Transition bibliographique

Qu'attendre des agences pour la FRBRisation des catalogues ?

Chantiers de FRBRisation par traitements algorithmiques de l'ABES et de la BnF

Olivier Rousseaux (Abes)
Etienne Cavalié (BnF)

Journée Systèmes & Données - 14/11/2017

Plan

- Les chantiers Abes
- Les chantiers BnF
- Quelle collaboration entre les deux agences ?
- Quels bénéfices pour les autres bibliothèques ?

Les chantiers Abes

1. D'où part-on ?
2. Quels outils disponibles, quelle méthodologie ?
3. Quels résultats attendus ?

Les chantiers Abes

D'où part-on ?

Un contexte (en 2013)

- Fin de vie annoncée du système informatique (CBS) hébergeant le Sudoc

Le projet de FRBRisation du Sudoc

- Pas de développements informatiques
- *Initialement :* travailler sur les seules consignes de catalogage (en attendant un nouvel environnement technique)
- *Evolution vers :* exploiter les algorithmes de regroupements de données développés dans le CBS par OCLC (service existant) - sans développement d'interface de recherche publique

… quelques (bouts de) bibliothécaires pour y travailler

Les chantiers Abes

Outils et méthodologie - Principes

Une expérimentation :

- adaptant le service aux spécificités du Sudoc
  ⇒ en étant le + conforme possible au modèle FRBR
  ⇒ en s'appuyant sur les évolutions d'Unimarc
  l'Abes a établi des spécifications
  - de format
  - de choix dans les types de notices à regrouper
  - de types d'information à générer ou à remonter dans les Tr

- regroupant les notices bibliographiques selon des calculs de comparaison basés sur des clés titre-auteur
  ⇒ pour obtenir **un corpus de d'œuvres** + **des liens de type Manifestation → Œuvre** (enrichissant les données existantes)

- menée dans l'environnement de production sur la totalité des données du catalogue

  ⇒ dans la limite de pertinence des résultats (évalués sur échantillons)

Les chantiers Abes

Résultats obtenus / attendus

Réalisations

- Mise en production le 23 octobre 2017
  → ~1,4 M de "pré-notices d'œuvre" (dites notices de regroupement) créées
  → ~4,2 M liens créés entre notices bibliographiques et notices de regroupement (pour 16,5 M notices bibliographiques)
- Fonctionnement en base de production (en mode "vitrine")
- Processus dynamique : enrichissement quotidien par calculs sur toutes les mises à jour du catalogue (300 à 600 notices de regroupement mises à jour) : les regroupements ne sont pas figés

