Affiliations et rapprochement avec le RNSR

Département des outils d'aide à la décision
(MENESR – SCSESR - SIES)
19 juin 2015

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

1 Le Répertoire National des Structures de Recherche (RNSR)

Répertorier et attribuer un identifiant unique à chaque structure de recherche

Suivre la démographie des structures de recherche

- Tutelles
- Localisation géographique
- Disciplines ERC
- Site web
- Directeur

Suivre la recomposition des structures de recherche

- Simple renouvellement (mais avec changement d'identifiant)
- Fusions / éclatements / intégrations

Problématique : rapprocher le réservoir de notices bibliographiques Conditor avec le RNSR

- L'affiliation des auteurs des travaux scientifiques mentionne très souvent la structure de recherche
- La mention de cette structure est par contre non-normalisée

La méthodologie d'appariement

2.1

Un travail conjoint avec l'équipe Conditor (INIST)

- Une normalisation en amont du RNSR
  - Libellé complet (suppression des mots vides)
  - Sigle (suppression de la ponctuation)
  - Ville postale (troncature sans le mot « cedex »)
- Pour chaque affiliation, une extraction des informations suivantes:
  - Pays
  - Code postal et ville
  - Adresse postale
- Sur le reliquat de l'affiliation, recherche de quatre types de libellé relatifs à un laboratoire
  - Libellé complet
  - Couple label-numéro
    - » Liste fermée de mots-clé pour le label (UMR, UMS, etc…)
    - » Numéro de 2 à 4 chiffres
    - » Possibilité d'intercaler de 0 à 5 mots entre le label et le numéro
  - Sigle
    - » Si le sigle est partagé par plusieurs structures, ajout de la ville pour créer la clé d'appariement
  - Numéro seul
    - » Le label peut entraîner un non-matching (UMR attendu Vs UMS dans l'affiliation)
    - » Si le numéro est partagé par plusieurs structures, ajout de la ville pour créer la clé d'appariement
- Contrôle systématique par la recherche de la ville du laboratoire dans l'affiliation

Les premiers résultats

2.2

Résultats liés aux cinq bases initiales constitutives de Conditor

- HAL, Web of Science , INIST, INRA, INRIA
- Environ 630 000 affiliations référencées

	Avant contrôle ville			Après contrôle ville		
Critère appariement	**Effectifs**	**Cumul***	**%**	**Effectifs**	**Cumul***	**%**
Libellé complet sans mot vide	195 083	195 083	45 %	153 295	153 295	45 %
Label + numéro	272 410	320 991	63 %	237 956	272 094	81 %
Sigle	248 943	425 792	98 %	193 951	328 882	97 %
Numéro de labo	267 109	**433 384**	100 %	240 071	**337 719**	100 %

*Logique additive : le matching par numéro seul n'est réalisé que si les trois autres méthodes sont infructueuses

	Effectifs	Taux d'appariement
Avant contrôle ville	433 384	**68,5 %**
Après contrôle ville	337 719	**53,4 %**
Nombre total d'affiliations	632 636	-

3 Des pistes pour améliorer le rapprochement Conditor-RNSR

Un enrichissement du RNSR

- La complétude des enregistrements existants
  - 8 089 structures ont une ville postale (87,3 %)
  - 4 446 structures ont un sigle (47,9 %)
  - 9 194 structures ont un libellé complet (99,2 %)
- L'ajout de structures car le RNSR ne couvre pas tous les couples label-numéro détectés dans les affiliations

L'étude et la prise en compte de la succession et du regroupement de structures

- Succession : pouvoir ventiler l'historique de la production scientifique selon la composition actuelle des structures de recherche
  - En l'état, non-exhaustivité des liens
  - Exemple de deux structures consécutives ayant les mêmes libellé, sigle et ville
- Regroupement : pouvoir régler le grain d'observation de la production scientifique
  - Exemple des équipes INRIA
  - Etude fine du RNSR à mener également

La détection des tutelles dans les affiliations

- Effectuer un nouveau contrôle de cohérence
- Découvrir (par déduction) de nouveaux appariements Conditor-RNSR
- Le travail est déjà engagé

Des stratégies complémentaires

4

Banques de CV et connexion avec les bases RH

- La source de données
  - Candidatures 2013 à la prime d'encadrement doctoral et de recherche (PEDR)
  - 5 800 CV au format PDF liés sans ambiguïté à une structure de recherche
  - Production scientifique pendant la période 2009-2012
- Traitement expérimental pour l'extraction automatique de référence bibliographiques
  - Reconstruction des paragraphes du CV
  - Détection des références bibliographiques grâce à Crossref
- Résultats
  - 55 886 travaux détectés dont 11 645 en 2011
  - Taux de couverture de la méthode estimé à **69,5 %**

Bases RH et référentiel ORCID

- Appariement de la base RH nominative du MENESR (Gesup – Persée) avec le référentiel ORCID
  - Focus sur l'année 2011
  - Prise en compte de l'homonymie dans les deux bases à rapprocher
- 65 000 enseignants-chercheurs dont les nom et prénom sont renseignés
- 6 800 comptes ORCID détectés (10,5 %)
- Dont 1 000 comptes ORCID référençant des travaux scientifiques (DOI)
- 34 700 travaux scientifiques associés
- Dont 3 500 en 2011