27/05/2021
Séminaire archives et histoire des pollutions urbaines
Chuanming.Dong@ign.fr

Construction d'une mémoire des sites pollués : Fusion de bases de données et extraction d'événements

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 : **Catherine Dominguès** (directrice, LASTIG, IGN), **Philippe Gambette** (encadrant, LIGM, Univ. Gustave Eiffel)

IGN INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE

ADEME Agence de l'Environnement et de la Maîtrise de l'Energie

Contexte

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

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

Objectifs scientifiques de la thèse

Construire une **mémoire des sites** industriels potentiellement pollués

= un ensemble d'événements

un **événement** = date + lieu + action + acteur(s) + ...

Outils : Bases de données; TAL (traitement automatique des langues)

Objectifs scientifiques de la thèse

1. Réunir les informations caractérisant un site dans différentes bases : fusionner les bases de données
2. Désambiguïser et structurer les connaissances diffusées par les différents acteurs : extraire les événements dans des textes
3. Construire une frise chronologique dynamique des évènements industriels

La société BRODARD GRAPHIQUE était installée depuis 1959 sur la zone industrielle de Coulommiers.

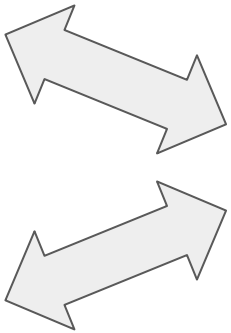
Applications possibles

- ADEME : évaluation du coût de réaménagement des sols pollués
- BRGM : consolidation de bases de données nationales
- DREAL : aide à la récapitulation des événements industriels par site
- Recherche en sociologie du risque, géographie : détection des risques de pollution à partir de ressources textuelles et analyse de leur impact
- Associations, citoyen·nes : analyse des risques de pollution identifier les risques

Construction d’une base de données unique

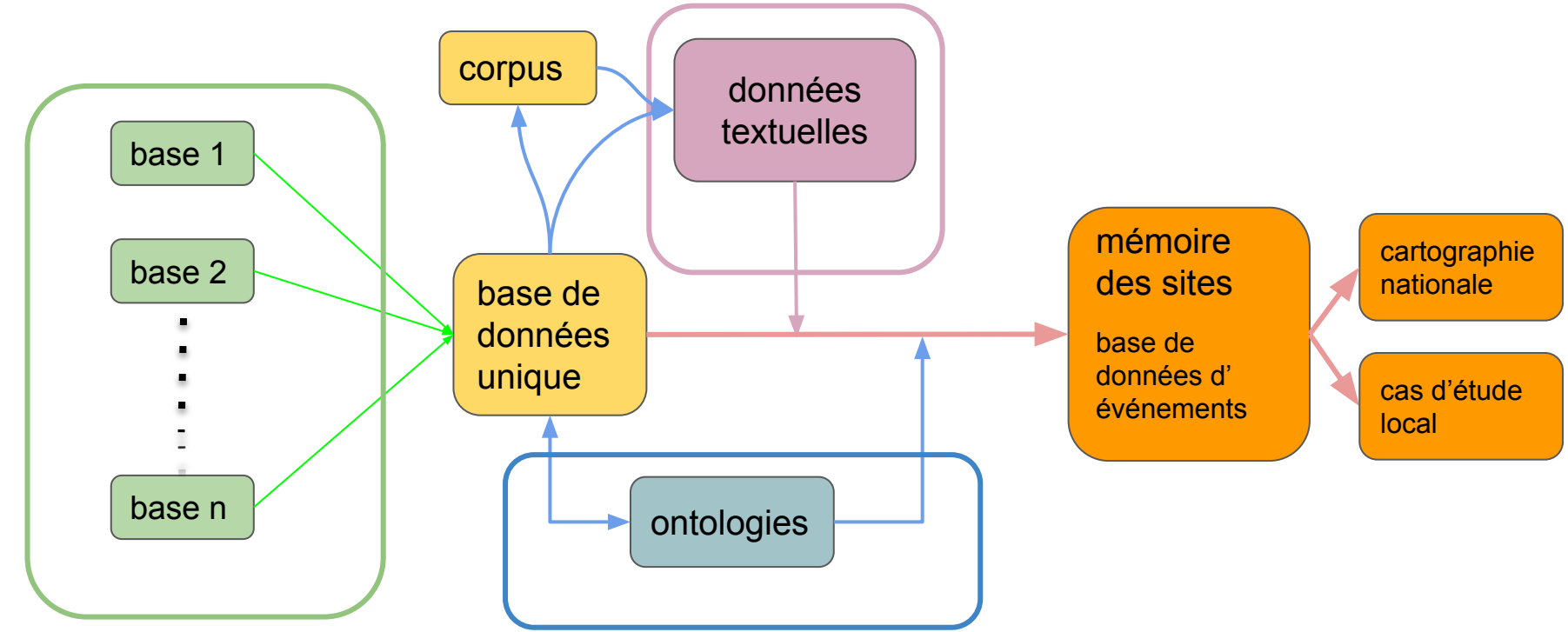
Apparier les éléments identiques dans différentes bases de données pour former une seule base

