

ALIGNER

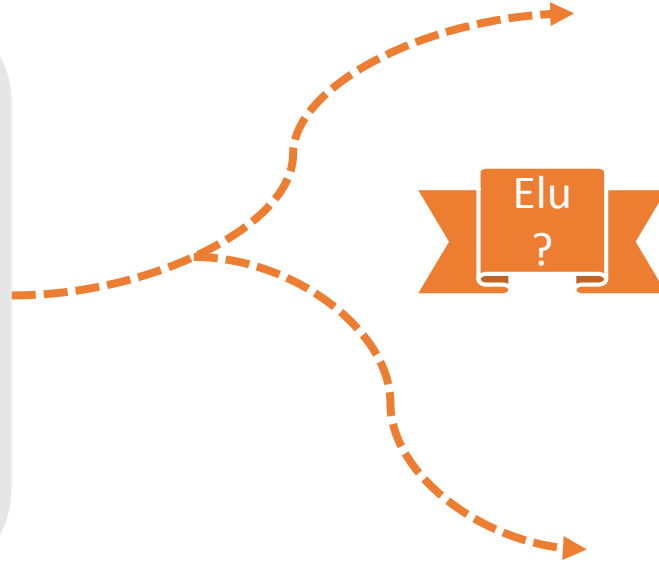
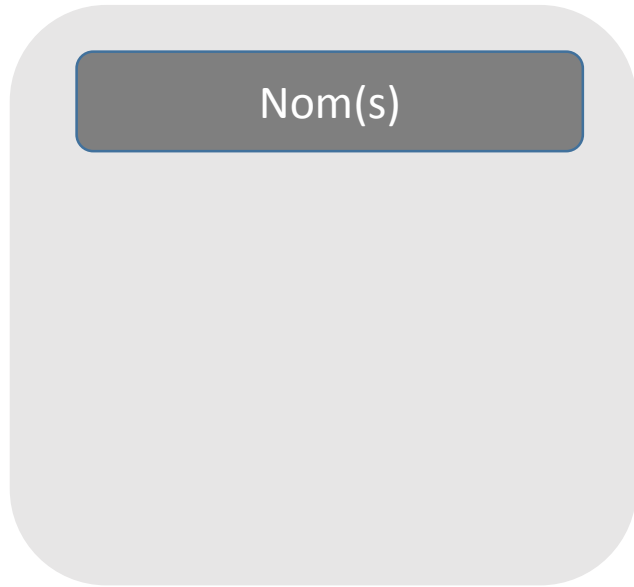
Le signalement augmenté

Aligner, en attendant Qualinca...

- En attendant la fin (2016) et la concrétisation (2017-2018?) du projet Qualinca...
- Solution
 - Maison
 - Provisoire
 - Partielle
 - **Fiable**

Aligner

Données d'entrée à lier



Référentiel de personnes

-

Candidats cibles en lice



Notice d'autorité

Forme retenue

Formes rejetées

Dates

Notices bibliographiques (1, n)

Titre

Contributeur(s)

Sujet(s)

Date

+ autres attributs

Titre

Contributeur(s)

Sujet(s)

Date

+ autres attributs

« Superautorité »

- . Persée
- . IdRef/Sudoc
- . IdHal
- . Etc.

Nom(s)

Titre(s)

Co-contributeur(s)

+ autres attributs

Sujets

Dates
biblio

Dates
de vie

Informations de la notice d'autorité enrichies
par les notices bibliographiques liées

Notice d'autorité

Forme retenue

Formes rejetées

Dates

Notices bibliographiques (1, n)

Titre

Contributeur(s)

Sujet(s)

Date

+ autres attributs

Titre

Contributeur(s)

Sujet(s)

Date

+ autres attributs

« Superautorité »

Cluster de notices biblio (cluster)

Nom(s)

Titre(s)

Co-contributeur(s)

+ autres attributs

Sujets

Dates biblio

Dates de vie

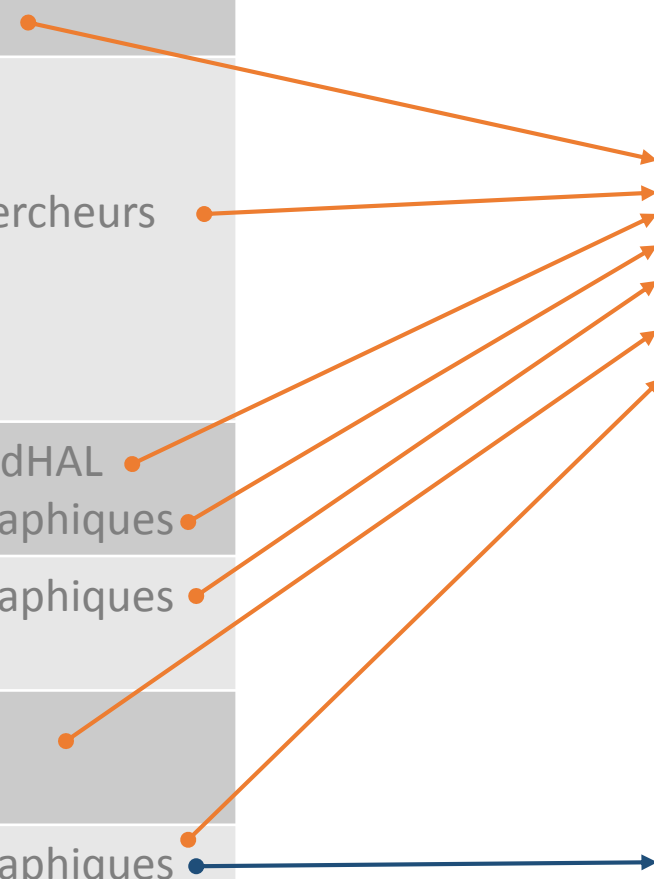
Agrégation de plusieurs notices bibliographiques

