Inférence semi-automatique et interactive de règles avec ou sans vérité terrain pour la reconnaissance de structure de documents

Cérès Carton
Encadrée par Bertrand Coüasnon et Aurélie Lemaitre

23 Mars 2016

UMR IRISA

intuidoc

INSA RENNES

Analyse d'images de documents

Adaptation rapide d'un système de reconnaissance de structure de documents

- Pour le concepteur
- En temps et en effort

Exploitation efficace documents numérisés

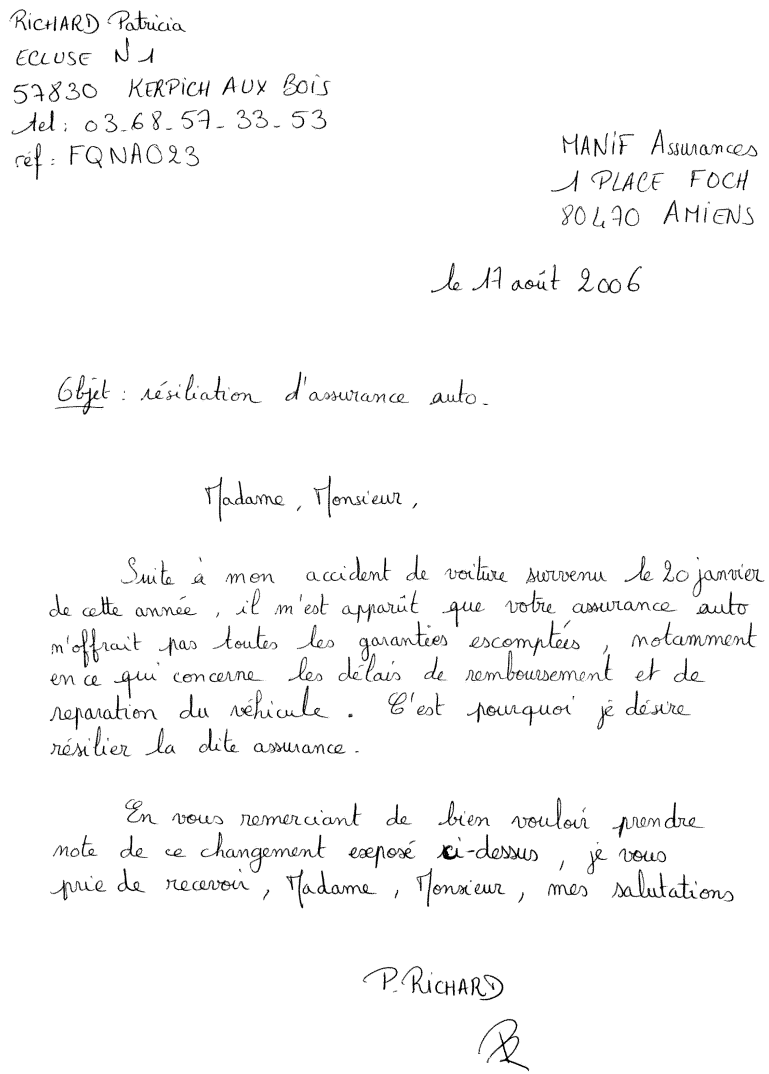
$\Rightarrow$ connaissance de la structure

- Indexation
- Navigation
- Classification
- Reconnaissance de caractères
- ...

Analyse de la structure de documents

Structure logique d'un document

- Structure physique
- Étiquetage logique
- Organisation hiérarchique



RICHARD Patricia
ECLUSE N 1
57830 KERPICH AUX BOIS
tel : 03-68-57-33-53
réf : FQNA023

MANIF Assurances
1 PLACE FOCH
80470 AMIENS

le 11 août 2006

Objet : résiliation d'assurance auto.

Madame, Monsieur,

Suite à mon accident de voiture survenu le 20 janvier de cette année, il m'est apparût que votre assurance auto n'offrait pas toutes les garanties escomptées, notamment en ce qui concerne les délais de remboursement et de reparation du véhicule. C'est pourquoi je désire résilier la dite assurance.