Enrichir les bases de données en extrayant des informations dans d’autres bases et ressources textuelles

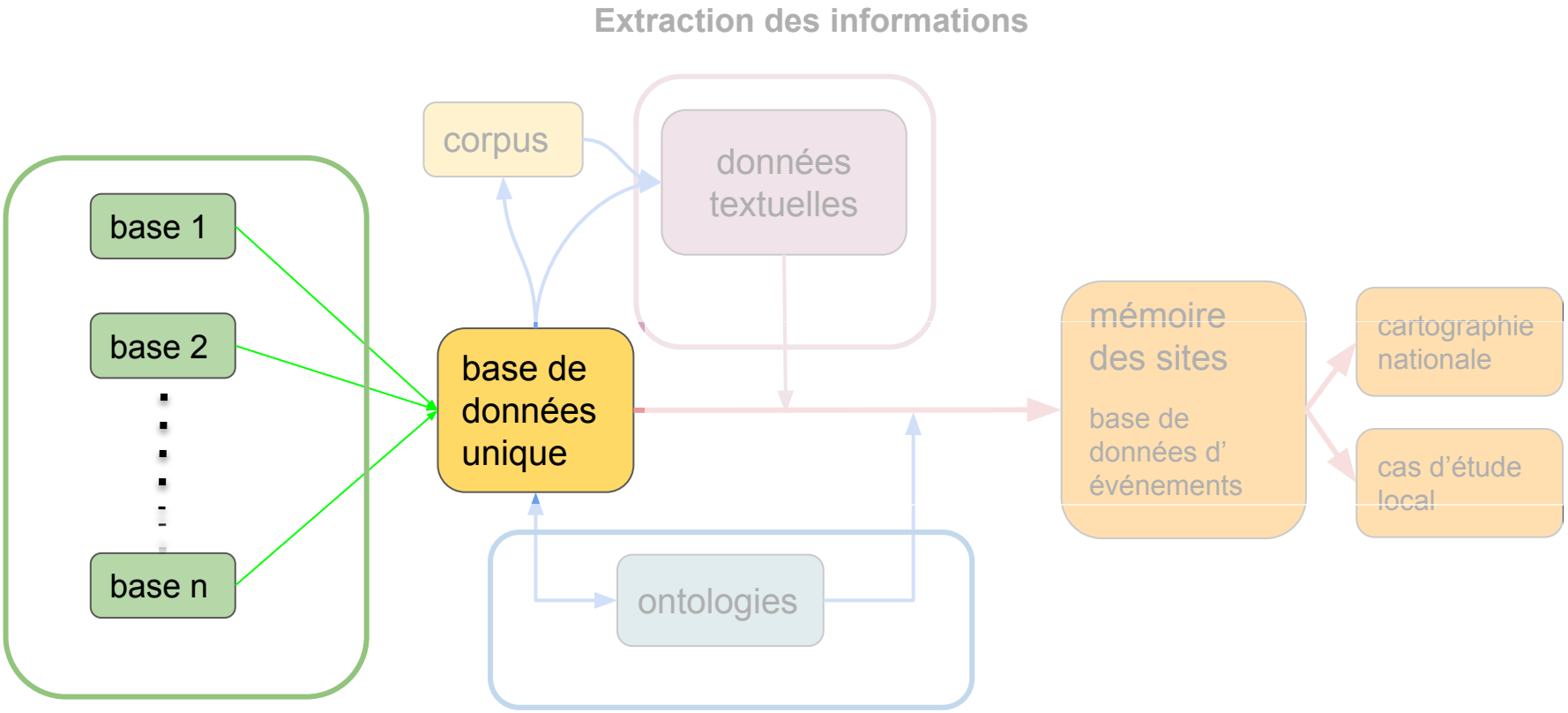


Standardiser les informations enregistrées dans les champs partagés, et définir les champs essentiels pour la mémoire des sites

Les étapes


Extraction des informations
corpus
données textuelles
base 1
base 2
base n
base de données unique
ontologies
mémoire des sites
base de données d' événements
cartographie nationale
cas d'étude local
Appariement entre les bases
Standardiser et structurer

Étape 1 : fusionner les bases de données



Appariement entre bases

- identification de plusieurs bases de données
  - bases de données de sites d’entreprises
  - bases de données de substances / polluants
- recherche des champs communs
- conception d’une méthode d’appariement entre champs
- évaluation de l’appariement

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 ActiviPoll :** typologies de substances potentiellement liées à des activités industrielles

Champs en commun pour un appariement :

