# Intuitions métier ➔ règles formelles

Un contributeur a tendance à republier chez le même éditeur

Un contributeur avec un rôle non principal est moins susceptible de republier chez le même éditeur

Un contributeur a tendance à republier avec les mêmes co-contributeurs

```
SI X a publié avec Y, et
SI X a publié un autre
livre avec Z qui a le
même nom que Y,
ALORS Z == Y
```

Qualinca

# Intégrer Qualinca dans les opérations ABES

- **Outils** :
  - Liage automatique NB ➔ NA
  - Aide à la décision intégrée à l'application professionnelle www.idref.fr (diapo suivante #1)
    - En temps réel
    - Après pré-traitement
  - Outil de vérification et de correction des liens internes au Sudoc (diapo suivante #2)
  - Web service
- **Corpus** :
  - Sudoc et autres bases ABES (theses.fr, Manuscrits)
  - Traitement de données d'éditeur licencesnationales.fr
  - Corpus extérieurs

Qualinca

IdRef
Le référentiel des autorités Sudoc

Identifiant * Mot de passe * Se souvenir

Rechercher | Résultats | Notice | Aide

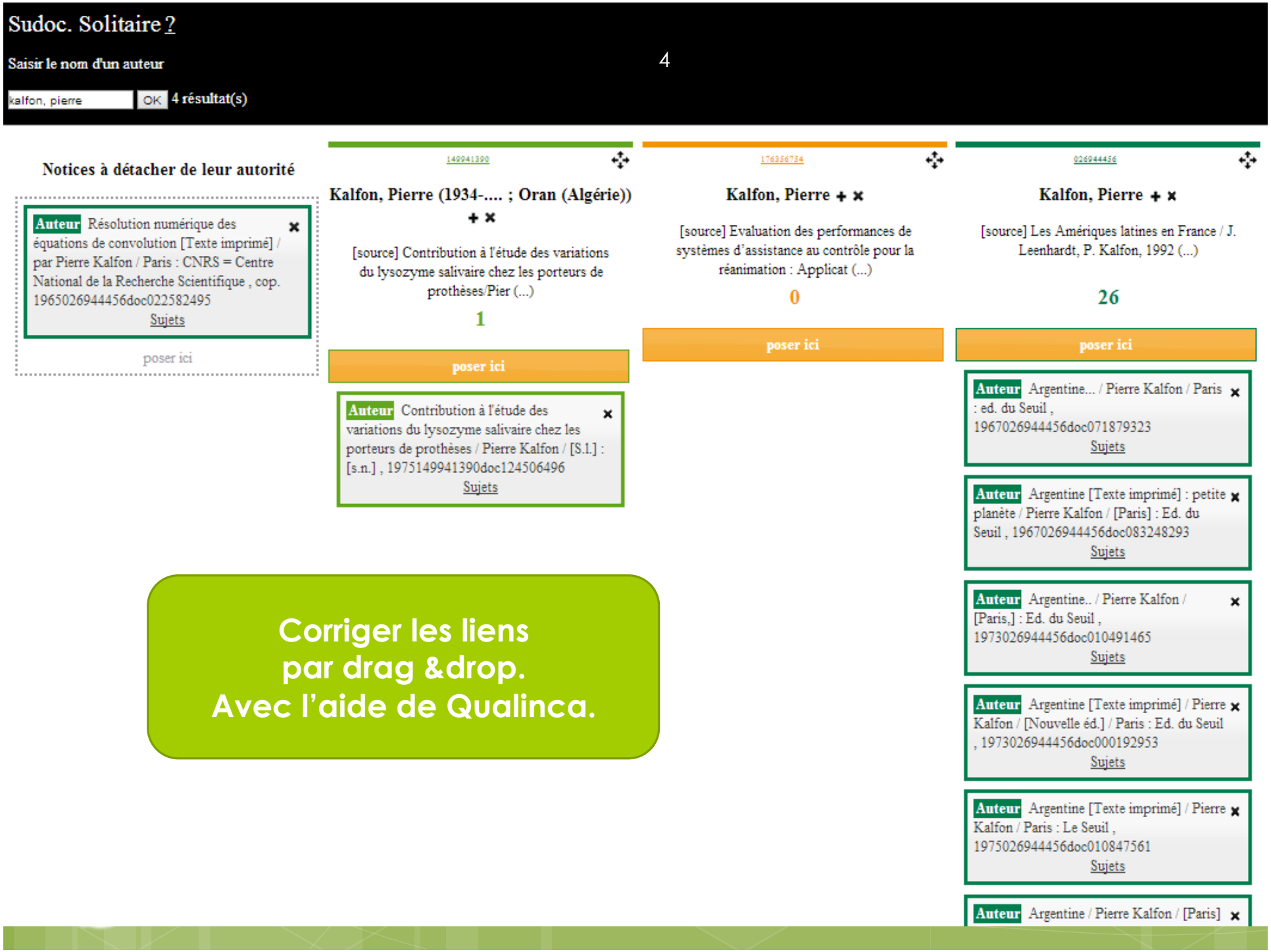
# Résultats

**53** résultats pour le terme " **Nom de personne=Bernard**

**Trop d'homonymes ! Qualinca fera le tri pour le catalogueur.**

| | | | | |
|---|---|---|---|---|
| 1. | 176652280 | BERNARD, ALAIN | | |
| 2. | 028051092 | Bernard, Alain (19..-.... ; traducteur) | | |
| 3. | 028051165 | Bernard, Alain (19..-.... ; ingénieur) | Bibliographie | Notice abrégée |
| 4. | 02805119X | Bernard, Alain (1941-....) | Bibliographie | Notice abrégée |
| 5. | 031926452 | Bernard, Alain (19..-.... ; biologiste) | Bibliographie | Notice abrégée |
| 6. | 034871640 | Bernard, Alain (1947-....) | Bibliographie | Notice abrégée |
| 7. | 035635290 | Bernard, Alain (1955-....) | Bibliographie | Notice abrégée |
| 8. | 057656673 | Bernard, Alain (19..-.... ; historien) | Bibliographie | Notice abrégée |
| 9. | 075050773 | Bernard, Alain (1953-....) | Bibliographie | Notice abrégée |

ls
ot en :
MarcXml
Trier les résultats par :
Pertinence
