

05/06/2020
DOING – MADICS
Chuanming.Dong@ign.fr

Alignement de bases de données pour l'extraction d'informations concernant les sols pollués

Auteur : **Chuanming Dong** (LASTIG, Univ Gustave Eiffel, ENSG, IGN, F-94160 Saint-Mandé, France ; Agence de l'environnement et de la Maîtrise de l'Énergie, ADEME, F-49004, Angers, France)

Encadrement : **Guillaume Touya** (directeur, LASTIG, IGN), **Catherine Domingues** (co-encadrante, LASTIG, IGN), **Philippe Gambette** (co-encadrant, LIGM, Univ. Gustave Eiffel)



Contexte et objectifs du projet doctoral

Les informations diffusées sur les sites pollués s'accumulent et se superposent dans le temps :

- diversité d'acteurs pour la connaissance des sites pollués, leurs suivi et réaménagement : ADEME, DREAL, BRGM, DG..., etc
- diversité de contenus :
 - structurées : plusieurs bases de données
 - non structurées : informations textuelles [documents de types variés]

→ Objectif : une **mémoire des sites** où événements = dates+lieux+activités+acteurs

Construction d'une base de données unique

- **Apparier** les éléments identiques dans différentes bases de données pour former une seule base
- **Enrichir** les bases de données en extrayant des informations dans d'autres bases
- **Standardiser** les informations enregistrées dans les champs partagés, et définir les champs essentiels pour la mémoire des sites

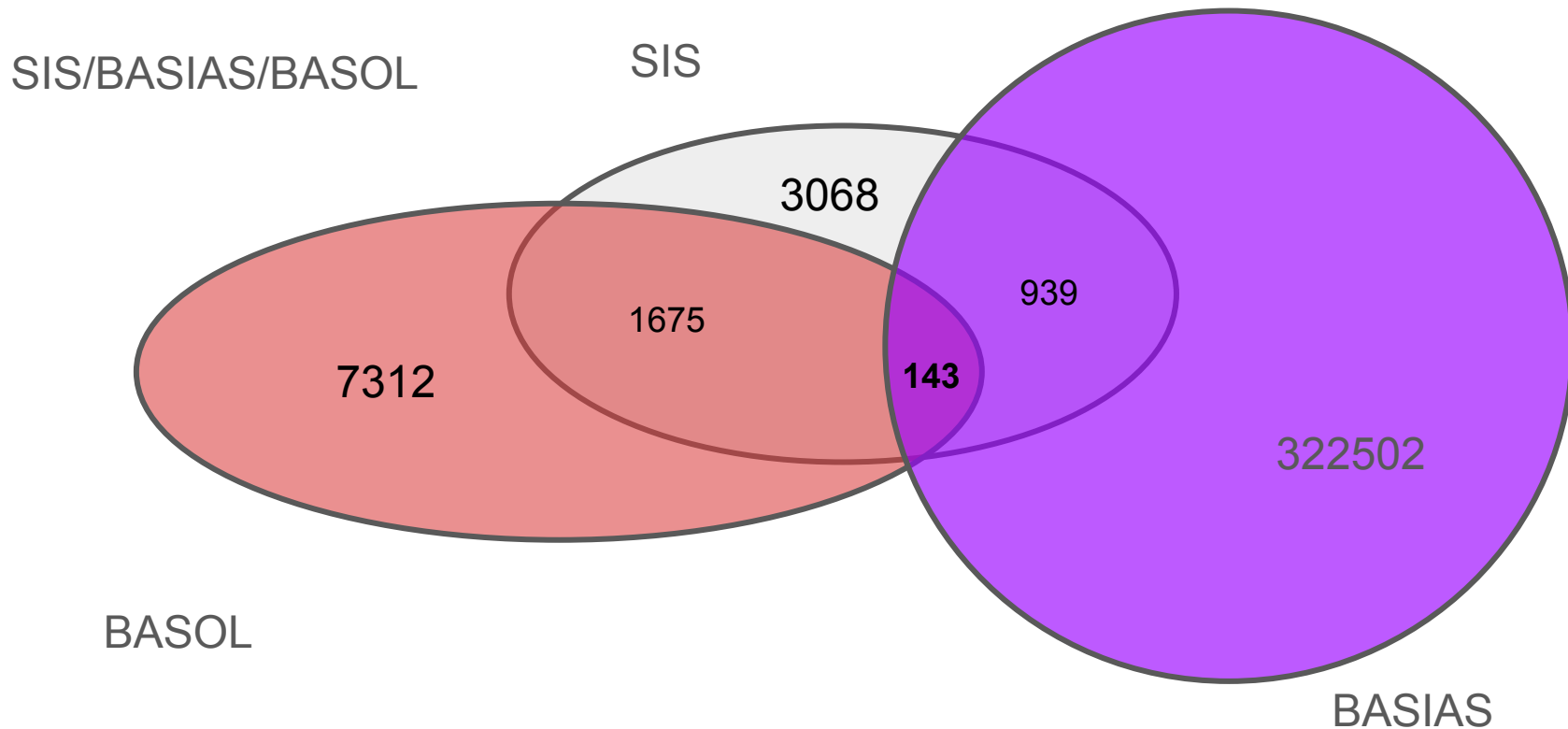
Bases de données disponibles sur la pollution industrielle