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

Critères d’appariement (par ordre d’importance)

- bases de sites / entreprises (BASOL, BASIAS, S3IC, SIS) :
  - nom d’entreprise ; exemple : Agence EDF-GDF Services
  - adresse ; exemple : 11 quai du batardeau
  - coordonnées ; exemple : (935104.0, 6294140.0)
  - commune / code postal : DRANCY, 93029
  - activité ; exemple : Travail des matières plastiques de 1949 à 1981
- base de polluants (BD ActiviPoll, polluants BASOL, polluants BASIAS) :
  - nom du polluant
  - abréviation du nom du polluant / formule chimique

BASOL

Description du site

Nom : Agence EDF-GDF Services

Adresse : 17 RUE MARC SANGNIER

Commune principale : 93046 LIVRY GARGAN

Code - Libellé NAF : J1 - Cokéfaction, usines à gaz

Description : Le terrain situé au centre-ville de Livry-Gargan, d'une superficie totale de 6906 m2, a accueilli de 1881 à 1949 une usine fabriquant du gaz à partir de la distillation de la houille. Dans les années 1953-1954, le site est transformé en station gazométrique.

...

Date de dernière mise à jour : *28/05/2018*

Polluant(s) suspecté(s) ou suivi(s) :[4] HAP (Hydrocarbures aromatiques, polycycliques, pyrolytiques et dérivés)
Hydrocarbures et indices liés

BASIAS

1 - Identification du site

Date de création de la fiche : (*) 19/05/2003

Raison(s) sociale(s) de l'entreprise : GDF - GAZ DE FRANCE

Adresses :

Numéro	Bis Ter	Type voie	Nom voie
17		rue	MARC SANGNIER

Code INSEE : 93046

Commune principale : LIVRY-GARGAN (93046)

5 - Activités du site

Etat d'occupation du site : Activité terminée

Date de première activité : (*) 01/01/1881

Date de fin d'activité : (*) 31/12/1949

Historique des activités sur le site :

N° activité	Libellé activité	Code activité	Date début (*)	Date fin (*)
1	Cokéfaction (cokerie, distillation de goudron,	C19.10Z	01/01/1881	31/12/1949

Méthodes d’appariement entre champs

- recherche et comparaison de sous-chaînes
  “AHLSTROM Paper Group” | *“AHLSTROM CHANTRAINE”*
- distances entre chaînes de caractères
  “NOVARTIS AGRO” / *“SYNGENTA AGRO”* = 50%
- normalisation des noms d’entreprises et des adresses

  [*“HAUBOURDIN SA”* , *“Ets HAUBOURDIN”* , …] => *“HAUBOURDIN”*
- conversions de coordonnées géographiques (LambertI, **Lambert93**, WGS84)
  (2.333333, 48.866667) → (651094.18, 6863166.15)
- utilisation de référentiels externes : symboles chimiques, formules moléculaires
  “Présence de COHV en concentrations significatives” => *“COHV”*

Exemples d’appariement

- BASOL et S3IC

NOM BASOL	Adresse BASOL	NOM Installation	Adresse Installation	Sentence	Distance (m)
CERPLEX	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
LOIRET AFFINAGE	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
BLAGDEN PACKAGING LYON	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

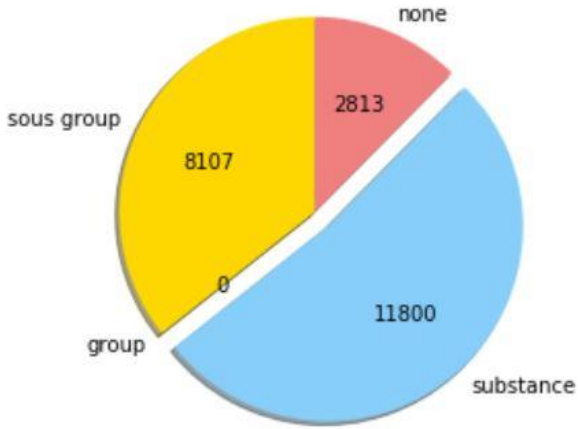
Résultat d'appariement BASOL-BASIAS

VP trouvé	VP annoté	FP(possible)	FN	NV	Précision	Gain
1553	1499	1254	1465	732	0,71	0,50