Filtrer les résultats par :




**Corriger les liens par drag &drop. Avec l'aide de Qualinca.**

# Echantillon de référence ABES

1. Sélectionner un lot de départ (NB+NA)
2. Corriger ce lot
   a) Exactitude des liens
   b) Complétude des liens

| Nb Autorités | Dont créées | Nb notices biblio |
|---|---|---|
| 133 | 61 | 578 |

| Nb liens OK |
|---|
| 342 |
| Nb Liens erronés corrigés 126 |
| NB Liens erronés supprimés (sans remplacement) 5 |
| Nb Liens manquants ajoutés 107 |

Qualinca

# Découverte de clés

# Découverte de clés pour l'identification de références d'entités

- Clé ou règles d'identification : combinaison de propriétés (ou inverses) qui identifie une entité.

  Ex. : même titre, même auteur ➔ même livre

- Utiles, en général, pour :
  - Les approches logiques et numériques d'identification de références
  - La présélection des références à comparer.
- Exploitables, dans Qualinca, pour : Inférer des liens de coréférence entre références contextuelles et entre références d'autorité.
- Problèmes : clés (composites) difficiles à spécifier pour un expert
- Solution : découverte automatique de clés à partir des faits
- Difficile : bases documentaires incomplètes et très volumineuses

Qualinca

# KD2R et SAKey [L3.1]: Deux algorithmes de découverte de clés OWL2

- **Hypothèses sur les données :**
  1. Sources de données conformes à des ontologies différentes
  2. Hypothèse du nom unique (UNA) dans chaque source
  3. Hypothèse du monde ouvert (OWA)

- **KD2R –Key Discovery for Reference Reconciliation**
  1. Trouver toutes les non clés maximales (inspiré de [Y. Sismanis et al. 2006])
  2. Dériver les clés minimales suivant la sémantique des clés de OWL2

- **Limites de KD2R :**
  - Passage à l'échelle de millions de références
  - Robustesse aux redondances dans les données

Qualinca

# KD2R et SAKey [L3.1]: Deux algorithmes de découverte de clés OWL2

- **Clé avec n exceptions (n-almost Key)** : une clé telle qu'il existe au **maximum n** références qui **violent cette clé**.