- **BASOL** : base de données sur les sites et sols pollués (ou potentiellement pollués)
- **S3IC** : base des installations classées
- **BASIAS** : Base de données d'Anciens Sites Industriels et Activités de Service
- **SIS** : Secteurs d'Information sur les Sols
- **ARIA** : Analyse, Recherche et Information sur les Accidents
- **BD ActiPoll** : typologies de substances potentiellement liées à des activités industrielles

Champs en commun pour un appariement :

- coordonnées géographiques
- adresse
- nom de l'entreprise
- polluants
- activités
- ...

Intersections entre les bases dans SIS



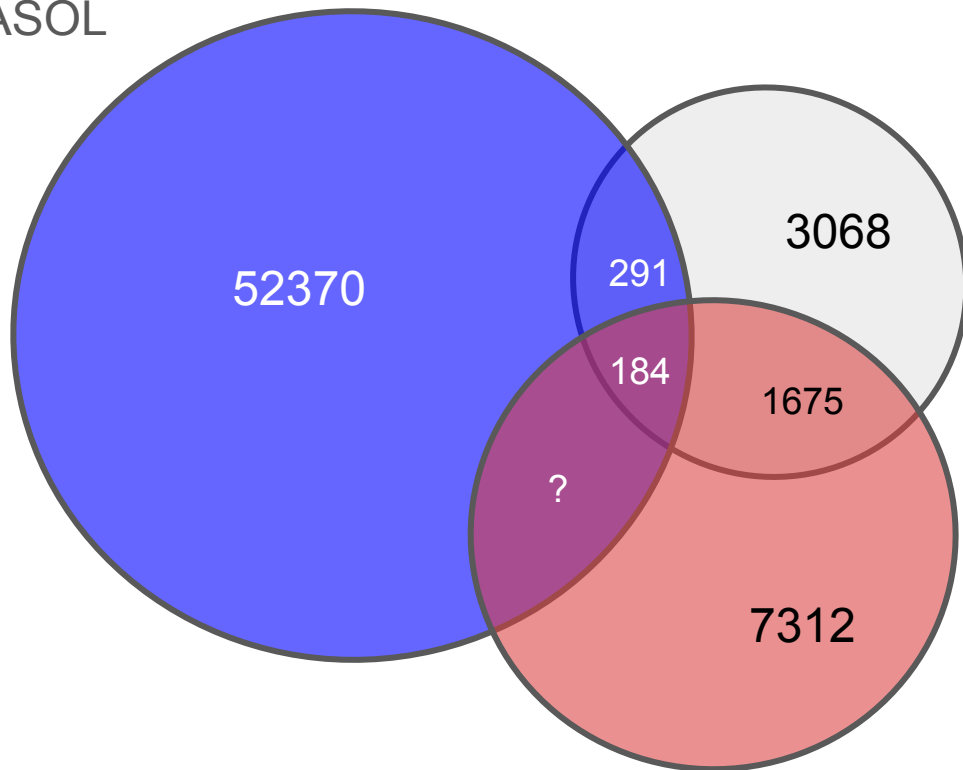
Insuffisance de recouvrement entre les bases

