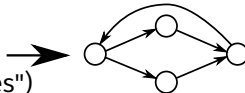


**Requête MSO**

*Y a-t-il une réunion pendant mes congés ?*

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

**Automate d'arbres**



# Données probabilistes : preuve de la borne supérieure

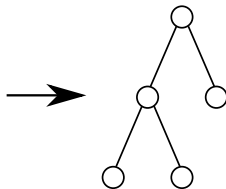
Généraliser les résultats de [Courcelle, 1990] à la provenance

**Données** de largeur bornée

**Calendrier**

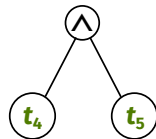
| id    | jour | type    |     |
|-------|------|---------|-----|
| $t_1$ | 9    | congés  | 95% |
| $t_2$ | 10   | réunion | 20% |
| $t_3$ | 11   | congés  | 30% |
| $t_4$ | 18   | congés  | 80% |
| $t_5$ | 18   | réunion | 90% |

**Encodage** en arbre



**linéaire**

**Circuit de provenance**

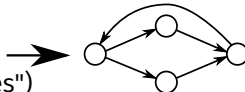


**Requête MSO**

*Y a-t-il une réunion pendant mes congés ?*

$\exists dt t' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

**Automate d'arbres**



# Données probabilistes : preuve de la borne supérieure

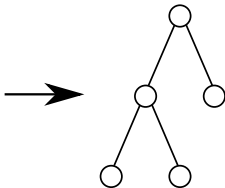
Généraliser les résultats de [Courcelle, 1990] à la provenance

**Données** de largeur bornée

**Calendrier**

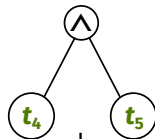
| id    | jour | type    |     |
|-------|------|---------|-----|
| $t_1$ | 9    | congés  | 95% |
| $t_2$ | 10   | réunion | 20% |
| $t_3$ | 11   | congés  | 30% |
| $t_4$ | 18   | congés  | 80% |
| $t_5$ | 18   | réunion | 90% |

**Encodage** en arbre



**linéaire**

**Circuit de provenance**



**linéaire**

**72%**

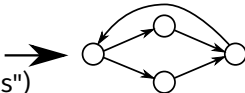
**Probabilité**

**Requête MSO**

*Y a-t-il une réunion pendant mes congés ?*

$\exists dt' \text{ Calendrier}(t, d, \text{"congés"})$   
 $\wedge \text{Calendrier}(t', d, \text{"réunion"})$

**Automate d'arbres**



# Projet de recherche

---

Raisonner avec la provenance sur les données du Web

# Web des données

- Bases de connaissances : Wikidata, YAGO, etc.

- Bases de connaissances : **Wikidata**, YAGO, etc.

## Inria (Q1146208)

Institut national de recherche en  
informatique et en automatique

► [Plus de langues](#)

### Déclarations

nature de  
l'élément



institut de recherche

▼ 0 référence

+ [ajouter](#)

Wikipédia [modifier](#)



fi

[INRIA](#)

fr

[Institut national de recherche en informatique  
et en automatique](#)

ja

[フランス国立情報学自動制御研究所](#)

date de fondation/création



1967

► 1 référence

+ [ajouter](#)

# Web des données

- Bases de connaissances : **Wikidata**, YAGO, etc.



# Web des données

- Bases de connaissances : **Wikidata**, YAGO, etc.
- Données ouvertes : **data.gouv.fr**, etc.

🏠 / Jeux de données 1 à 20 sur 25484

Trier par  

**Recensement des équipements sportifs, espaces et sites de pratiques**  
 01/01/2005 à 16/02/2015  Trimestrielle  France

**Population**  
 01/1901 à 07/2015  Mensuelle  France

**Liste et localisation des Musées de France**  
 2011 à 2014  Annuelle  France  Autre  22

# Web des données

- Bases de connaissances : **Wikidata**, YAGO, etc.
- Données ouvertes : **data.gouv.fr**, etc.