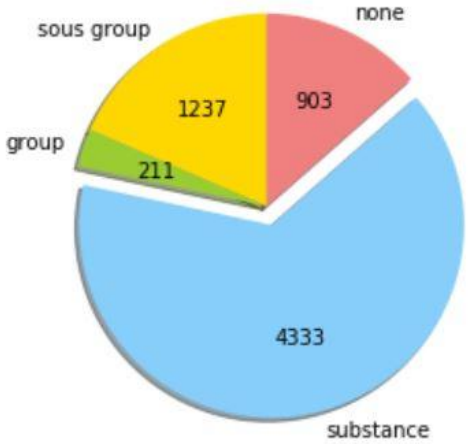
- **VP trouvé** (VPT) : « Vrais positifs trouvés », les liens BASOL-BASIAS retrouvés par notre algorithme et validés par les références BASOL dans la base BASIAS.
- **VP annoté** (VPA) : « Vrais positifs annotés », les liens BASOL-BASIAS qui n'existent pas dans BASIAS, mais retrouvés par notre algorithme et validés par la vérification manuelle.
- **FP** (possible) : « Faux positifs possibles », les liens BASOL-BASIAS qui n'existent pas dans BASIAS, retrouvés par notre algorithme mais ne peuvent pas être validés par la vérification manuelle.
- **FN** : « Faux négatifs », les liens BASOL-BASIAS valides qui existent dans BASIAS, mais ne peuvent pas être retrouvés par notre algorithme.
- **NV** : « Non valides », les liens BASOL-BASIAS qui existent dans BASIAS, mais ne peuvent pas être retrouvés par notre algorithme à cause des identifiants BASOL invalides.

Apparier 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

Ontologie Polluants et activités

- **BD ActiviPoll**

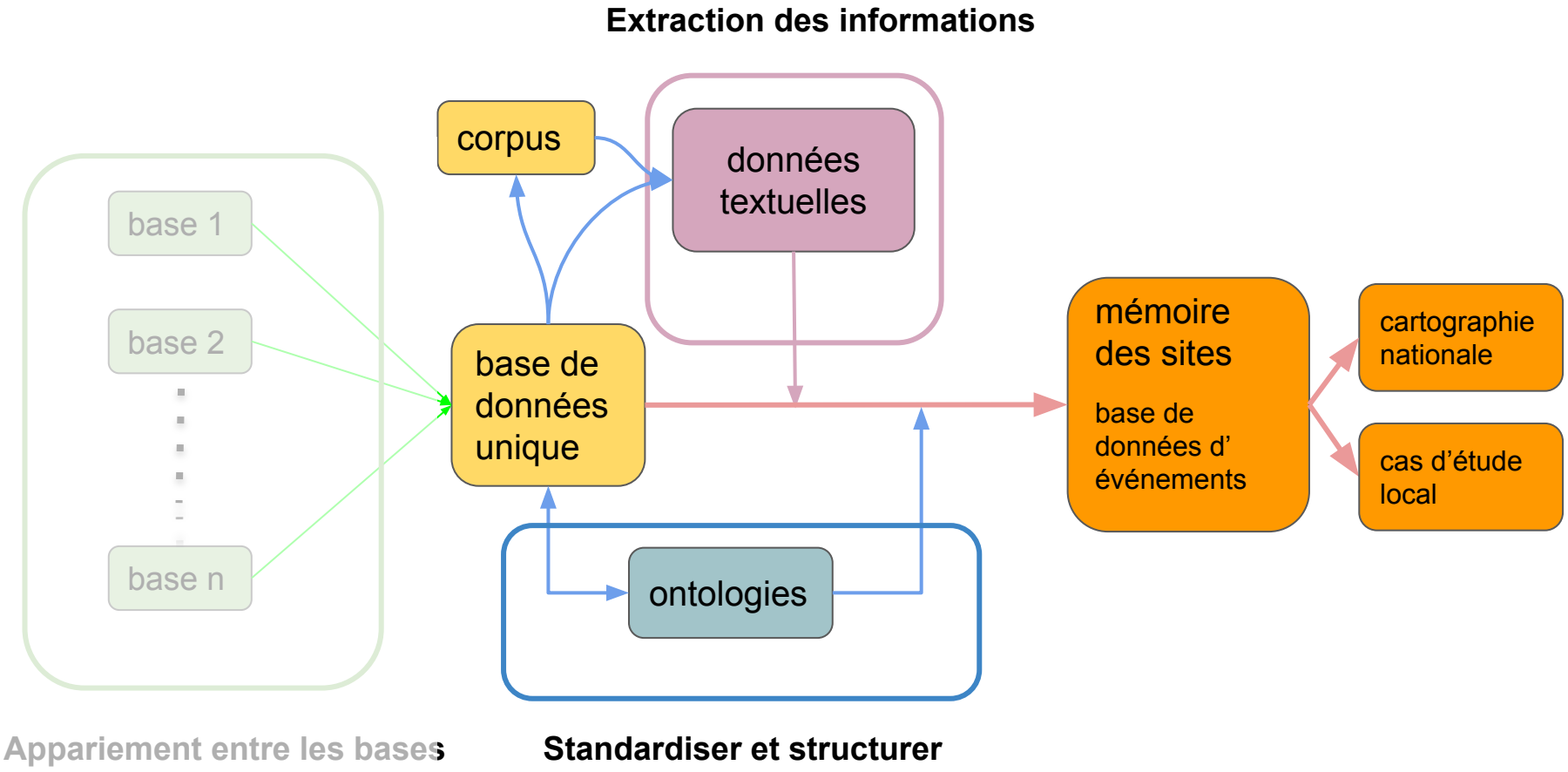
  - 3 niveaux
  - 6 groupes, 72 sous-groupes et 2654 substances