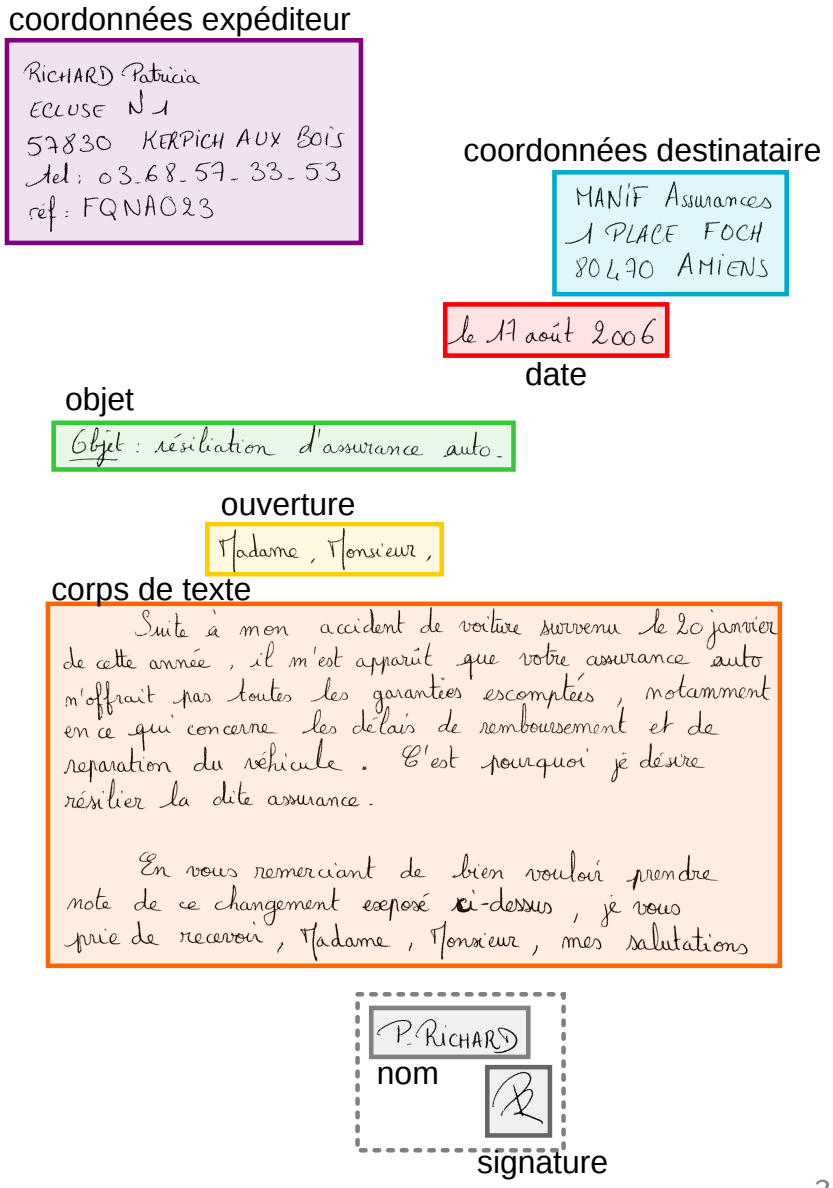
En vous remerciant de bien vouloir prendre note de ce changement exposé ci-dessus, je vous prie de recevoir, Madame, Monsieur, mes salutations

P. RICHARD

Analyse de la structure de documents

Structure logique d'un document

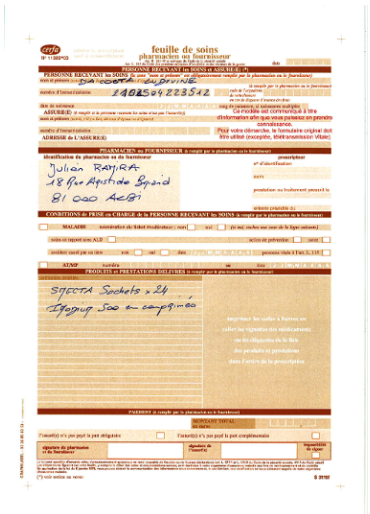
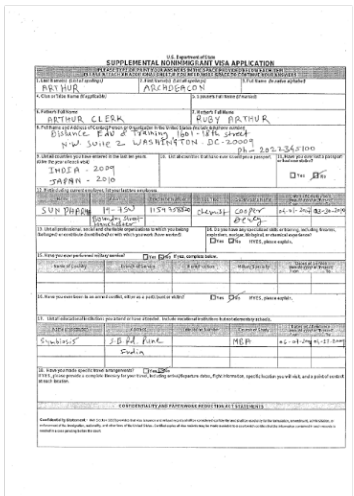
- Structure physique
- Étiquetage logique
- Organisation hiérarchique