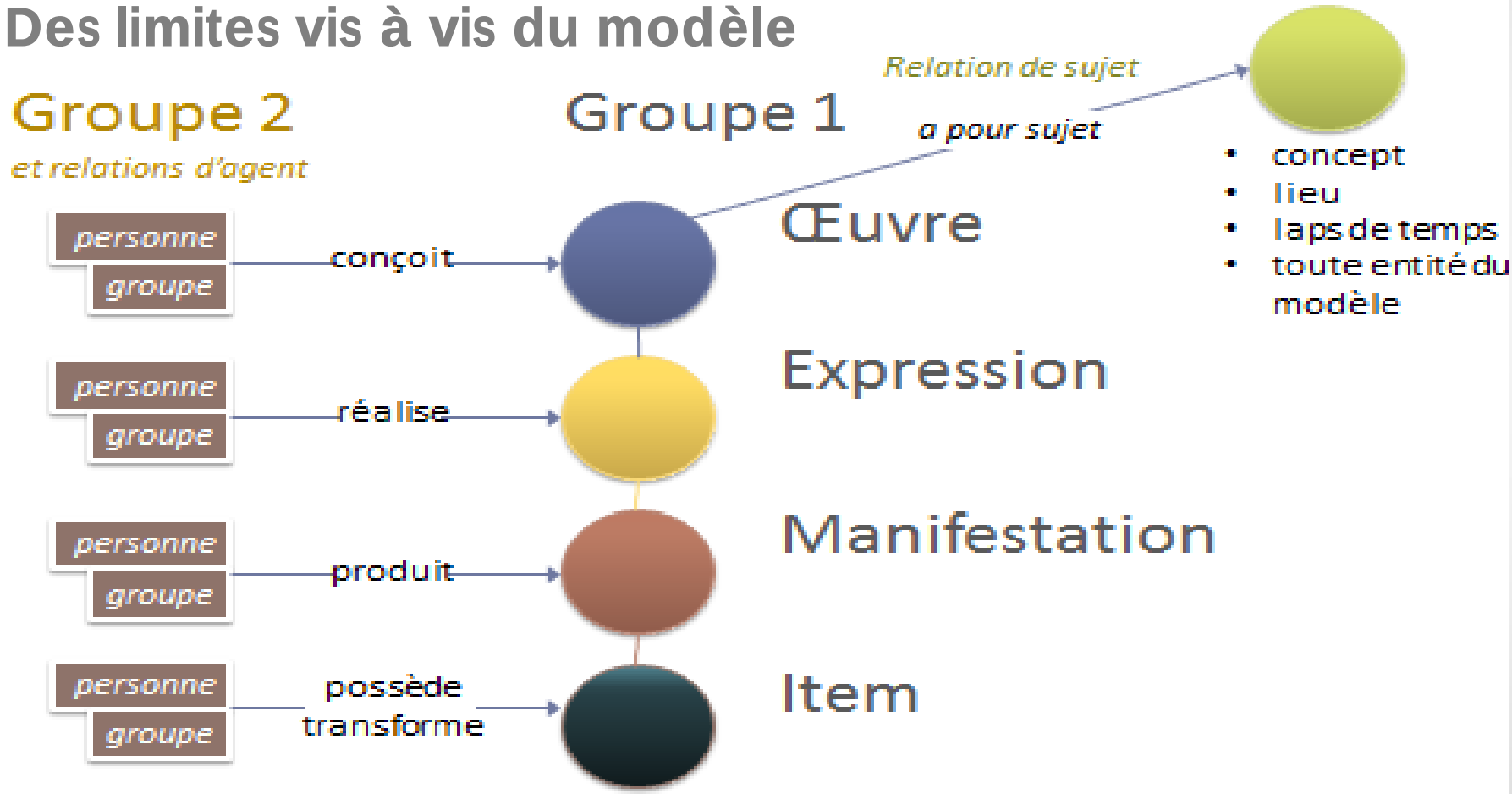
Des limites sur les données

- Traitement impossible des agrégats sans titre d'ensemble
- Traitement trop peu pertinent des ressources continues
- Pas de notice de regroupement générée pour une notice bibliographique seule à représenter une oeuvre
- ...