Données à lier

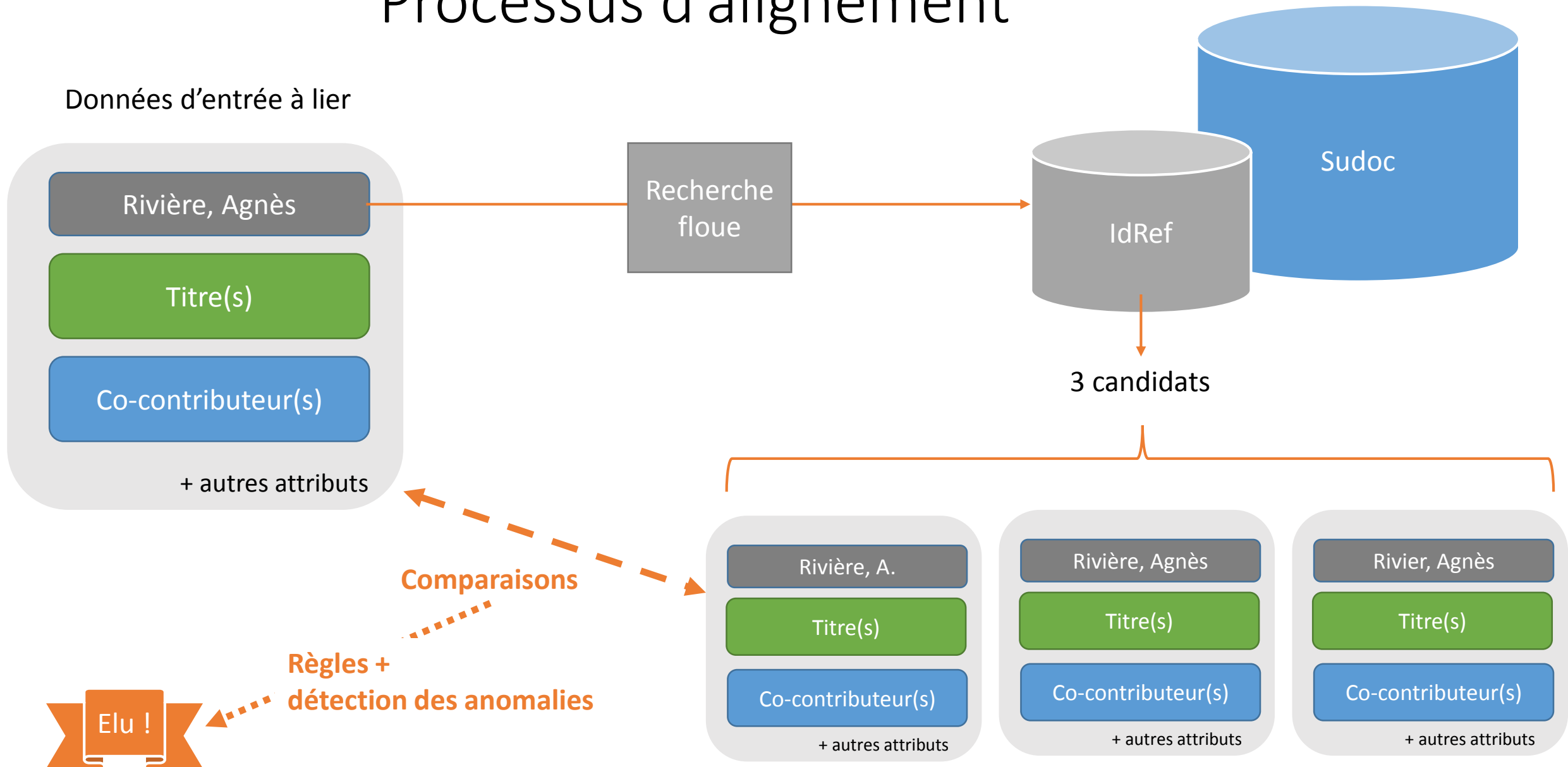
Persée	. <i>Superautorités</i>
. ADUM . Toulouse2 . SAMPRA . OKINA . Lorraine Etc.	. Annuaire de chercheurs (nom + équipe)
HAL	. <i>Superautorités</i> IdHAL . Notices bibliographiques
CAIRN ebooks	. Notices bibliographiques Sudoc
ORCID	. <i>Superautorités</i>
Wiley (ISTEX)	. Notices bibliographiques (articles)