- **SAKey – Scalable Almost Key Discovery**
  1. Trouver toutes les ''n-non keys'' maximales
  2. Dériver les ''n-almost keys'' minimales
  3. Exploiter des dépendances sémantiques pour précompiler les données et/ou élaguer l'espace de recherche.

- **Collaboration entre le LIRMM, le LRI et le LIG :** comparaison théorique et expérimentale de deux sémantiques de clés :
  - la sémantique OWL2 et
  - la sémantique considérée dans Atencia et al. 2012

Qualinca

# KD2R et SAKey : Quelques résultats de l'évaluation

| Bases | Type de références | #références | #propriétés | Algo | #clés | Temps |
|---|---|---|---|---|---|---|
| INA | Notices documentaires | 44779 | 11/82 | KD2R-O | 3 | 18h:22 mins |
| | Notices d'autorité (Personne) | 7444 | 7/44 | | | |
| ABES | Références contextuelles (Personne) | 5671 | 9 | KD2R-O | 12 | 6mins:21s |
| | | | | SAKey | 12 | 1min:3s |

- Résultats INA (KD2R) : sélection de propriétés fréquentes pour découvrir des clés ➔ clés non pertinentes.
- Résultats ABES (KD2R et SAKEY) :
  - Une clé qui semble valide [[apourcoauteur, apourappellation]]
  - 11 autres qui seraient pertinentes, mais avec des exceptions

    ➔ nécessité **d'évaluation de la qualité** des liens produits en utilisant ces clés.
- Autres évaluations (LOD et OAEI) : 6 jeux de données

Exemple (temps sur dbpedia:Lake) : KD2R en **208 minutes** et SAKey en **5 secondes**