🏠 / Jeux de données 1 à 20 sur 25484

Trier par  

**Recensement des équipements sportifs, espaces et sites de pratiques**  
📅 01/01/2005 à 16/02/2015 ⌚ Trimestrielle 📍 France

**Population**  
📅 01/1901 à 07/2015 ⌚ Mensuelle 📍 France

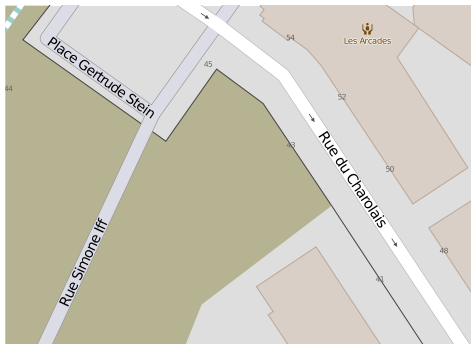
**Liste et localisation des Musées de France**  
📅 2011 à 2014 ⌚ Annuelle 📍 France 🌐 Autre ↺ 22



|     |                                                                                |
|-----|--------------------------------------------------------------------------------|
| CSV | <b>vehicules 2014.csv</b><br>Dernière modification le jeudi 6 août 2015        |
| CSV | <b>usagers 2014.csv</b><br>Dernière modification le jeudi 6 août 2015          |
| CSV | <b>lieux 2014.csv</b><br>Dernière modification le jeudi 6 août 2015            |
| CSV | <b>caracteristiques 2014.csv</b><br>Dernière modification le jeudi 6 août 2015 |
| CSV | <b>vehicules 2013.csv</b><br>Dernière modification le jeudi 6 août 2015        |

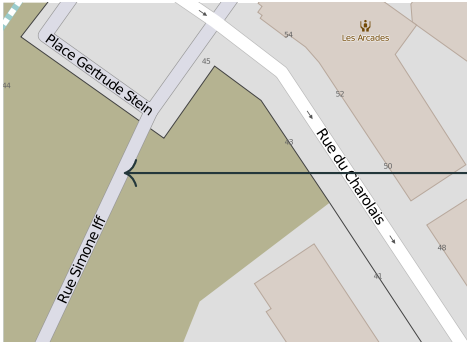
# Web des données

- Bases de connaissances : **Wikidata**, YAGO, etc.
- Données ouvertes : **data.gouv.fr**, etc.
- Données géographiques : **OpenStreetMaps**, etc.



# Web des données

- Bases de connaissances : **Wikidata**, YAGO, etc.
- Données ouvertes : **data.gouv.fr**, etc.
- Données géographiques : **OpenStreetMaps**, etc.



|                                  |                                                                                 |
|----------------------------------|---------------------------------------------------------------------------------|
| <a href="#">addr.city</a>        | Paris                                                                           |
| <a href="#">addr.housenumber</a> | 2                                                                               |
| <a href="#">addr.postcode</a>    | 75012                                                                           |
| <a href="#">addr.street</a>      | Rue Simone Iff                                                                  |
| <a href="#">name</a>             | Centre de recherche<br>Inria de Paris                                           |
| <a href="#">office</a>           | <a href="#">research</a>                                                        |
| <a href="#">opening_hours</a>    | Mo-Fr 07:30-19:00                                                               |
| <a href="#">phone</a>            | +33 1 80 49 40 00                                                               |
| <a href="#">website</a>          | <a href="http://www.inria.fr/centre/paris">http://www.inria.fr/centre/paris</a> |

# Web des données

- Bases de connaissances : [Wikidata](#), YAGO, etc.
- Données ouvertes : [data.gouv.fr](#), etc.
- Données géographiques : [OpenStreetMaps](#), etc.
- Annotations sémantiques : [Web Data Commons](#), etc.

|                             |                |
|-----------------------------|----------------|
| <b>Crawl Date</b>           | Winter 2014    |
| <b>Total Data</b>           | 64 Terabyte    |
| <b>Parsed HTML URLs</b>     | 2,014,175,679  |
| <b>URLs with Triples</b>    | 620,151,400    |
| <b>Domains in Crawl</b>     | 15,668,667     |
| <b>Domains with Triples</b> | 2,722,425      |
| <b>Typed Entities</b>       | 5,516,068,263  |
| <b>Triples</b>              | 20,484,755,485 |

# Web des données

- Bases de connaissances : **Wikidata**, YAGO, etc.
- Données ouvertes : **data.gouv.fr**, etc.
- Données géographiques : **OpenStreetMaps**, etc.
- Annotations sémantiques : **Web Data Commons**, etc.