Cibles

IdRef/Sudoc	Superautorités : <i>IdRef + Notices Sudoc</i> (pour l'instant)
ISNI	Superautorités
ORCID	Superautorités
Viaf	Superautorités
Wiley	On aligne les données vers elles-mêmes = clusterisation

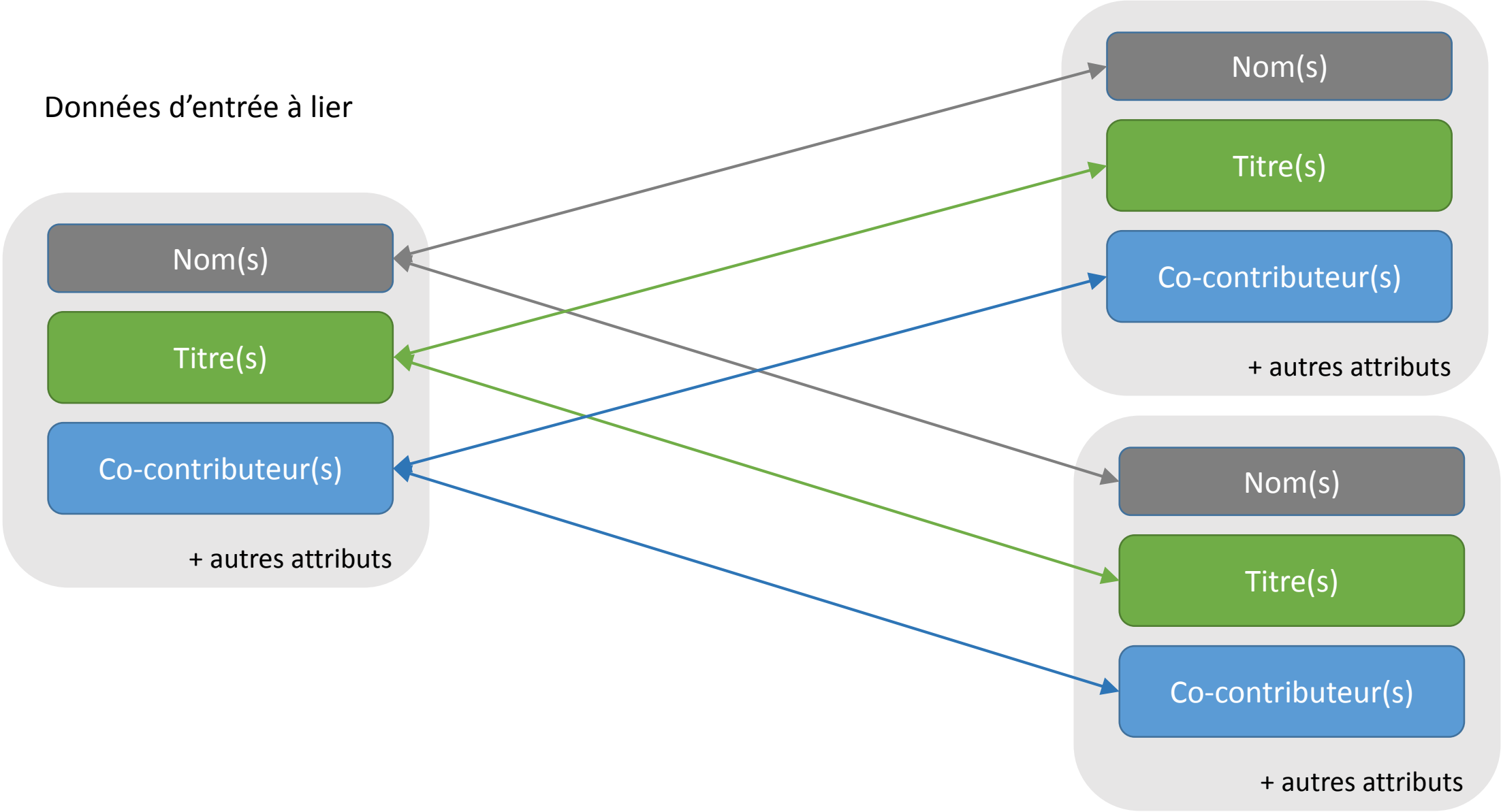


Processus d'alignement

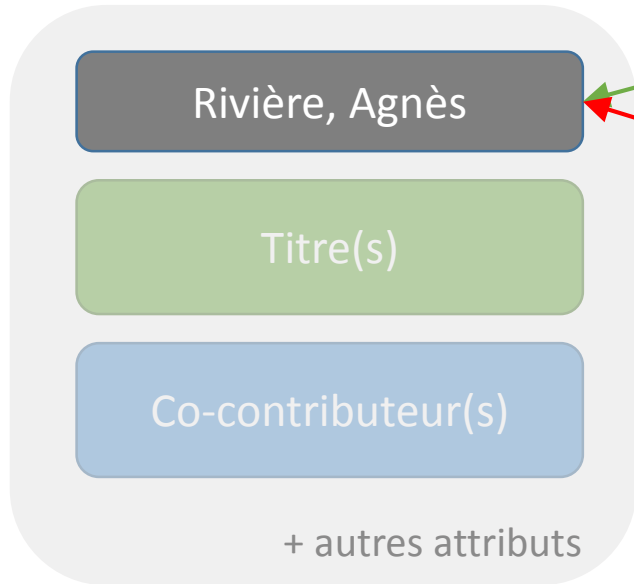


Candidats cibles en lice