Qualinca

# Réconciliation de références par enrichissement à partir du LOD

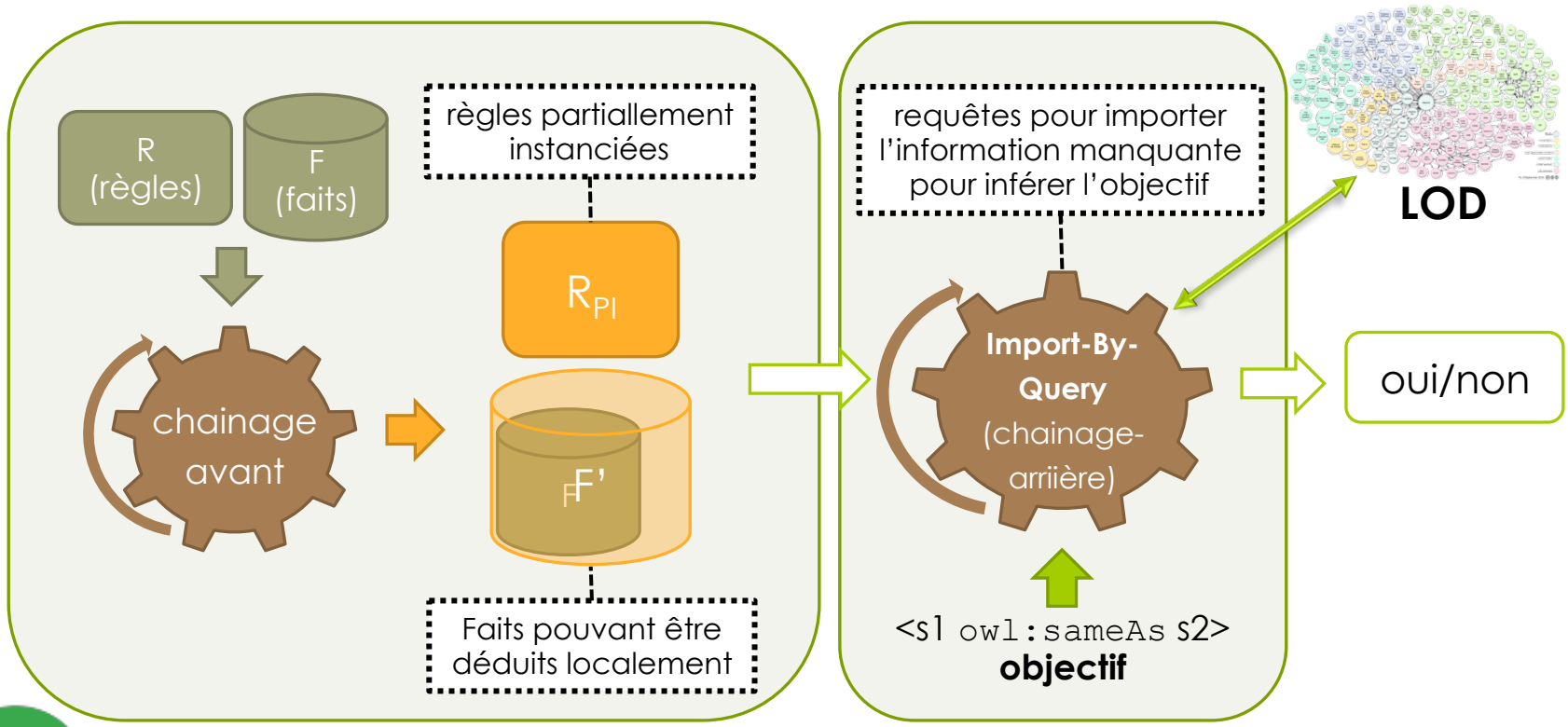
Qualinca

# Réconciliation de références par enrichissement à partir du LOD

- **objectif** : réconciliation de références RDF
  - dans un jeu de données RDF ou entre plusieurs jeux de données.
- **approche** : fondée sur des règles logiques pour inférer des faits `owl:sameAs`/ `owl:differentFrom` en générant des requêtes sur le LOD
  - les règles proviennent : des contraintes du schéma, des connaissances du domaine ou de la sémantique des constructeurs RDFS/OWL (ex. transitivité du `owl:sameAs`)
- **méthode : algorithme « import-by-query »**
  - **input** : requête booléenne de type **<s1 `owl:sameAs` s2>** (ou **<s1 `owl:differentFrom` s2>**) et une base de règles.
  - construction itérative de requêtes SPARQL pour importer du LOD les données externes pertinentes à la requête, **uniquement**.
  - notre algorithme est une adaptation de l'algorithme d'évaluation de programmes Datalog « Query-Subquery » combinant **chainage-avant** et **chainage-arrière.**

Qualinca

# Réconciliation de références par enrichissement à partir du LOD




Qualinca

# Réconciliation de références par enrichissement à partir du LOD

- **Evaluation** : expérimentations sur un corpus de données RDF fourni par l'INA et enrichi par des règles traduisant des connaissances des experts (par ex. clefs) et la sémantique des constructeurs RDFS/OWL
  - passage à l'échelle de base de connaissances constituées de plusieurs **millions de faits RDF** et de plus de **30 règles** (dont certaines sont récursives)
  - enrichissement des données locales par import de données provenant de DBpedia
  - réconciliation de **533 paires de personnes** homonymes (47 paires par raisonnement local) : en particulier des références d'autorité et, à une moindre échelle, des références contextuelles (liens d'autorités)
  - le nombre de faits importés est diminué de manière considérable par rapport à celui de l'approche standard (**7160 faits vs 500,000 faits**)
- ce travail a été soumis à la conférence ECAI 2014

Qualinca

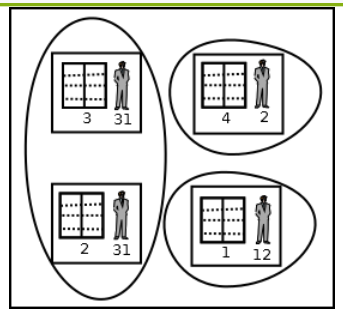
# Algorithme « import-by-query » adapté à l'incertitude