**Objectif** : combiner et utiliser ces données :

- Répondre à des requêtes logiques **complexes**
- Calculer des **visualisations** et des **statistiques**
- **Recouper** des informations ou trouver des **contradictions**
- Connecter ses **propres données** aux **jeux de données existants**

# Web des données

- Bases de connaissances : **Wikidata**, YAGO, etc.
- Données ouvertes : **data.gouv.fr**, etc.
- Données géographiques : **OpenStreetMaps**, etc.
- Annotations sémantiques : **Web Data Commons**, etc.

**Objectif** : combiner et utiliser ces données :

- Répondre à des requêtes logiques **complexes**
- Calculer des **visualisations** et des **statistiques**
- **Recouper** des informations ou trouver des **contradictions**
- Connecter ses **propres données** aux **jeux de données existants**

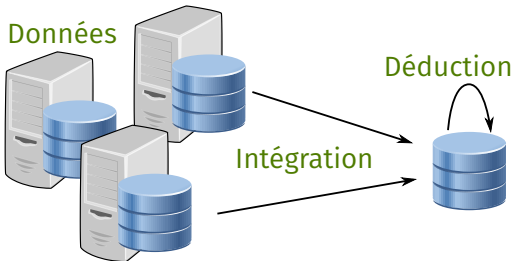
Ces données posent de nombreuses **difficultés** :

- **hétérogènes**    • **distribuées**    • **incomplètes**    • **peu fiables**

# Problème 1 : Intégration

Données **incomplètes** et **hétérogènes** issues de sources **multiples**

- Raisonnement avec des règles **logiques**
  - **Intégrer** les différences sources
  - **Déduire** les faits manquants



→ **Approches existantes** : OBDA, data integration, data exchange...

## Problème 2 : Fiabilité

Données produites **collaborativement** ou par des processus **faillibles**

- **Vandalisme**

## Problème 2 : Fiabilité

Données produites **collaborativement** ou par des processus **faillibles**

- **Vandalisme**

### Dans **Wikidata**

**Version du 1 avril 2014 à 07:22 (remercier)**

[Qlsusu](#) ([discussion](#) | [contributions](#))

(Affirmation ajoutée : *identifiant bibliothèque du Congrès (P244)*:

*Michael Jackson tops global download sales list*)

[Modification suivante](#) →

**propriété / identifiant bibliothèque du Congrès**

+

Michael Jackson tops global download sales list

## Problème 2 : Fiabilité

Données produites **collaborativement** ou par des processus **faillibles**

- **Vandalisme**

### Dans Wikidata

 This is your **last warning**. The next time you harm Wikidata, you **may be blocked from editing without further notice**.

**Version du 1 avril 2014 à 07:22 (remercier)**

[Qlsusu](#) (discussion | contributions)

(Affirmation ajoutée : [identifiant bibliothèque du Congrès \(P244\)](#):

*Michael Jackson tops global download sales list*)

[Modification suivante](#) →

**propriété / identifiant bibliothèque du Congrès**

+

Michael Jackson tops global download sales list

# Problème 2 : Fiabilité

Données produites **collaborativement** ou par des processus **faillibles**

- **Vandalisme**
- **Controverses**

## Dans Wikidata

### Crimée (Q7835)

péninsule d'Ukraine qui s'avance dans la mer Noire

[/modifier](#)

Aucun alias défini.

[• Plus de langues](#)

|                           |                                                                                           |                                         |
|---------------------------|-------------------------------------------------------------------------------------------|-----------------------------------------|
| pays                      |  Russie  | <a href="#">/modifier</a>               |
|                           | - 0 référence                                                                             | <a href="#">• ajouter une référence</a> |
|                           |  Ukraine | <a href="#">/modifier</a>               |
|                           | - 0 référence                                                                             | <a href="#">• ajouter une référence</a> |
| <a href="#">• ajouter</a> |                                                                                           |                                         |

## Problème 2 : Fiabilité

Données produites **collaborativement** ou par des processus **faillibles**

- **Vandalisme**
- **Controverses**
- **Extraction**

### Dans **YAGO**



#### Macquarie Group Limited

|                                                                                                    |                                                            |
|----------------------------------------------------------------------------------------------------|------------------------------------------------------------|
|  <b>MACQUARIE</b> |                                                            |
| <b>Type</b>                                                                                        | Public                                                     |
| <b>Founded</b>                                                                                     | <a href="#">Sydney</a> , New South Wales, Australia (1970) |
| <b>Revenue</b>                                                                                     | <a href="#">AUD</a> A\$8.1 billion (2014) <sup>[1]</sup>   |