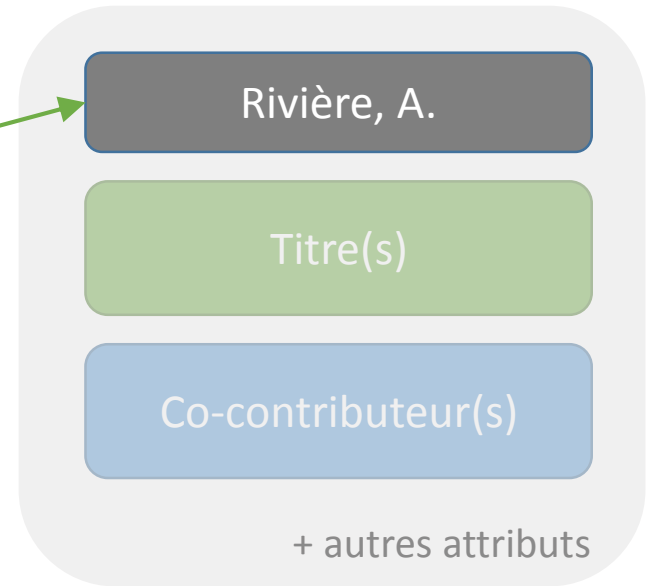
Données d'entrée à lier



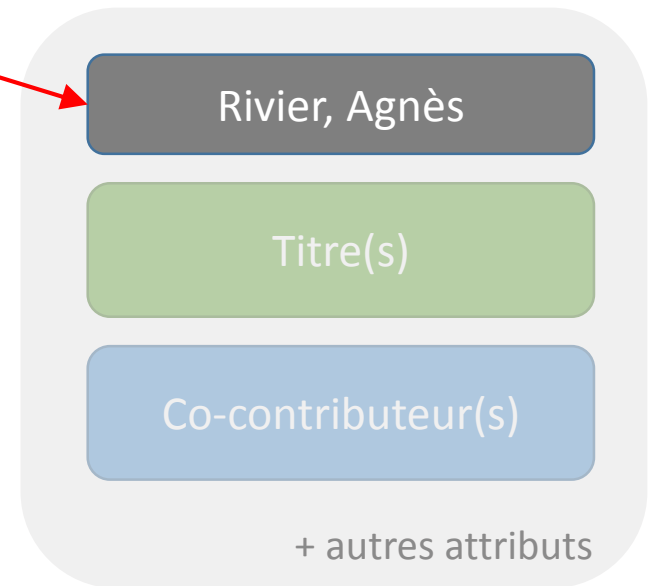
Données d'entrée à lier



100



-50



Candidats cibles en lice

Données d'entrée à lier

Nom(s)

Histoire des *soyeux*

Co-contributeur(s)

+ autres attributs

Nom(s)

Œil de lynx

Co-contributeur(s)

+ autres attributs

Nom(s)