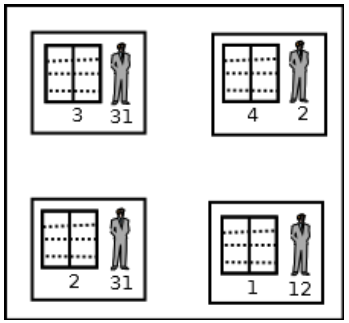
- nécessité de tenir compte de l'**incertitude**
  - *des faits incertains* dérivés de, par ex. valeurs de similarité entre littéraux, ou jugements de confiance sur une multitude de sources de données
  - *règles incertaines* provenant de, par ex. pseudo-clefs, ou alignements d'ontologies
- adaptation de l'algorithme « import-by-query »
  - une vue unifiée de l'incertitude en modélisant les valeurs d'incertitude comme valeurs de **probabilité** (probabilité qu'un fait/règle soit vrai)
  - fondée sur « probabilistic Datalog »
  - cadre général : « bootstrap » d'enrichissement de données
  - objectif : réconcilier plus de références contextuelles dans le corpus de l'INA.

Qualinca

# Vérification des liens par partitionnement

partition induite par les autorités

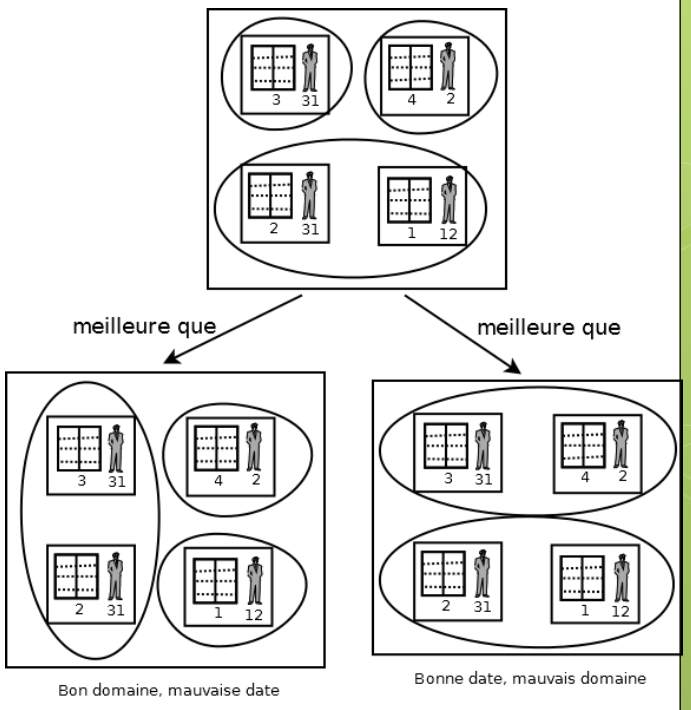



Appartient aux meilleures ?

Sémantique de partitionnement

Critères de similarité




Qualinca

# Evaluation

- Performance de l'approche
  - polynomiale en nombre de références et exponentielle en nombre de critères
    - ~6 sec. pour 650 références contextuelles et 6 critères
  - nécessité de travailler par sous-ensemble de références de la base
    - « blocking » par nom des références
- Pertinence de l'approche
  - Expérimentation sur le jeu de donnés ABES
    - 536 références contextuelles réparties en 7 sous-ensembles
    - 7 critères utilisés : appelation, titre, autresContributeurs, these, directeurDeThese, datePublication, rôle
  - Résultat :
    - Partition « jeu corrigé » meilleure que partition « jeu initial »
    - Partition « jeu corrigé » appartient aux « meilleures partitions » dans 5 cas sur 7