Limites des systèmes actuels

Difficulté croissante de la reconnaissance de structure

- Documents variés
- Documents complexes

Adaptation d'un système de reconnaissance

$\Rightarrow$ modélisation de connaissances complexes

- Constitution d'un corpus étiqueté bien sélectionné
  - Très coûteux en temps
- Expression manuelle de connaissances
  - Bonne vue sur les données et leur variabilité

Illustration 1 : corpus avec vérité terrain

- Courriers manuscrits en français
- Corpus de la compétition internationale RIMES
- Tâche de reconnaissance de structures
  - Localisation de 8 types de blocs
- Vérité terrain au niveau bloc
  - Boîtes englobantes
  - Transcription

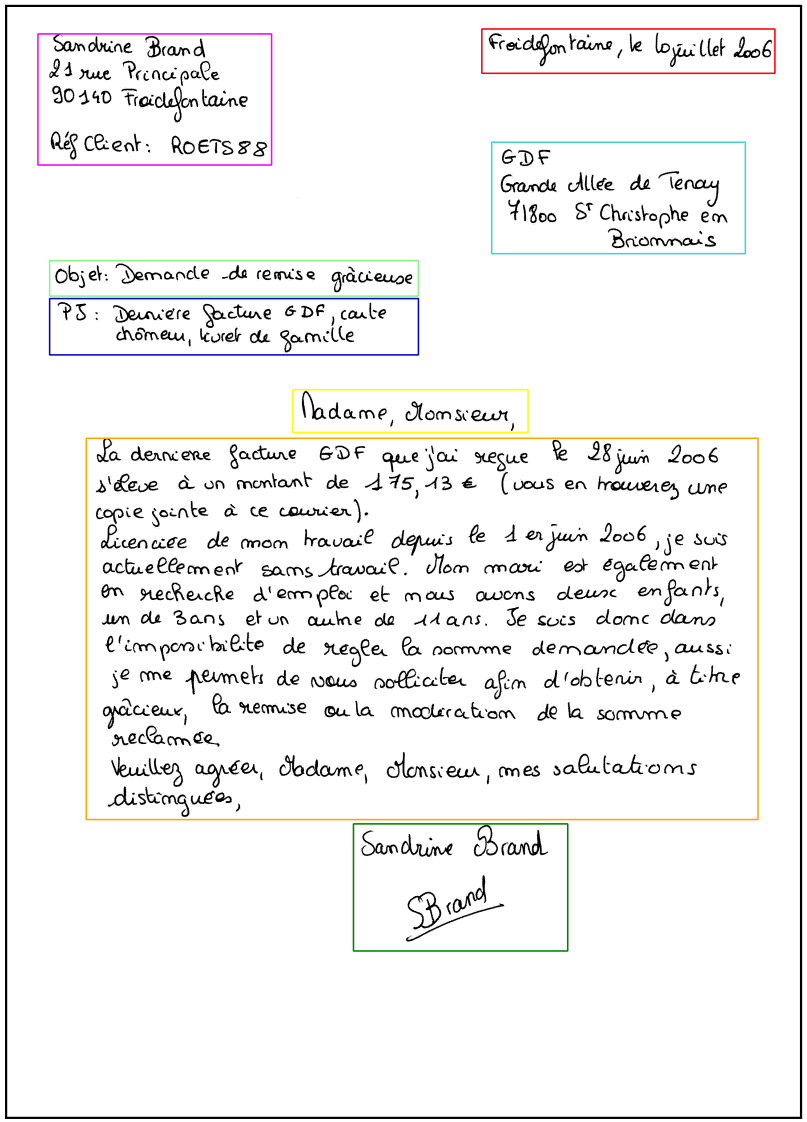
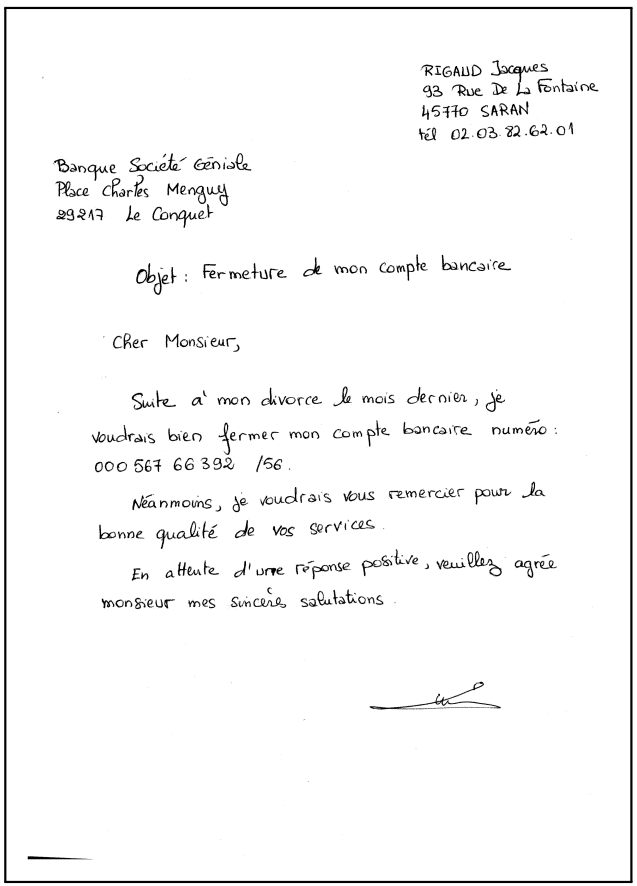