Les *soyeux*, une vocation

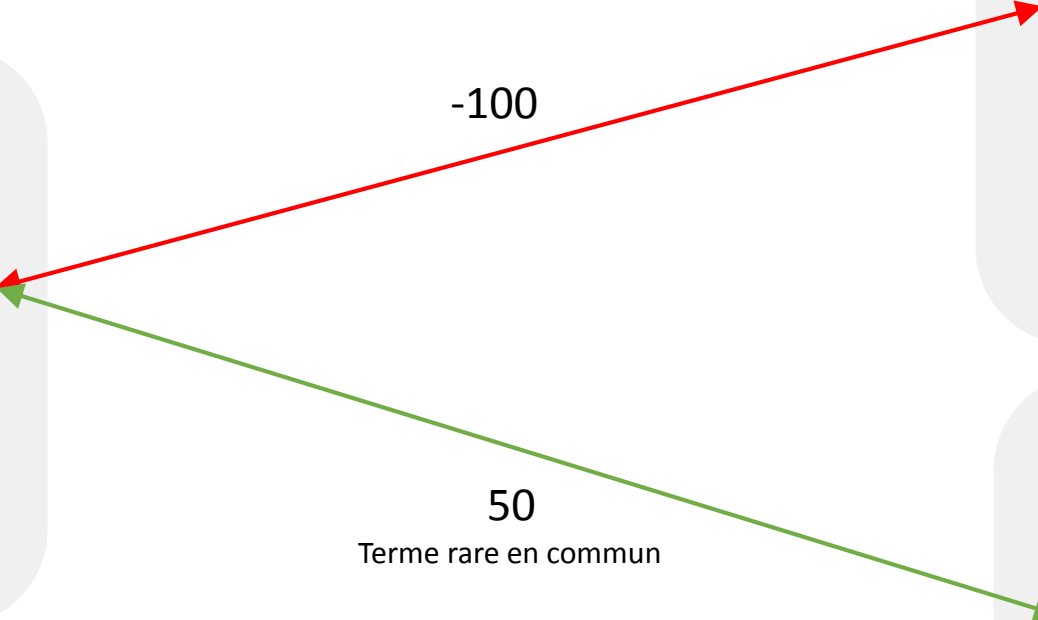
Co-contributeur(s)

+ autres attributs

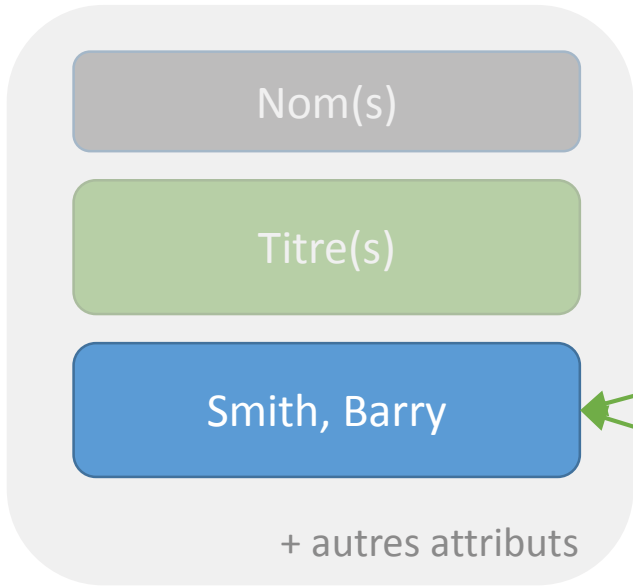
-100

50
Terme rare en commun

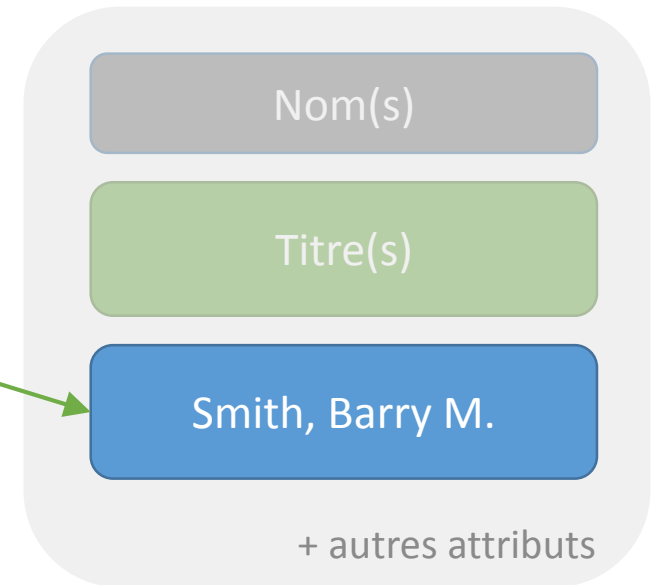
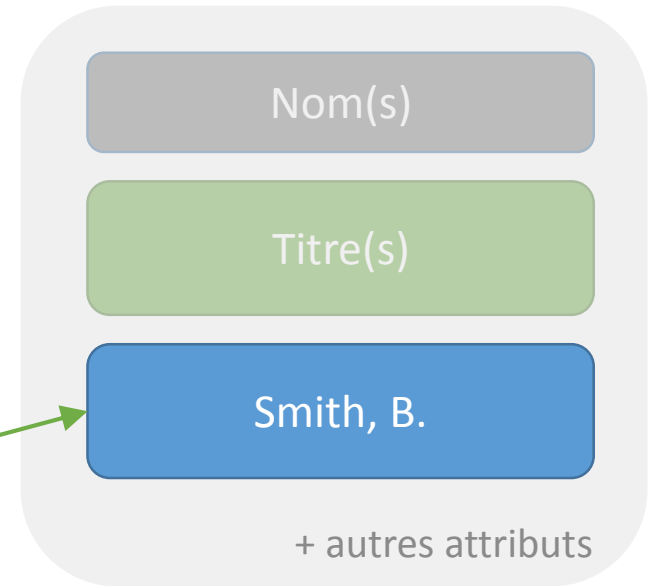
100
Titres identiques