- **codification ICPE des activités**

Codes ICPE	intitulé ICPE	NAF 2008	intitulé NAF 2008
A11	A11 - Cultures	A01	Culture et production animale, chasse et services annexes
F	F - Industries extractives	B	Industries extractives
F11	F11 - Houillères	B05.10Z	Extraction de houille
D11	D11 - Extraction de pétrole et gaz naturel	B06	Extraction d'hydrocarbures
F2	F2 - Minerais métalliques (extraction de)	B07	Extraction de minerais métalliques

...

Étape 2 : extraire les événements dans des textes



Entraînement d’un extracteur

Corpus disponible : BASOL, BASIAS, ARIA, rapports préfectoraux

Qu'est-ce que nous extrayons ?

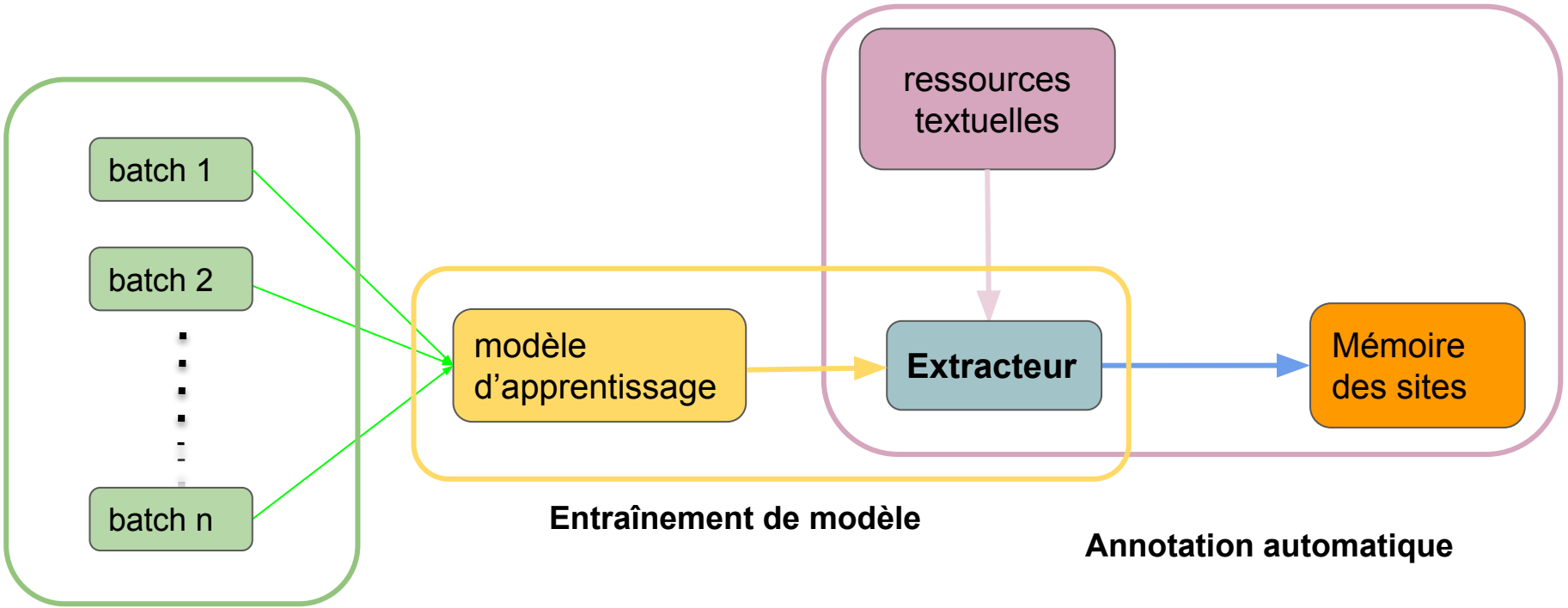
- Événement
  - Déclencheur d'événement
  - Acteur
  - Indicateur de temps
  - Indicateur de lieu
- Activité
- Pollution
  - substance
  - polluant

Entraînement de l’extracteur

Comment extraire ?

Méthodes :

- Annotation manuelle
- Annotation semi-automatique : extraction par patrons (avec Unitex)
- Annotation automatique : deep learning (Bi-LSTM et CRF) + CamemBERT


batch 1
batch 2
batch n
modèle d'apprentissage
ressources textuelles
Extracteur
Mémoire des sites
Entraînement de modèle
Annotation automatique
Annotation (semi-)manuelle

Données d’entraînement et de test

Pourquoi corpus BASOL ?