Illustration 1 : corpus avec vérité terrain

Difficultés

- Segmentation de la page en blocs
- Règles usuelles non respectées



RIGAUD Jacques
93 Rue De La Fontaine
45770 SARAN
tél 02.03.82.62.01

Banque Société Générale
Place Charles Mengus
29217 Le Conquet

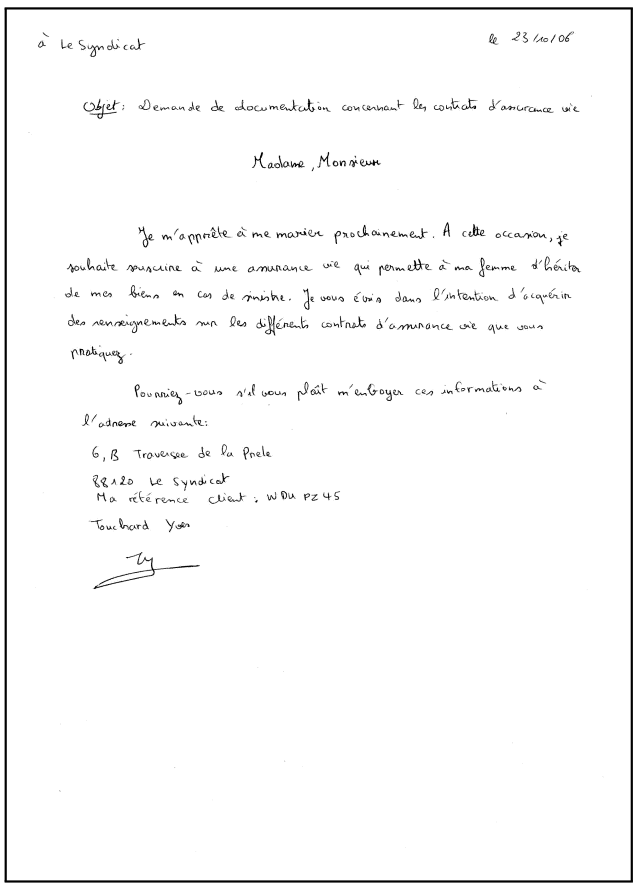
Objet : Fermeture de mon compte bancaire

Cher Monsieur,

Suite à mon divorce le mois dernier, je voudrais bien fermer mon compte bancaire numéro : 000 567 66 392 /56.

Néanmoins, je voudrais vous remercier pour la bonne qualité de vos services.

En attente d'une réponse positive, veuillez agréer monsieur mes sincères salutations.