Données d'entrée à lier



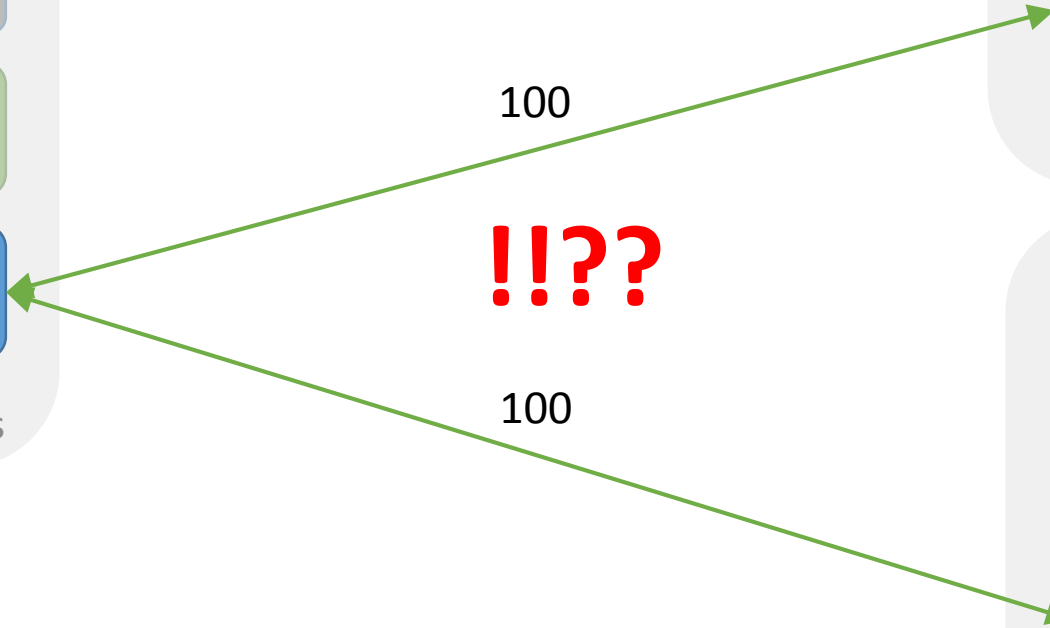
Candidats cibles en lice



100

!!??