- Sols pollués : champ lexical pertinent
- Texte narratif : présence des évènements assurée
- Rapport gouvernemental : langage officiel

Unité d’entrée : phrase

données d’entraînement : 347 phrases annotées manuellement

données de test : 130 phrases annotées semi-manuellement

Annotation manuelle

Éléments annotés : groupe de mots

- **Format** : étiquette B (Begin), I (In) ou E (End) suivie d'un indicateur de catégorie

- **Catégories**
  - **N** : noyau évènement
  - **A** : activité industrielle
  - **T** : expression temporelle
  - **L** : localisation
  - **O** : objet
  - **I** : installation
  - **S** : substance
  - **U** : polluant non substance
  - **D** : dépôt de polluant
  - **R** : relation

Exemple d'annotation :

La	société	BRODARD	GRAPHIQUE
BI	II	II	EI

était	installée	depuis	1959
BN	EN	BT	ET

sur	la	zone	industrielle	de	Coulommiers	.
BL	IL	IL	IL	IL	EL	O

Résultats

	Rappel		Précision		F-mesure
Activité (A) :	0.6382	120 / 188	0.6250	120 / 192	0.6315
Déclencheur (N) :	0.6952	308 / 443	0.7738	308 / 398	0.7324
Localisation (L) :	0.6277	86 / 137	0.5512	86 / 156	0.5870
Temps (T) :	0.9282	194 / 209	0.8899	194 / 218	0.9086
Objet (O) :	0.8257	763 / 924	0.8903	763 / 857	0.8568
Relation (R) :	0.5376	157 / 292	0.6108	157 / 257	0.5719
Total :	**0.7423**	**1628/2193**	**0.7834**	**1628 / 2078**	**0.7623**