Les chantiers Abes

Résultats obtenus / attendus

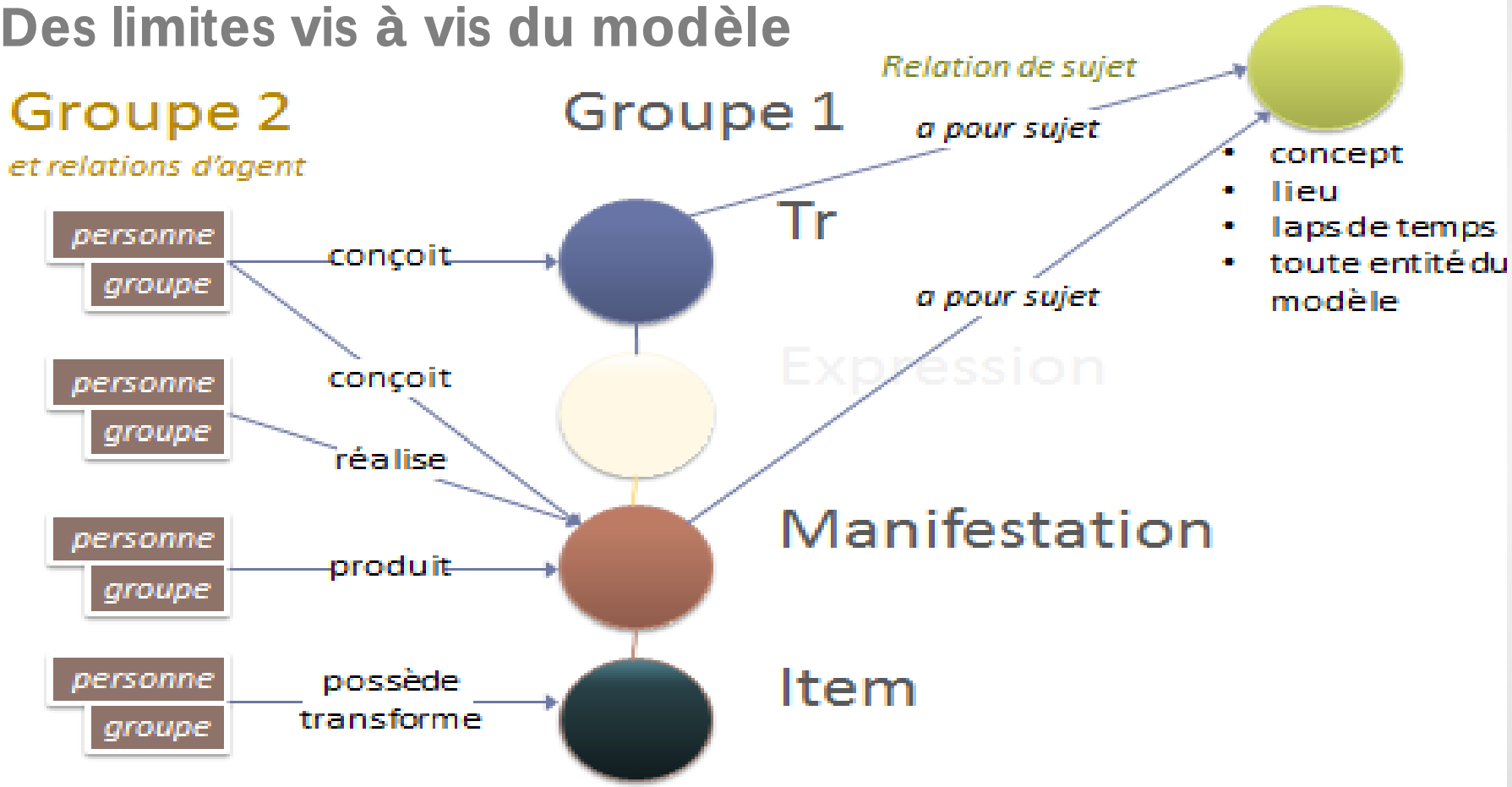
Des limites vis à vis du modèle



Les chantiers Abes

Résultats obtenus / attendus

Des limites vis à vis du modèle



Les chantiers Abes

Résultats obtenus / attendus

Expérimentation… en cours

- Evaluation qualitative à mener
  - sur les données existantes du catalogue
  - sur les données générées :
    → paramètres à ajuster
    → limites et insuffisances des algorithmes.

Chantiers à prévoir

- Amélioration des algorithmes existants
- Amélioration / enrichissement des données dans d'autres contextes (internes Abes / mode collaboratif [BnF] / …)
- Modalités de validation des données générées (dont établissement de règles de description des œuvres conformes à RDA-FR)
- Export / Exposition des résultats (dont développement d'une interface de consultation publique)
- Extension du traitement à tous les types de ressources
- Niveau de l'expression
- Et au-delà… processus de production conforme au modèle LRM

Les chantiers BnF

1. D'où part-on ?
2. Quels outils disponibles, quelle méthodologie ?
3. Quels résultats attendus ?

Les chantiers BnF

d'où on part

310.000 notices d'autorité Titre

Titre conventionnel : 100.000

Titre uniforme textuel : 10.000

Titre uniforme musical : 200.000

Une équipe d'experts Biblio et Autorités

Le projet data.bnf.fr

- Liens notices bib / notices d'autorité à la volée
- Fin 2015 : Reversement au catalogue
- RobotDonnées

Les chantiers BnF

Outils et méthodologie

Principes

- Partir de ce qu'on maîtrise
- Partir des données les plus exploitables
- Partir des outils existants (et élargir leur usage)
- Exploiter le travail des autres !
- Expérimenter par calculs à la volée, puis réinscrire les données en dur

Constat

Continuum entre l'automatique et le manuel

Les chantiers BnF

Outils et méthodologie

Réalisations

Liens notices d'autorité Titre <-> notices bibliographiques : ~200.000

Dédoublonnage de notices d'autorité Personnes physiques

Chantiers en cours

- Création d'œuvres : imprimés français du XXe siècle
- Agrégats XXe siècle
- Liens œuvres Films <-> œuvres Livres
- Liens œuvres Films <-> bande originale
- Dédoublonnage Auteurs