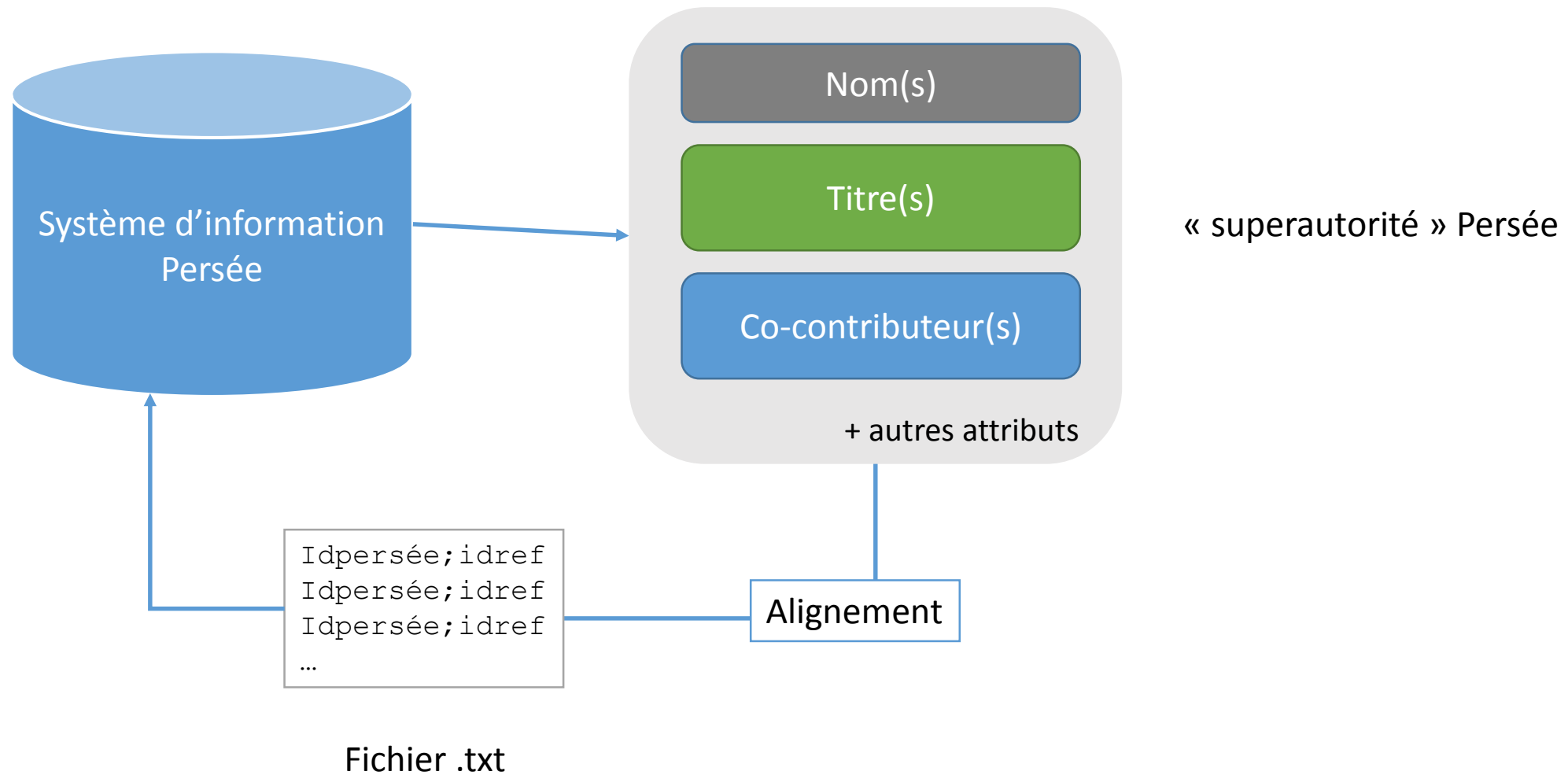
100



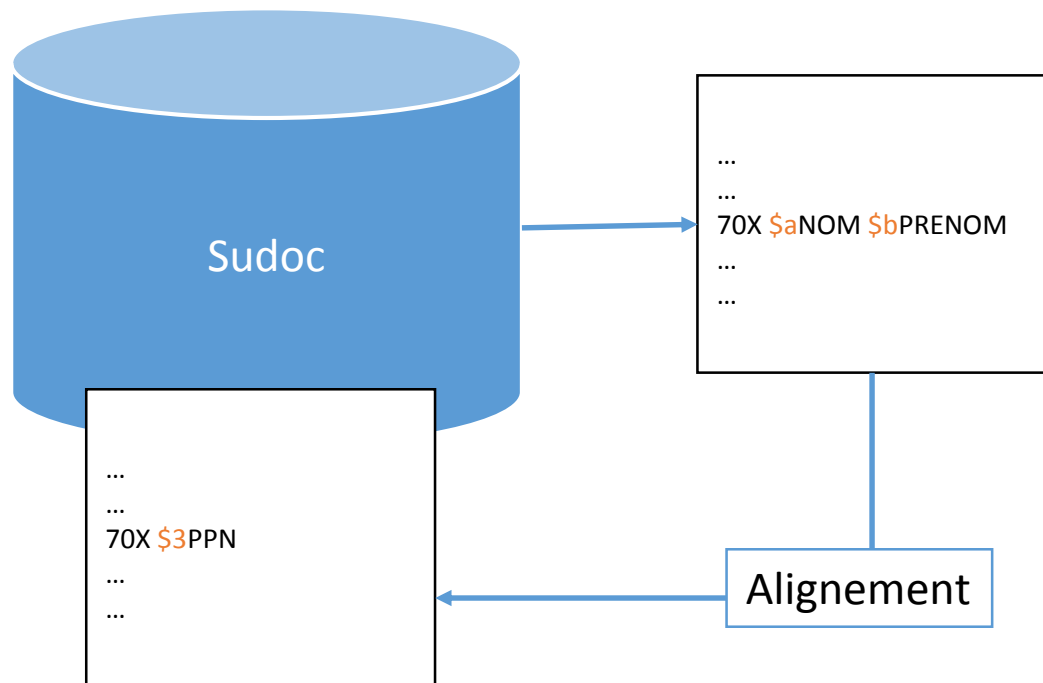
Scénarios variés

- Auteurs Persée
- Points d'accès dans les notices d'Ebooks CAIRN
- Mentions d'auteur dans les notices d'Articles Wiley (ISTEX)