à Le Syndicat

le 23/10/06

Objet : Demande de documentation concernant les contrats d'assurance vie

Madame, Monsieur

Je m'apprête à me marier prochainement. A cette occasion, je souhaite souscrire à une assurance vie qui permette à ma femme d'hériter de mes biens en cas de malheur. Je vous écris dans l'intention d'acquérir des renseignements sur les différents contrats d'assurance vie que vous pratiquez.

Pourriez-vous s'il vous plaît m'envoyer ces informations à l'adresse suivante:

6, B Traversée de la Prele

88120 Le Syndicat
Ma référence client : WDU PZ45

Touchard Yves

Illustration 2 : corpus sans vérité terrain

- Registres de mariages mexicains
  - 30 000 documents
- Compétition organisée par Family Search à HIP 2013
- Localisation des champs manuscrits mois et année dans un acte pré-imprimé

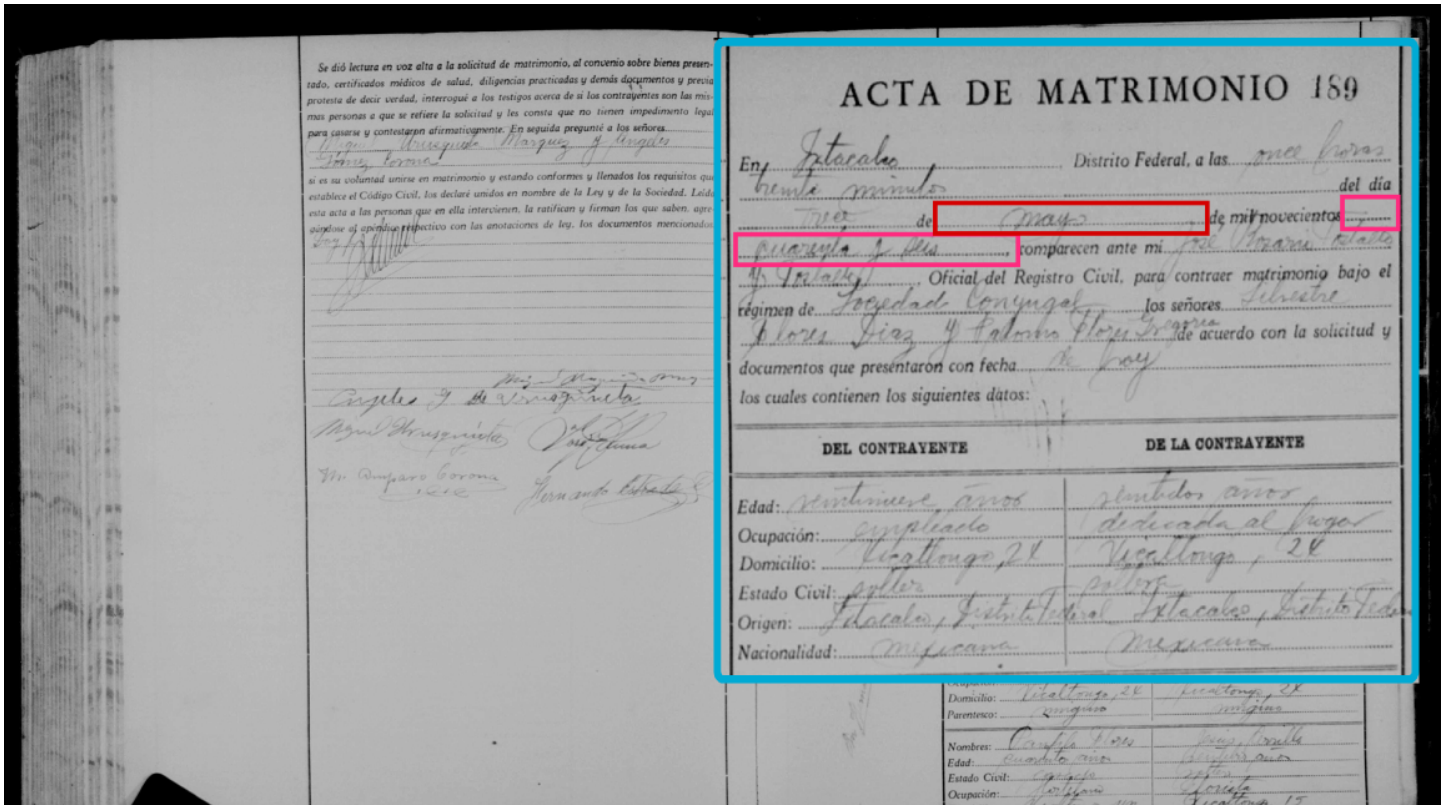
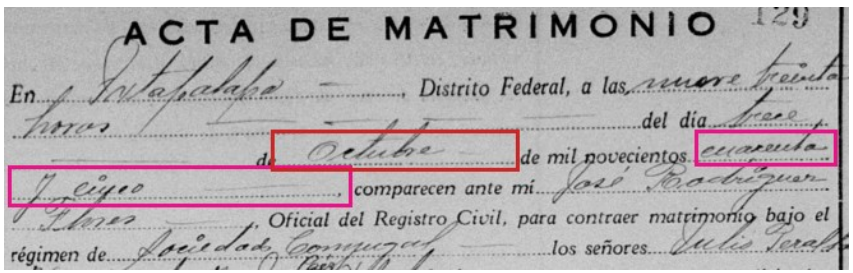
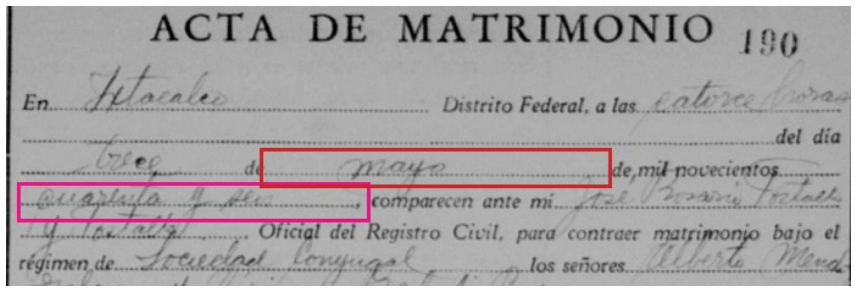
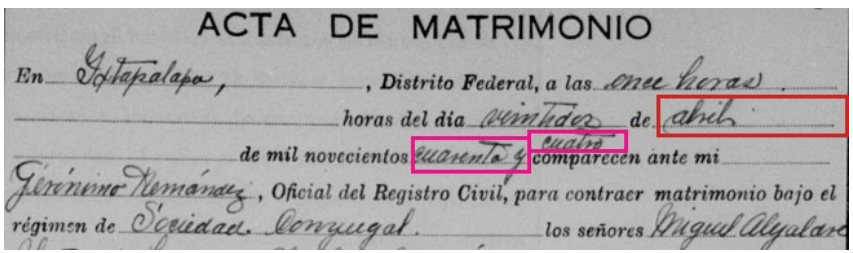