## Problème 2 : Fiabilité

Données produites **collaborativement** ou par des processus **faillibles**

- Vandalisme
- Controverses
- **Extraction**

Dans **YAGO**



**Macquarie Group Limited**

|  <b>MACQUARIE</b> |                                                            |
|----------------------------------------------------------------------------------------------------|------------------------------------------------------------|
| Type                                                                                               | Public                                                     |
| Founded                                                                                            | <a href="#">Sydney</a> , New South Wales, Australia (1970) |
| Revenue                                                                                            | <a href="#">AUD A\$8.1 billion (2014)</a> <sup>[1]</sup>   |

**extraction  
automatique**

<Macquarie>  $\xrightarrow{\text{<hasRevenue>}}$  "8100000000"^^<dollar>

## Problème 2 : Fiabilité

Données produites **collaborativement** ou par des processus **faillibles**

- Vandalisme
- Controverses
- Extraction

Dans **YAGO**



**Macquarie Group Limited**

|         |                                                                                                    |
|---------|----------------------------------------------------------------------------------------------------|
|         |  <b>MACQUARIE</b> |
| Type    | Public                                                                                             |
| Founded | <a href="#">Sydney</a> , New South Wales, Australia (1970)                                         |
| Revenue | <a href="#">AUD A\$8.1 billion (2014)</a> <sup>[1]</sup>                                           |

**extraction  
automatique**

<Macquarie>  $\xrightarrow{\text{<hasRevenue>}}$  "8100000000"^^<dollar>

→ **Approches existantes** : truth finding, data cleaning, data repair...

**Objectif :** Intégrer les données du Web et raisonner sur ces données en estimant leur fiabilité grâce à des annotations de provenance

Le raisonnement et la fiabilité vont de pair :

- Les données d'une source peuvent provenir d'autres sources
- Les règles d'intégration elles-mêmes ne sont pas fiables
- Il faut étudier la fiabilité des résultats du raisonnement

**Objectif :** Intégrer les données du Web et raisonner sur ces données en estimant leur fiabilité grâce à des annotations de provenance

Le raisonnement et la fiabilité vont de pair :

- Les données d'une source peuvent provenir d'autres sources
  - Les règles d'intégration elles-mêmes ne sont pas fiables
  - Il faut étudier la fiabilité des résultats du raisonnement
1. Prendre en compte la provenance initiale des faits
  2. Propager la provenance au cours du raisonnement

# Provenance existante des données

**Wikidata** : Plus de **40M** faits ont une source (près de 50%)

|                      |              |
|----------------------|--------------|
| identifiant BnF      | 118629315    |
| - 1 référence        |              |
| importé de           | data.bnf.fr  |
| date de consultation | 26 août 2015 |
| affirmé dans         | data.bnf.fr  |

Version du 30 août 2015 à 12:28 (restaurer)  
(annuler)  
ShonagonBot (discussion | contributions)  
(Ajout d'une référence à une affirmation : *identifiant BnF*  
(P268): 118629315)  
Modification suivante →  
propriété / **identifiant BnF** : **118629315** / référence

importé de : data.bnf.fr

date de consultation :

26 août 2015

Horodatage

+2015-08-26T00:00:00Z

Zone horaire

+00:00

Calendrier

Grégorien

affirmé dans : data.bnf.fr

# Provenance existante des données

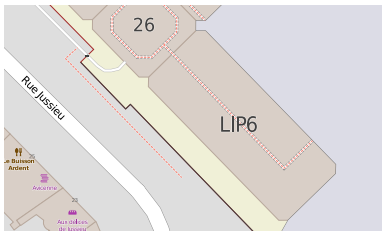
**Wikidata** : Plus de **40M** faits ont une source (près de 50%)

|                      |              |
|----------------------|--------------|
| identifiant BnF      | 118629315    |
| - 1 référence        |              |
| importé de           | data.bnf.fr  |
| date de consultation | 26 août 2015 |
| affirmé dans         | data.bnf.fr  |

Version du 30 août 2015 à 12:28 (restaurer)  
(annuler)  
**ShonagonBot** (discussion | contributions)  
(Ajout d'une référence à une affirmation : *identifiant BnF*  
(P268) : 118629315)  
Modification suivante →  
propriété / **identifiant BnF** : 118629315 / référence

|                            |                       |
|----------------------------|-----------------------|
| importé de : data.bnf.fr   |                       |
| date de consultation :     |                       |
| 26 août 2015               |                       |
| Horodatage                 | +2015-08-26T00:00:00Z |
| Zone horaire               | +00:00                |
| Calendrier                 | Grégorien             |
| affirmé dans : data.bnf.fr |                       |