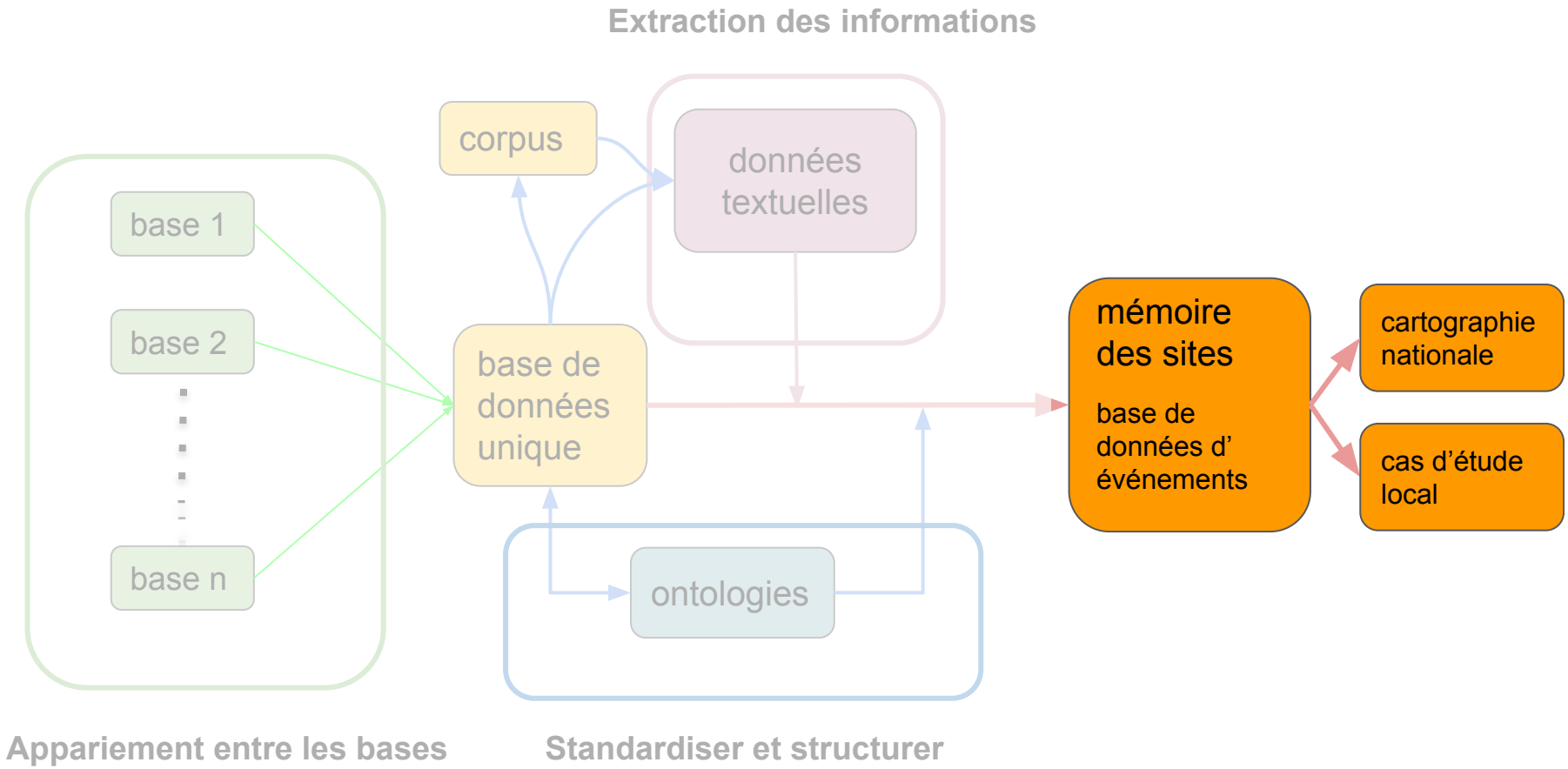
Résultats

	Rappel		Précision		F-mesure
Installation (I) :	0.7684	146 / 190	0.9480	146 / 154	0.8488
Polluant (U) :	0.4655	54 / 116	0.7500	54 / 72	0.5744
Substance (S) :	0.8189	95 / 116	0.8962	95 / 106	0.8558
Dépôt (D) :	0.1538	2/13	0.2500	2/8	0.1904
Total :	**0.6827**	**297 / 435**	**0.8735**	**297 / 340**	**0.7664**

Futur travail

- Améliorer l'extracteur des informations
  - ajouter les traits morpho-syntaxiques et textométriques en entrée de l'entraînement
  - ajuster le learning rate et le nombre d'"époques" d'entraînement pour obtenir un meilleur résultat
- Augmenter la taille de corpus annoté pour l'entraînement par méthode de bootstrapping
- Entraîner un classifieur pour catégoriser les informations extraites selon les ontologies
- Réorganiser ces données pour les intégrer dans la mémoire des sites pollués
  - lier les éléments extraits par les relations syntaxiques
  - ordonner les événements extraits par le temps pour former une frise chronologique
  - grouper les activités extraites selon le schéma d'ICPE
  - ...