Qualinca

# Organisation, résultats et perspectives

# Fonctionnement

- Réunions plénières du projet : 5
  - cadre général de travail et points d'étape
  - Prochaine : Vendredi 11 avril
- Réunions thématiques par petits groupes : 14
  - Résolution par requêtes externes : LIG, INA
  - Construction de jeu de données : ABES, INA, LIRMM, LRI
  - Complétion et contrôle des liens : LIRMM, ABES
  - Fusion : LIRMM, LRI
  - Découvertes clés : LRI, LIG, LIRMM
- Site Web pour affichage et échanges internes : www.lirmm.fr/qualinca
- Avancement en conformité avec les prévisions :
  - Tâche 2 : Modèle de qualité (2 livrables)
  - Tâche 3 : Réconciliation de données (2 livrables)
  - Tâche 4 : enrichissement (3 livrables)
  - Tâche 6 : évaluation et démonstrateurs (2 livrables)

Qualinca

# Résultats obtenus

- Publications
  - 2 revues internationales : *JWS, KIS*
  - 5 conférences internationales (dont 2 multipartenaires) : *IPMU, Fuzzy IEEE, KSE, SGAI, WOD*
    - 2 soumissions en cours (dont 1 multipartenaire)
  - 2 conférences nationales (dont 1 multipartenaire) : *IC*
    - 1 soumission en cours

- Diffusion, valorisation
  - Organisation d'ateliers sur le thème des données liées : SOS-SLWD 2013, SoWeDo 2014
  - Tutoriel sur le web de données : EGC 2014
  - Participation au groupe de travail international : FRBROO-CRM SIG
  - Présentation du projet à la session industrielle d'ICCS 2014
  - Plusieurs séminaires dans différents cadres : invité, labex, laboratoire

- 5 Prototypes :
  - Liage par règles à critères (LIRMM /ABES)
  - Liage par enrichissement à partir de base externe (LIG)
  - Partitionnement (LIRMM)
  - Découverte de Clés (LRI)
  - Création d'autorités (INA)

# Perspectives

① Evaluation des méthodes sur les problèmes de qualité
- Spécifier des scénarios d'utilisation des méthodes : *évaluation de la qualité d'un catalogue, réparation d'un catalogue, fusion de catalogues, insertion d'une notice documentaire dans un catalogue…*
- Eliciter des règles/critères/attributs pour les données ABES et INA en exploitant les algorithmes de découverte de clés
- Evaluer dans un contexte métier

② Prise en compte d'incohérences dans le liage
- Utiliser des techniques de raisonnement en présence d'incohérences

③ Prise en compte d'une notion de confiance dans les données ou dans les règles
- Utiliser des principes d'annotation et des modèles de préférence

④ Extension des méthodes de fusion de données
- Construction d'autorités enrichies et dédoublonnage d'autorités

Qualinca

# Merci !

Qualinca

# Annexes INA

# Exemple d'extraction « métier »

**Local view for "2843095001"**

| | |
|---|---|
| ina:a pour thématique | ina:Information[1] |
| ina:aPourDateDiffusion | "2005-05-26"^^xsd: |
| ina:aPourLieu | Paris[1] |
| ina:aPourParticipant | Delanoë, Bertrand[2] |
| | Delanoë, Bertrand[1] |
| ina:aPourPresentateur | Laborde, Françoise[3] |
| | Laborde, Françoise[1] |
| ina:aPourRealisateur | Oudin, Jean Philippe |
| | Oudin, Jean Philippe[1] |
| ina:aPourResume | "Françoise LABORDE reçoit Bertrand DELANOE, maire de Paris, pour parler "Oui" à la Constitution, il parle de sa défense du "Oui", de la démocratie, des pro rassemblement sportif prévu le 5 juin sur les Champs Elysées pour soutenir la car |
| | constitution[1] |
| | référendum[1] |

**La notice EMISSION**

**La description contextuelle: « personne »**

**Local view for "PERSON_2843095001_3"** URI de référence contextuelle

| Predicate | Value (sorted: def |
|---|---|
| skos:preferred label | "Delanoë, Bertrand" |
| rdf:type | core:Personne physique |
| core:hasOccupation | maire de Paris (core:confidence->"1.0"^^xsd:float) (core:source->", maire de Paris, pour parler du "Oui "a |

**« occupation »**

**Local view for "PERSON_2843095001_3_occup_1"**

| Predicate | Value (sorted: default) |
|---|---|
| rdfs:label | "maire de Paris" |
| rdf:type | core:Fonction, qualités et titres |
| core:hasType | dbpedia:Mayors of Paris |
| | Maire de Paris |