Chantiers prévus 2018

- Autres périodes, Autres types de documents (audiovisuel, cartes...)
- Nettoyages : préparer alignements et création d'oeuvres
- Les Expressions dans data.bnf.fr

Les chantiers BnF

RobotDonnées

Extraction des algorithmes testés sur data.bnf.fr

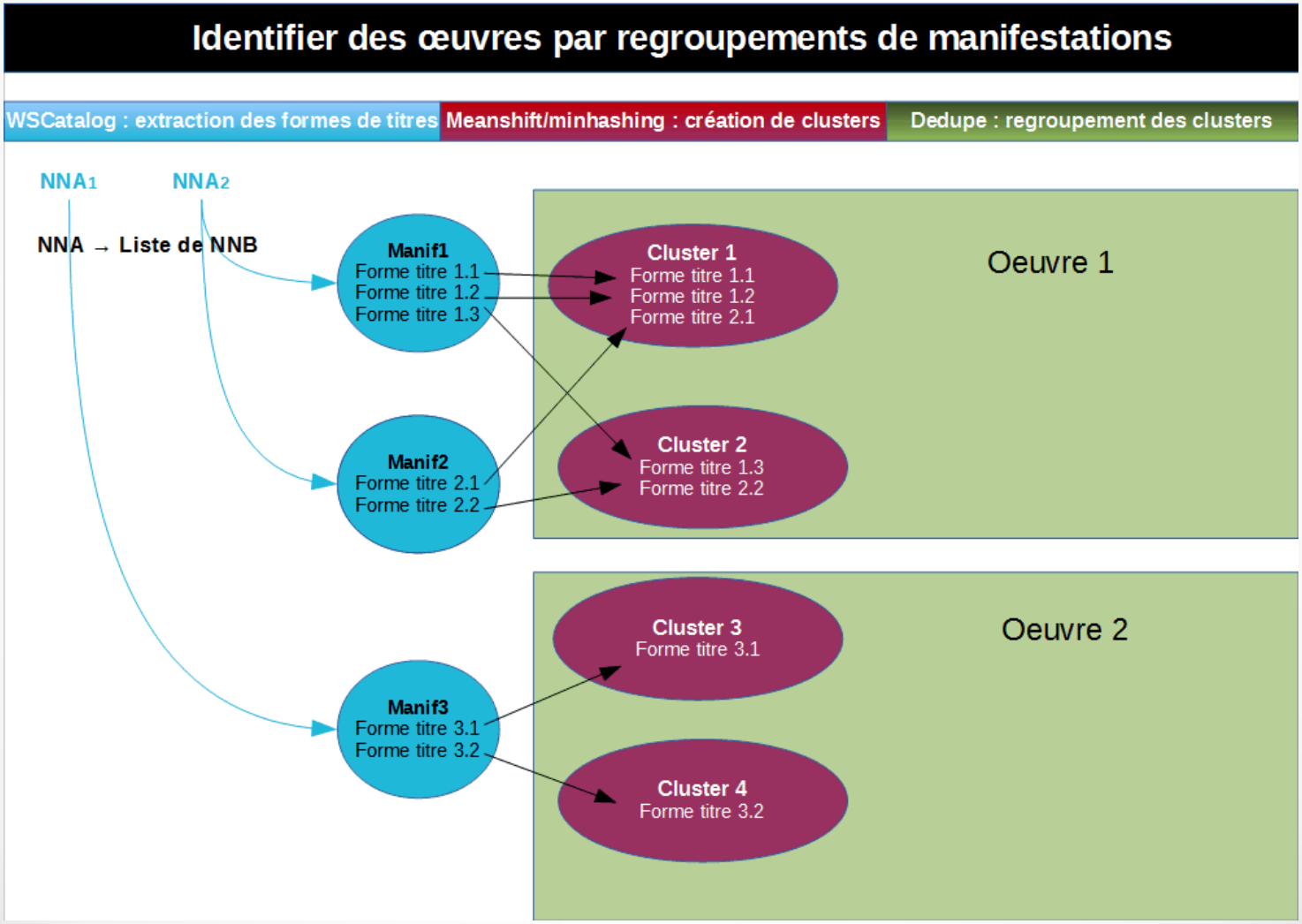
- regrouper les formes de titre identiques
- identifier les liens Titre original <-> Titre traduit
- générer des notices d'œuvres
- travailler sur des corpus