**OpenStreetMaps** : **>40M** points et **>120M** voies avec source (>35%)



**Way: LIP6 (42741440)**

3d shapes in UPMC, footways

Edited about 1 year ago by **Goffredo**

Version #9 · Changeset #28100197

|        |                                                                                                |
|--------|------------------------------------------------------------------------------------------------|
| source | cadastre-dgi-fr source :<br>Direction Générale des<br>Impôts - Cadastre.<br>Mise à jour : 2009 |
|--------|------------------------------------------------------------------------------------------------|

# Provenance existante des données

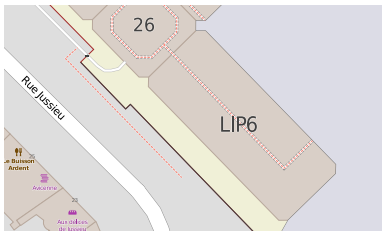
**Wikidata** : Plus de **40M** faits ont une source (près de 50%)

|                      |              |
|----------------------|--------------|
| identifiant BnF      | 118629315    |
| - 1 référence        |              |
| importé de           | data.bnf.fr  |
| date de consultation | 26 août 2015 |
| affirmé dans         | data.bnf.fr  |

Version du 30 août 2015 à 12:28 (restaurer)  
(annuler)  
**ShonagonBot** (discussion | contributions)  
(Ajout d'une référence à une affirmation : *identifiant BnF*  
(P268) : 118629315)  
Modification suivante →  
propriété / **identifiant BnF** : **118629315** / référence

|                            |                       |
|----------------------------|-----------------------|
| importé de : data.bnf.fr   |                       |
| date de consultation :     |                       |
| 26 août 2015               |                       |
| Horodatage                 | +2015-08-26T00:00:00Z |
| Zone horaire               | +00:00                |
| Calendrier                 | Grégorien             |
| affirmé dans : data.bnf.fr |                       |

**OpenStreetMaps** : **>40M** points et **>120M** voies avec source (>35%)



**Way: LIP6 (42741440)**

3d shapes in UPMC, footways

Edited about 1 year ago by **Goffredo**

Version #9 · Changeset #28100197

|        |                                                                                                |
|--------|------------------------------------------------------------------------------------------------|
| source | cadastre-dgi-fr source :<br>Direction Générale des<br>Impôts - Cadastre.<br>Mise à jour : 2009 |
|--------|------------------------------------------------------------------------------------------------|

**Autres indices** : **Historique** des faits, **utilisateurs** qui ont édité

# Propager la provenance au cours du raisonnement

→ Comment définir la **provenance** d'une réponse **certaine**?

## Données :

| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |

## Requête logique :

*Y a-t-il des réunions  
pendant mes congés?*

## Règle logique :

Je ne reviens pas  
pour **un seul jour**

## Provenance :

$t_1 \wedge t_2 \wedge t_3 \wedge \text{règle?}$

# Propager la provenance au cours du raisonnement

→ Comment définir la **provenance** d'une réponse **certaine**?

**Données :**

| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |

**Requête logique :**

*Y a-t-il des réunions  
pendant mes congés?*

**Règle logique :**

Je ne reviens pas  
pour **un seul jour**

**Provenance :**

$t_1 \wedge t_2 \wedge t_3 \wedge \text{règle?}$

→ Généraliser les **semianneaux de provenance** [Green et al., 2007]  
des bases de données au raisonnement?

# Propager la provenance au cours du raisonnement

→ Comment définir la **provenance** d'une réponse **certaine** ?

**Données :**

| id    | jour | type    |
|-------|------|---------|
| $t_1$ | 9    | congés  |
| $t_2$ | 10   | réunion |
| $t_3$ | 11   | congés  |

**Requête logique :**

*Y a-t-il des réunions  
pendant mes congés ?*

**Règle logique :**

Je ne reviens pas  
pour **un seul jour**

**Provenance :**

$t_1 \wedge t_2 \wedge t_3 \wedge \text{r\grave{e}gle} ?$

→ Généraliser les **semianneaux de provenance** [Green et al., 2007]  
des bases de données au raisonnement ?

→ Comment calculer **efficacement** cette provenance  
et la représenter de façon **concise**, selon le langage de règles ?