SIS/S3IC/BASOL

SIS

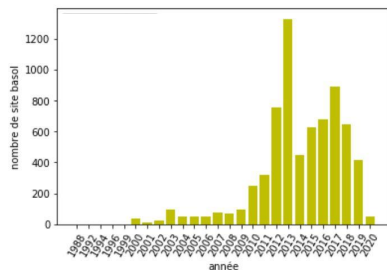
S3IC

BASOL

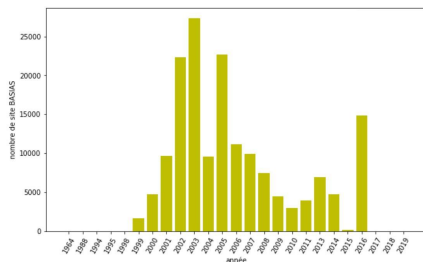


Difficultés d'appariement

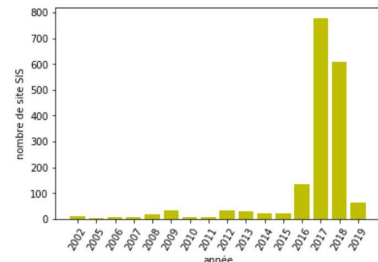
- Différences de format des données :
 - structurées (BASOL, BASIAS, S3IC) / texte brut (BASOL, documents S3IC, ARIA)
- Différences de temporalité / distributions des dates de création des fiches :



BASOL : potentiellement pollués



BASIAS : sites anciens



SIS : information sur les sols

- Différences de désignation dans les champs communs aux bases :
 - nom de l'entreprise
 - définition de l'activité
 - adresse, coordonnées géographiques
 - polluants impliqués

Problématique

- Quels critères utiliser pour définir les appariements ?
- Quels champs pour les critères ?
- Priorité entre les critères ?
- Comment adapter les méthodes d'appariement ?
- Comment évaluer la qualité des appariements réalisés ?

Critères

- bases de sites / entreprises (BASOL, BASIAS, S3IC, SIS) :
 - coordonnées ; exemple : (935104.0, 6294140.0)
 - adresse ; exemple : 11 quai du batardeau
 - nom d'entreprise ; exemple : Agence EDF-GDF Services
 - activité ; exemple : Travail des matières plastiques
 - polluants ; exemple : Hydrocarbures
- base de polluants (BD ActiviPoll, polluants BASOL, polluants BASIAS) :
 - nom du polluant
 - abréviation du nom du polluant / formule chimique

Méthodes d'appariement entre champs

- conversions de coordonnées géographiques (LambertI, Lambert93, WGS84)
- recherche et comparaison de sous-chaînes
- distances entre chaînes de caractères
- grammaires locales de normalisation des noms d'entreprises et des adresses
- utilisation de référentiels externes : symboles chimiques, formules moléculaires

Premiers résultats sur les sites

critère de distance

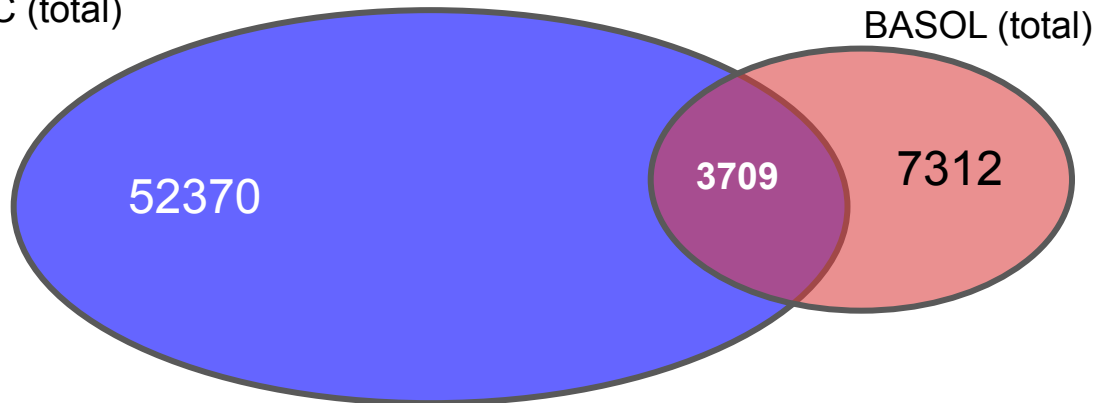
+

similarité des noms

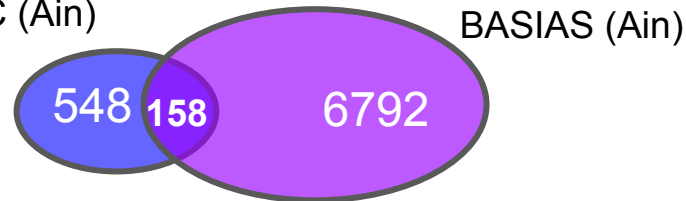
d'entreprise et

des adresses

S3IC (total)