Petite équipe pilote

- premières utilisations sur les œuvres textuelles françaises du XXe siècle

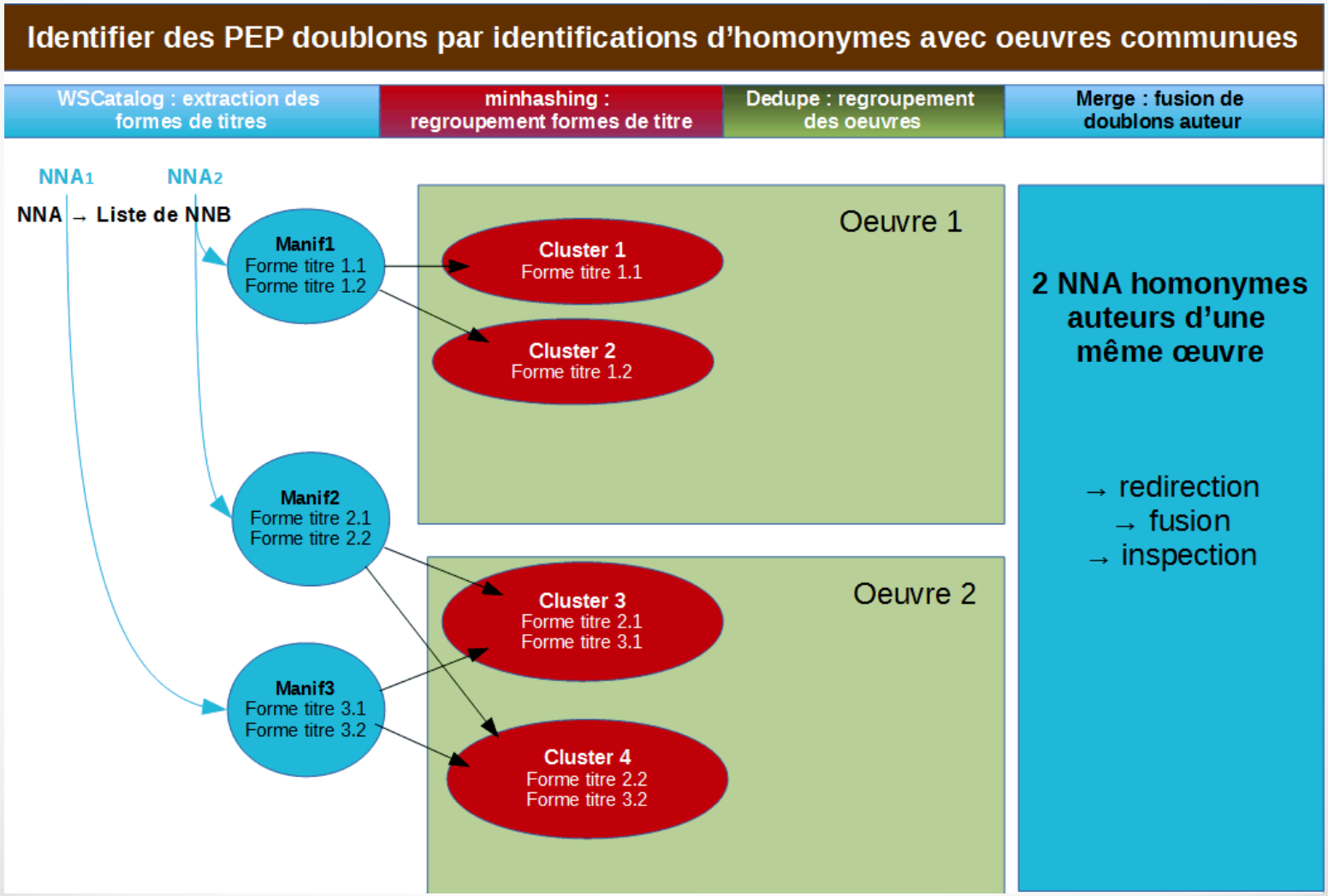
Les chantiers BnF

RobotDonnées - Principes



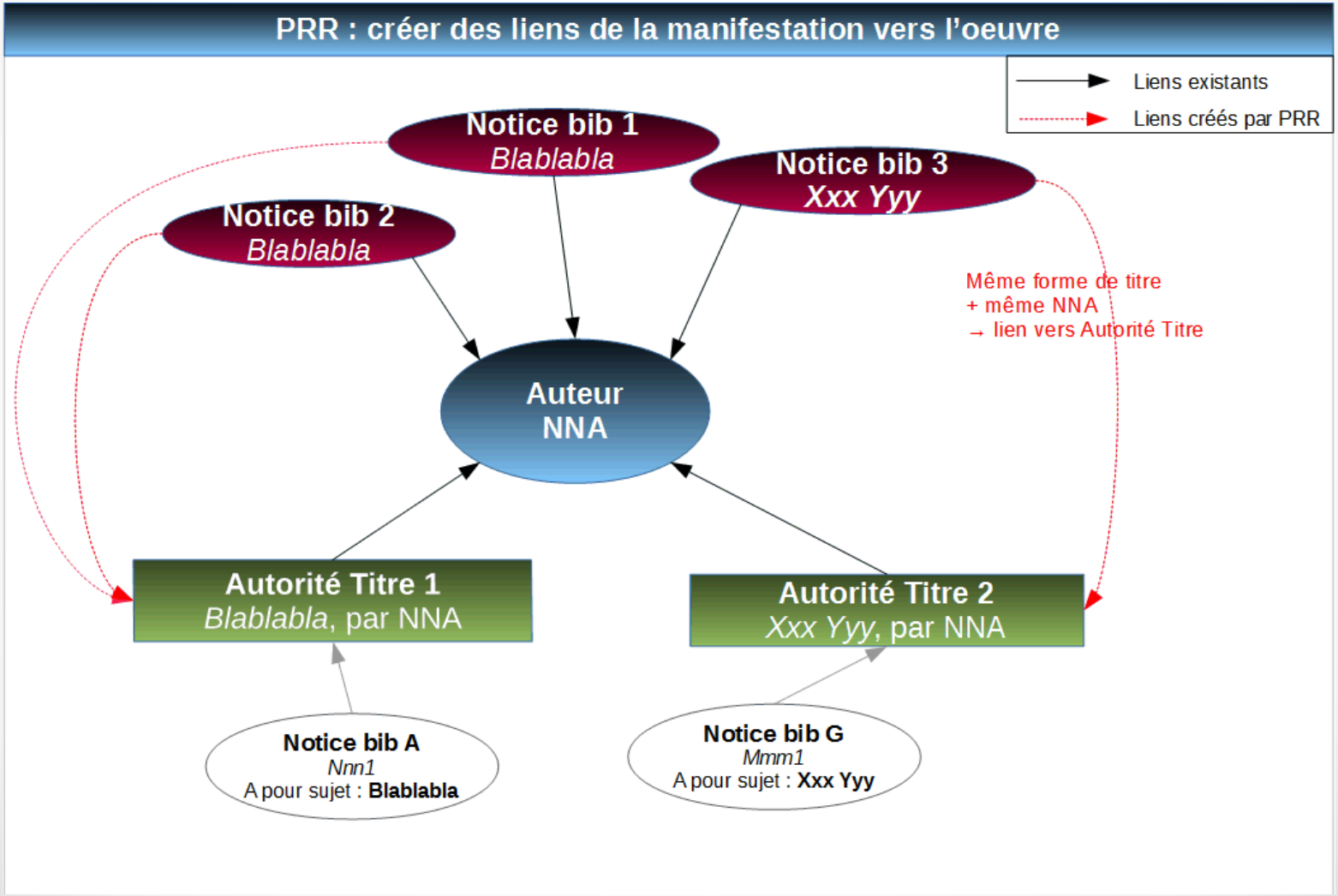
Les chantiers BnF

RobotDonnées - Principes



Les chantiers BnF

RobotDonnées - Principes



Les chantiers BnF

RobotDonnées - en vrai

WsCatalog

Authors
liste des numéros de

11928159;1212824

Complete cluster
rapatrie les autres PE

Notice types
m
types de notices (Gui

Doc types
a
types de documents

Rules

Add

Author rules

Add

Title zones
245$a+245$i; 245$
zones de titre, e.g 2

Limit
2000
nombre maximal de

Stop words (optional

liste de mots à suppr

Min word size
3
taille minimale d'un

Responsibility keywo
, par; trad. de
liste de motifs-clés (

Skip keywords (optic
, roman; , comédie
Nettoyer le contenu p

Clean subzones (opti
245$a

Run 72671 of *worksbuilder.py* (👍)

wscatalog.py #70155 > minhashing.py #70202 > worksbuilder.py #72671

Parameters | Log | **Result** | Statistics

+ create prr run from this result

Related works (⤓ zip)