Persée → IdRef



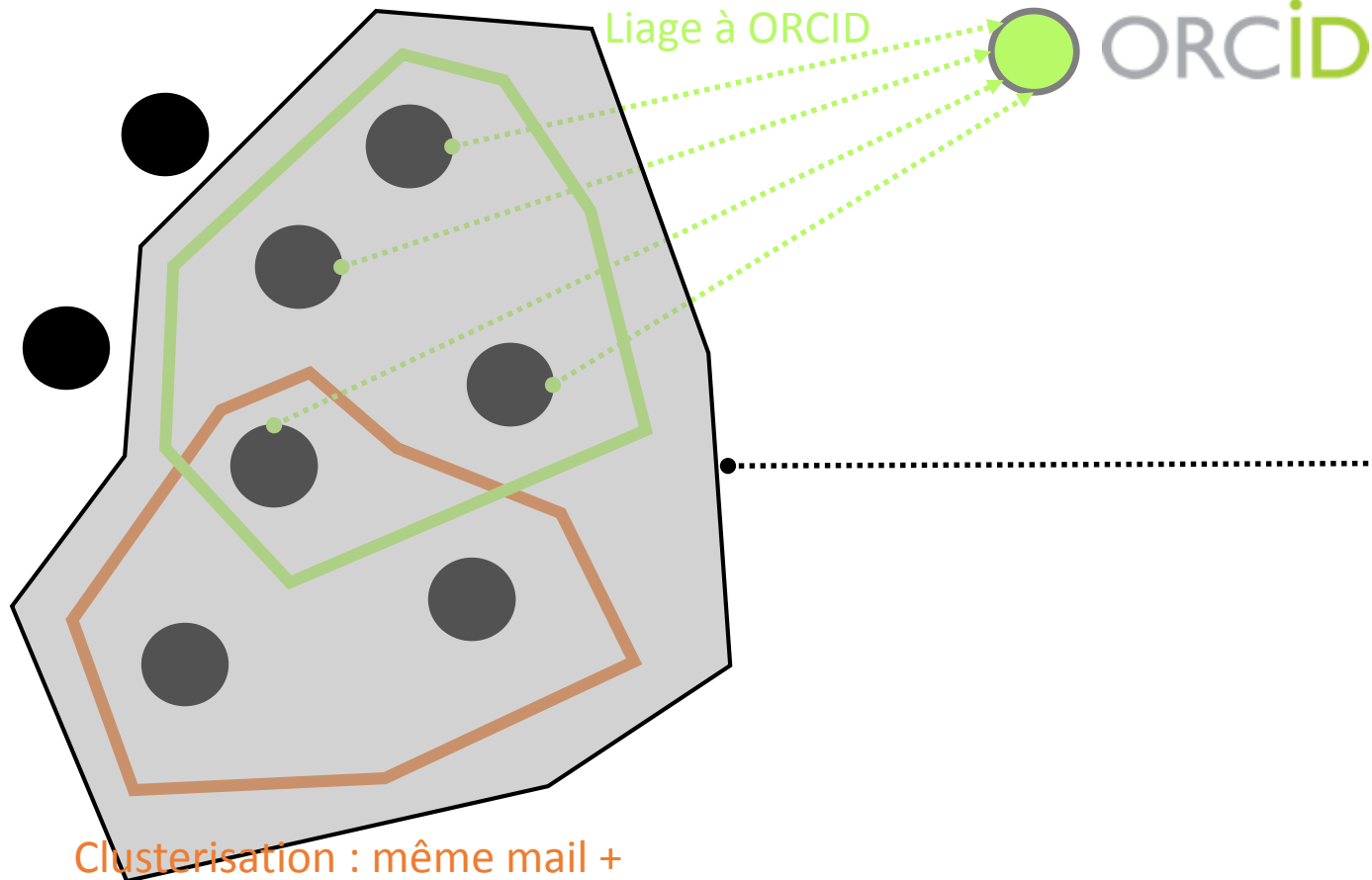
Pour aider le Cercle Ebooks CAIRN (oct, 2016)



89% des points d'accès liés
5 000 ebooks
(corpus facile...)

Auteurs d'articles Wiley (ISTEX)

Plein de Robert Sullivan



En cours

11 millions de mentions d'auteur

400 000 liens vers 100 000 auteurs ORCID

Alignement vers IdRef en cours (bcp moins)

Clusterisation : même mail +
nom identique ou proche

IdRef

Anomalies

- Puces à l'oreille :
 - *Deux candidats différents avec cocontributeurs ou titres qui matchent*
 - ➔ 2 types d'anomalie possibles :
 - doublons IdRef
 - mauvais lien Sudoc ➔ IdRef (= superautorité polluée)
- On écarte ces cas
 - ➔ À corriger
- Mais certaines anomalies dans les données sont silencieuses
 - Et causent de mauvais alignements (et impactent le taux de précision) ☹️
 - ➔ Problème profond pris en compte par Qualinca

Limites : Taux de précision < 100%

- Création de quelques liens erronés
 - Mais taux de précision proche de 100% (variable selon les corpus)
 - égal voire supérieur au liage par les catalogueurs en situation réelle
- Ces faux positifs
 - Ne sont pas dues au programme
 - Mais sont dus à des erreurs dans les données (doublons ou faux liens)
 - ➔ Qualinca

Limites : Taux de rappel < 100%

- Vrais négatifs : la bonne autorité n'est pas dans IdRef
 - Dc : la créer dans IdRef ou lier vers une autre source !
 - question stratégique ouverte ?
- Faux négatifs : elle existe dans IdRef, mais on ne trouve pas la bonne autorité. Ou on la trouve mais on n'ose pas lier...
 - Prudence légitime
 - Que faire ?
 - Il faut plus d'attributs
 - Il faut des règles plus nombreuses et complexes
 - L'humain doit intervenir, pour les cas difficiles ... mais avec quel outil ?
 - ➔ Qualinca