S3IC (Ain)



Exemples d'appariement

- BASOL et S3IC

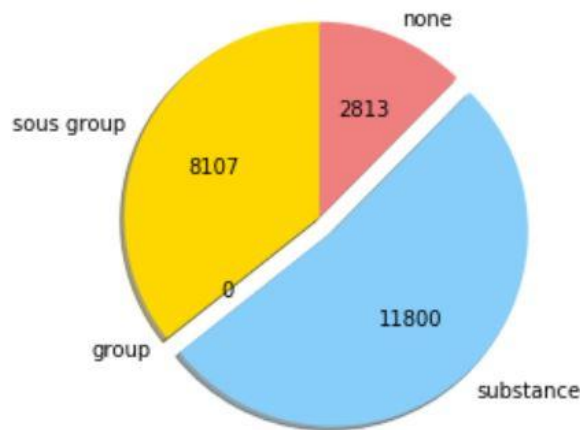
NOM BASOL	Adresse BASOL	NOM Installation	Adresse Installation	Sentence	Distance (m)
<u>CERPLEX</u>	Z.I. de Neuville en Ferrain rue du Vertuquet , Neuville-en-Ferrain	SARBEC	Zone Industrielle, Rue du Vertuquet BP 64, 59531 NEUVILLE EN FERRAIN	- Ancien site Rank Xerox devenu Cerplex, puis SARBEC.	674
<u>LOIRET AFFINAGE</u>	ZONE ARTISANALE DE VAUGOUARD RN 7 , Fontenay-sur-Loing	LOIRET AFFINAGE	Les Stations, RN 7, 45210 FONTENAY SUR LOING	LOIRET AFFINAGE est une société d'affinage d'aluminium localisée à Fontenay-sur-Loing (45) au lieu dit "Les Stations".	128
<u>BLAGDEN PACKAGING LYON</u>	112 chemin de Mure , Saint-Pierre-de-Chandieu	GRS VALTECH	112, Chemin de Mure, Zac du Dauphiné, 69780 ST PIERRE DE CHANDIEU	Le site est repris en 2004 par la société GRS VALTECH, spécialisé dans la valorisation des terres polluées par voie thermique.	4

Exemples d'appariement S3IC et BASIAS

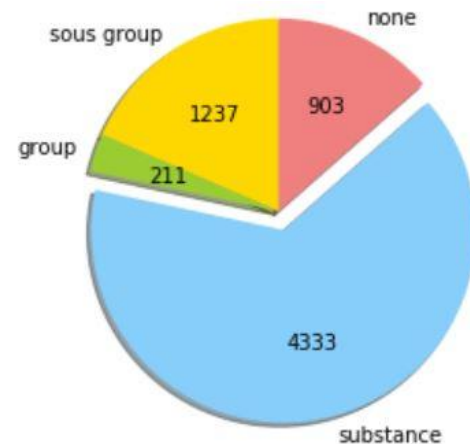
Nom BIC	Activité BIC	Adresse BIC	Nom Entreprise Basias	Activité Basias	Adresse Basias	Distance (m)
<u>SOCIETE BELLEGARDIENNE D'ABATTAGE SAS</u>	Activité Transformation et conservation de la viande de boucherie	6 rue Louis Armand 01200, BELLEGARDE-SUR-VALSERINE	<u>SEGAB, anc. SNUAB, anc. SEFA</u>	Abattoir	rue Louis Armand, , BELLEGARDE-SUR-VALS ERINE	8
<u>PELICHET ALBERT S.A- station traitement</u>	Activité Travaux de terrassement spécialisés ou de grande masse	Lieu-dit L'Ouche 01170, GEX	<u>Albert PELICHET SA</u>	Décharge sauvage	lieu dit "Grand Chauvilly", , GEX	630
<u>BFM RECUPERATION</u>	Activité	Zone ACTIPARC Nord 01990, CHANEINS	<u>BFM RECUPERATION SARL</u>	Dépôt de ferraille, ferrailleur	Zone Actiparc Nord, , CHANEINS	14

Premiers résultats sur les polluants

recherche de
sous-chaîne



occurrence dans ActiviPoll des polluants
cochés dans BASOL



occurrence dans ActiviPoll des polluants
dans le champ « Autres » de BASOL

Polluant Basol	Groupe	Sous-Groupe	Substance
métaux	métaux		
nickel			Nickel
phénol		phénol	
Antimoine Hydrocarbures dans les sédiments Dichlorométhane dans les gaz du sol		hydrocarbures	Dichlorométhane Antimoine