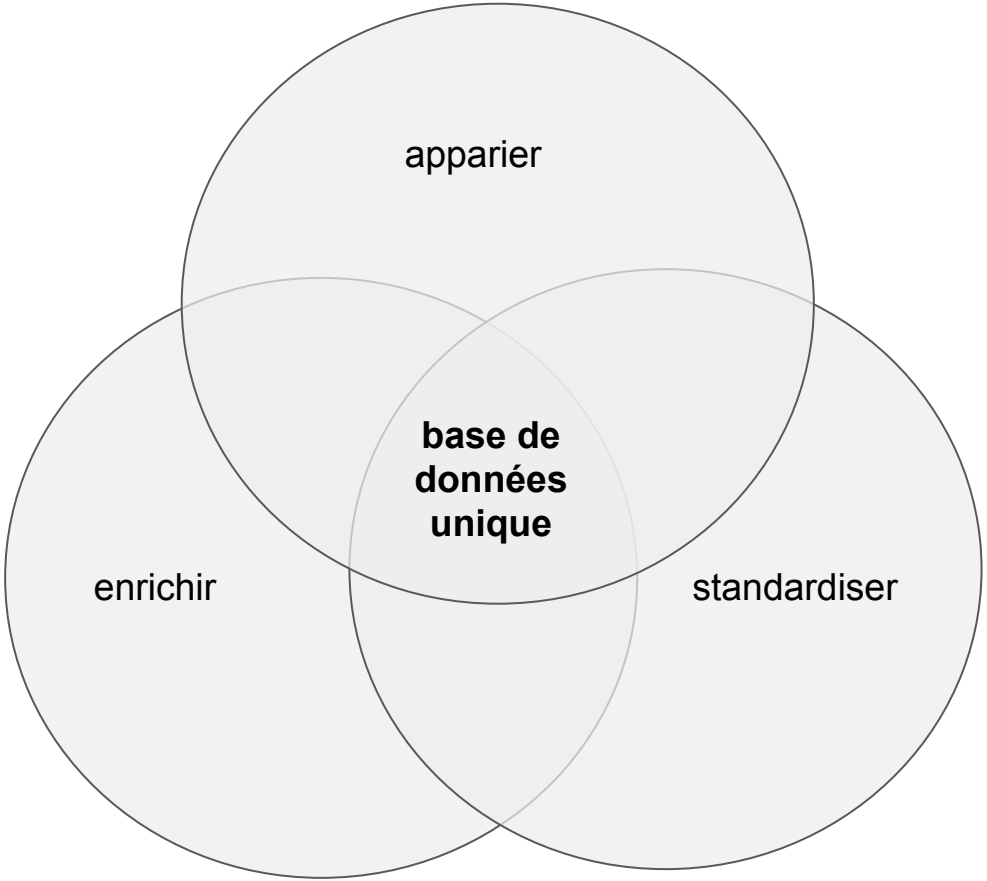
Étape 3 : valorisation des résultats



Perspectives de valorisation des résultats

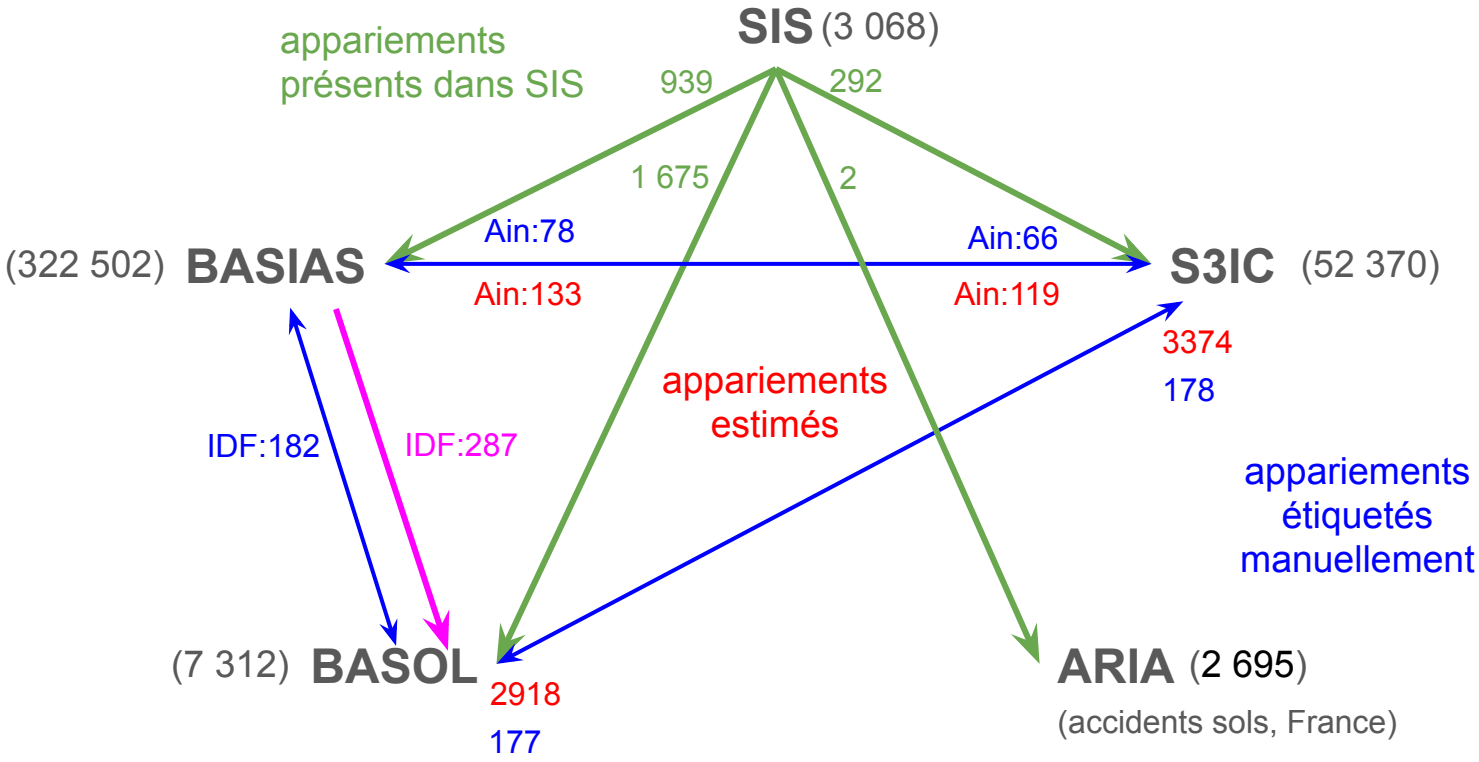
- représentation à l’échelle nationale des résultats automatiques
  - cartographie de la mémoire des sites
  - recherche avancée multicritère
- démonstrateur web de l’extracteur d’événement
  - envoi d’un texte
  - extraction des évènements
- prototype d’interrogation et de visualisation de la mémoire d’un site
  - focus sur une zone plus précise
  - vérification manuelle des données extraites automatiquement

Merci pour votre attention !



📧 chuanming.dong@ign.fr

Appariements



Communications et posters

- DOING – MADICS 2020 : Alignement de bases de données pour l'extraction d'informations concernant les sols pollués
- AGEE - MADICS 2020 : Fusion entre bases de données hétérogènes concernant la pollution des sols
- Journée de Recherche IGN 2020 : Construction d'une mémoire des sites pollués (Poster)
- Journée de Recherche IGN 2021 : Construction d'une mémoire des sites pollués (Poster)