Qualinca

# Exemple sur une œuvre (1)

**Le champ textuel « Œuvres » de la notice EMISSION**

| | |
|---|---|
| ina:aPourLieu | salle de spectacle[1] (core:precision->"Bobino") |
| ina:aPourOeuvres | "Henri TACHAN chante "Dans les wagons de première classe", "Qui trop embrassse mal étreint", "Un mur", "Bosco" et "La table habituelle", accompagné par Jean LESAGE au piano. Roland ROMANELLI joue des airs d'accordéon. Colette RENARD chante "Chagrin d'amour", "Marine, là-bas", "Autopsie d'un amour", "Le rôdeur de Paris", "Tu me plais et je t'aime", "Bateau de femmes", "Je t'aime en français", "Le marin et la rose" et "Irma la douce" accompagnée par Jacques LALU au piano, Fernand GARBASI à la guitare, Gibert ROUSSEL à l'accordéon et Pierre NICOLAS à la contrebasse."[1] |
| ina:aPourProducteur | Rezvilsoff, Dominique[1] |

**Extraction autour de l'Œuvre: Un Mur**

Qualinca

# Exemple sur une œuvre (2)

**La notice EMISSION**

champ « évènement »

**Local view for "CPF86654140"**

| | |
|---|---|
| core:event | Dans les wagons de première classe[1]<br>Qui trop embrassse mal étreint[1]<br>Un mur[1]<br>Bosco[1]<br>la table habituelle[1]<br>Chagrin d ' amour[1]<br>Marine . là - bas[1]<br>Autopsie d ' un amour[1]<br>le rôdeur de Paris[1]<br>Tu me plais et je t ' aime[1]<br>Bateau de femmes[1]<br>Je t ' aime en français[1]<br>le marin et la rose[1]<br>Irma la douce[1] |
| core:lastUpdateDate | "2010-10-06"^^xsd:date[2] |

champ « interprétation »

**Local view for "interpretation_CPF86654140_204065"**

| Predicate | Value (sorted: default) |
|---|---|
| rdfs:label | "Un mur"@fr[1] |
| rdf:type | core:Performance[1] |
| core:chanteur | <Henri TACHAN>[1] |
| core:piano | <Jean LESAGE>[1] |
| core:work | Un mur[2] |

**La référence contextuelle à l'œuvre**

**Local view for "work_204065"**

| Predicate | Value (sorted: default) |
|---|---|
| skos:preferred label | "Un mur"@fr |
| rdf:type | core:Work |

**Sa description contextuelle**

Qualinca

# Autorité homonyme Jean Lesage

**Local view for "http://www.ina.fr/thesaurus/pp/concept_10073132"**

| Predicate | Value (sorted: default) |
|---|---|
| skos:alternative label | "Jean Lesage" (core:confidence->"0.75"^^xsd:float) (core:comesFrom->skos:prefLabel) |
| skos:preferred label | "Lesage, Jean" |
| rdf:type | ina:Homme |
| ina:aPourNoteQualite | "Homonymes : 1 - Homme politique, avocat québecois. Canada ; 2 - Auteur musical. Canada" |
| ina:aPourStatut | "Z" |
| skos:hidden label | "LESAGE JEAN" |

**L'auteur musical**

**Local view for "http://www.ina.fr/thesaurus/pp/concept_10073132_2"**

| Predicate | Value (sorted: default) |
|---|---|
| skos:alternative label | "Jean Lesage" |
| skos:preferred label | "Lesage, Jean" |
| rdf:type | ina:Personne Physique INA |
| ina:aPourHomonyme | Lesage, Jean |
| ina:aPourNationalite | Canada (core:confidence->"1.0"^^xsd:float) (core:comesFrom->ina:aPourNoteQualite) |
| ina:aPourNationaliteString | "Canada" |
| core:hasOccupation | Auteur musical (core:confidence->"0.46153846153846156"^^xsd:float) (core:comesFrom->ina:aPourNoteQualite) |

Qualinca