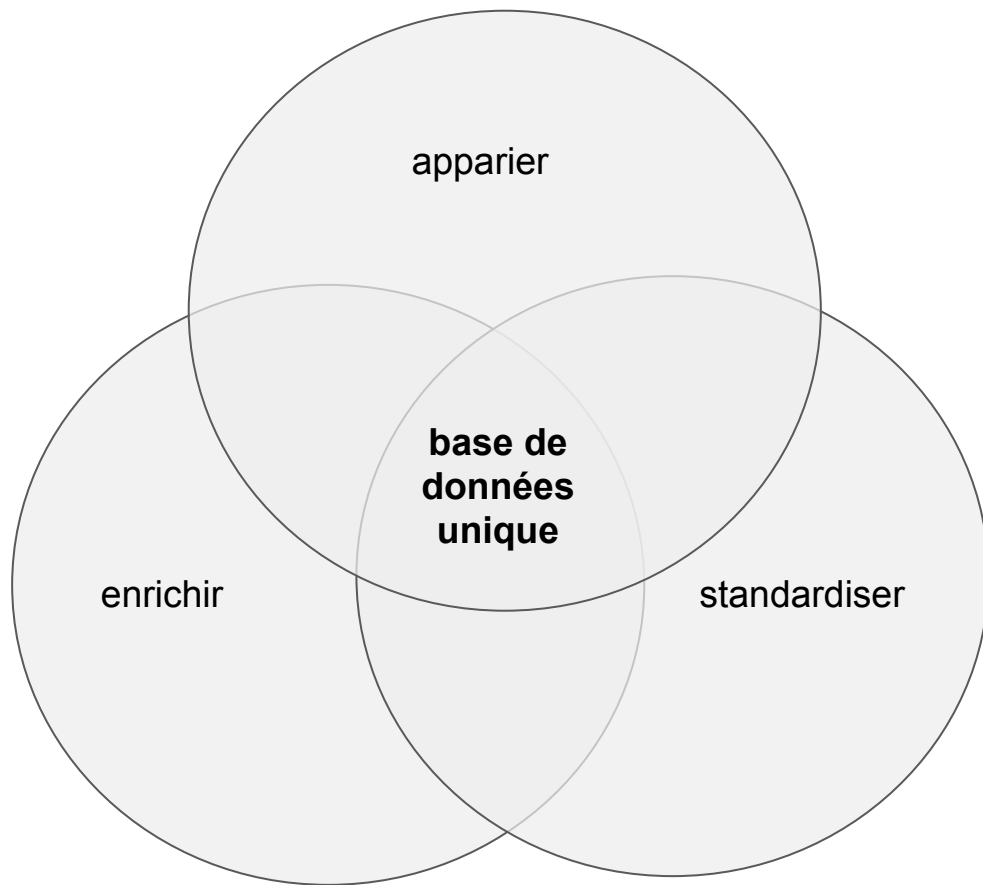
Évaluation

- étiquetage manuel d'un échantillon de 200 résultats d'appariement des sites :
 - estimation de la précision :
 - 92% pour BASOL / S3IC ;
 - 73% pour BASIAS Ain / S3IC Ain
- utilisation des fiches SIS avec lien vers BASOL et vers S3IC :
 - estimation du rappel :
 - 72% pour un appariement BASOL / S3IC avec 57% de précision (sur 29 fiches avec liens valides)
- échanges à venir avec des chimistes
 - évaluation des résultats d'appariement des polluants

Futur travail

- Améliorer les appariements :
 - grammaires locales pour les noms d'entreprises et les adresses
 - utilisation de référentiels externes polluants
 - méthodes d'appariement plus adaptées à la proximité attendue des chaînes de caractères
- Augmenter le corpus de validation pour l'évaluation des appariements
- Finaliser la fusion entre les bases des sites (potentiellement) pollués
- Entraîner sur le corpus de BASOL un extracteur des informations pertinentes (agent polluant, date et lieu de l'activité polluante, ancien exploitant, etc.) figurant dans les données textuelles :
 - l'extraction utilise l'appariement des bases de données
 - les informations extraites pourront enrichir la base de données de la mémoire des sites

Merci pour votre attention !



 chuanming.dong@ign.fr