- Les Stèles de la Ghorfa
- L. Poinssot et R. Lantier. Un sanctuaire de Tanit à Carthage
- Les Ruines de Thugga et de Thignica au XVIIe siècle
- Les Fouilles de Dougga en avril-mai 1903
- Cratères et candélabres de marbre trouvés en mer près de Mahdia
- Trois inscriptions chrétiennes de Tubernoc
- Notice sur une statue de saint Jean-Baptiste découverte en 1898 dans l'église de Rouvres. (Jan
- Bronzes trouvés en mer près de Mahdia (Tunisie
- Mosaïques d'El-Haouria (Plaine de Sidi Nasseur Allah
- L. Poinssot. Inscriptions Libyques de Tunisie
- Inferno
- Zama
- Les Inscriptions de Thugga
- Les Mosaïques de la "Maison d'Ariadne" à Carthage
- Inscriptions de Bulgarie
- L'autel de la Gens Augusta à Carthage
- L. Poinssot et Ch. Saumagne. Les "Piscines romaines" de Gafsa
- Épotides de bronze trouvées en mer près de Mahdia
- Statuettes trouvées en mer près de Mahdia (Tunisie
- Guide du Musée du Bardo (Musée Alaoui
- Quelques inscriptions de Tunisie
- La Carrière de trois proconsuls d'Afrique
- M. L. Poinssot. Moules de Carthage en terre cuite
- Les inscriptions d'"Uchi Majus" d'après les recherches du capitaine Gondouin
- Fouilles à Tubernuc
- L. Poinssot,... Macomades-Iunci
- A. Merlin et L. Poinssot. Le Proconsul d'Afrique Dardanius
- L'Archéologie chrétienne en Tunisie (1920-1932). Relatori : Proff. L. Poinssot et R. Lantier
- A. Merlin et L. Poinssot. Marbres trouvés en mer près de Mahdia (Tunisie
- Inscriptions inédites de Lambèse et de Timgad
- Inscriptions de Dougga
- Note sur une statue de saint Jean-Baptiste découverte en 1898 dans l'église de Rouvres

Données de l'oeuvre


```
{
    "notice_type": "TIC",
    "alternative_titles": [],
    "manifs": [
        "ark:/12148/cb43878956r",
        "ark:/12148/cb438505282",
        "ark:/12148/cb435902694",
        "ark:/12148/cb442175370",
        "ark:/12148/cb43659605h",
        "ark:/12148/cb43661260t"
    ],
    "language": [
        "eng"
    ],
    "preferred_title": {
        "lang": "eng",
        "zone": null,
        "title": "Inferno"
    },
    "date": null,
    "authors": [
        [
            "Dan Brown",
            70,
            100,
            14529449
        ]
    ],
    "id": "dbd6a75020c18153cd0615a8a02e6130",
    "creation_date": "2017/05/03"
}
```


le	date
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305
hing	20171012-1305

Les chantiers BnF

Résultats attendus

Enrichissements du catalogue

Création d'œuvres : imprimés du XXe siècle

Clusters d'agrégats

Liens entre œuvres

Amélioration des notices

Zones codées

Nettoyage des informations

Evolution des métiers & des compétences

Collaboration entre agences

L'hybridation des FRBRisations respectives esquissera les contours d'une base d'oeuvres commune

Des identifiants qui se (re)connaissent déjà

- (ré)alignement des données

  → enrichissements réciproques

D'autres registres de convergence

- via la normalisation (autour de RDA-FR)
- avec la réflexion sur un modèle de données pour une production commune (projet de Fichier national d'entités (FNE))

Récupérer les données FRBRisées

- Partager : ne pas refaire/racheter ce qui a déjà été fait
- Phase de transition
- Taux de couverture des données enrichies (FRBRisées ?) pour un catalogue de bibliothèque de lecture publique

Vers un catalogage LRM-natif

Chantier “prodMD”

Intermarc-NG

Pour toute question

https://www.transition-bibliographique.fr/nous-contacter/