## ***Raisonner avec la provenance sur les données du Web*** *pour l'intégration et la fiabilité*

1. Provenance symbolique pour le raisonnement
  - **Définition** abstraite à différents niveaux d'expressivité
  - **Calcul** et **représentation** efficace
2. Propager des relations de **fiabilité** à travers la provenance
3. Calculer des confiances **quantitatives** et probabilistes
4. **Réviser** les jugements sur les sources primaires  
avec des retours utilisateurs, en remontant la provenance

# Application : Compléter et vérifier Wikidata

Ajouter à Wikidata des faits issus d'autres sources et les vérifier :

**Valeria Pereyra** [Q16227045]  
Valeria Pereyra is a Argentine artistic gymnast.  
She was born on February 12, 1996.

This challenge is about the place of birth of this Wikidata entity.

Buenos Aires, Argentina

Skip

Reject

Règles logiques **déclaratives** : extraction, intégration, conflits

→ Calculer la **provenance** et estimer la **fiabilité** avec :

- **Sources originales** (Wikipédia, etc.)
- **Extracteurs** (comme pour Yago) et **intégration**
- **Contradictions** entre faits
- Jugements de la **foule** (pas toujours fiables)

# Intégration

---

Équipe DAHU : Vérification en bases de données

**Gestion de connaissances distribuées :** Langage **WebdamLog** :

- **Contrôle d'accès** fondé sur la provenance
- Gestion de **contradictions** et **explication** de réponses

**Données du Web :** Fondements des bases de données **graphe** :

- Étude de problème de la **détermination de vues**

⇒ Ces problèmes peuvent être **capturés** avec la provenance

⇒ Ils peuvent être **étendus**, notamment pour gérer la **confiance**

**Projet :** *Raisonnement avec la provenance sur les données du Web*  
pour l'intégration et la fiabilité

1. Provenance symbolique pour le raisonnement
  - **Définition** abstraite à différents niveaux d'expressivité
  - **Calcul** et **représentation** efficace
2. Propager des relations de **fiabilité** à travers la provenance
3. Calculer des confiances **quantitatives** et probabilistes
4. **Réviser** les jugements sur les sources primaires  
avec des retours utilisateurs, en remontant la provenance

**Thèse :** *Tirer parti de la structure des données incertaines*  
[ICALP'15], [LICS'15], [PODS'16]; après-thèse [IJCAI'16]

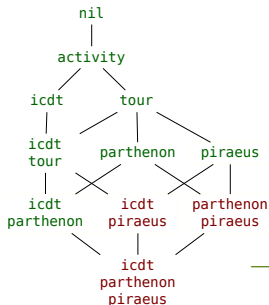
**Pré-doc :** À Tel Aviv [ICDT'14] et à Oxford [IJCAI'15]

# Crowdsourcing

**Fouille de données :** Trouver des motifs fréquents

**Crowdsourcing :** Poser des questions à la foule

→ Questions à la foule : quels **ensembles d'objets** sont fréquents ?

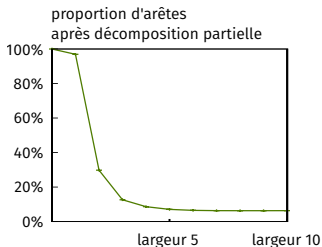


- Utiliser une **taxonomie** sur les objets
- Les ensembles forment un **treillis distributif**
- **Compromis** entre le **coût** des questions posées et le **coût** de calculer quelles questions poser

→ Bornes de **complexité** sur ce problème

# Applicabilité pratique de la largeur d'arbre

- Travail avec S. Maniu : les jeux de données réels peuvent être **partiellement décomposés** en arbre



Décomposition **partielle** en arbre  
du **graphe OSM de Paris**

- **4.3 M** nœuds et **5.4 M** arêtes
  - Largeur totale  $\leq$  **521**
- Stage de M. Monet : les méthodes à base d'automates peuvent être **implantées en pratique**
  - Travail avec M. Monet : compilation efficace en automates pour des **langages restreints** de requêtes



COURCELLE, Bruno (1990). “The Monadic Second-Order Logic of Graphs. I. Recognizable Sets of Finite Graphs”. In : *Inf. Comput.*

GREEN, Todd J., Grigoris KARVOUNARAKIS, Val TANNEN (2007).  
“Provenance Semirings”. In : *Proc. PODS*.