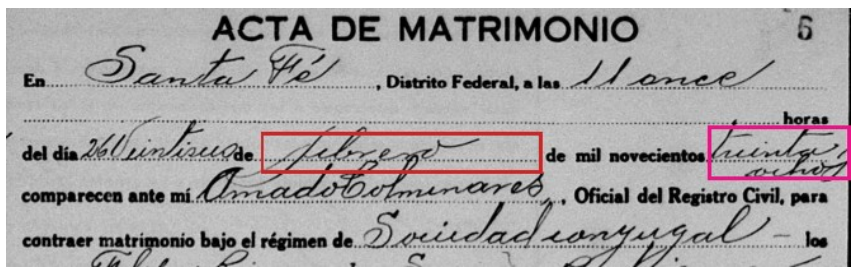
Illustration 2 : corpus sans vérité terrain

- Difficulté : mise en page variable des pré-imprimés
  - Position des champs varie
- Combien de variantes ?









- Comment détecter les variantes ?
  - Regarder un par un les 30 000 documents ?

Méthodes statistiques

- Apprentissage automatique de l'étiquetage par un modèle statistique
  - Niveau bloc
  - Niveau pixel
- À partir d'un échantillon étiqueté

Différents modèles possibles

- Champs aléatoire de Markov [Lemaitre, 2007]
- Réseaux bayésiens [Le Bourgeois, 2001]
- Champs aléatoires conditionnels [Montreuil, 2010] [Chaudhury, 2009]
- Réseaux de neurones [Rangoni, 2008]

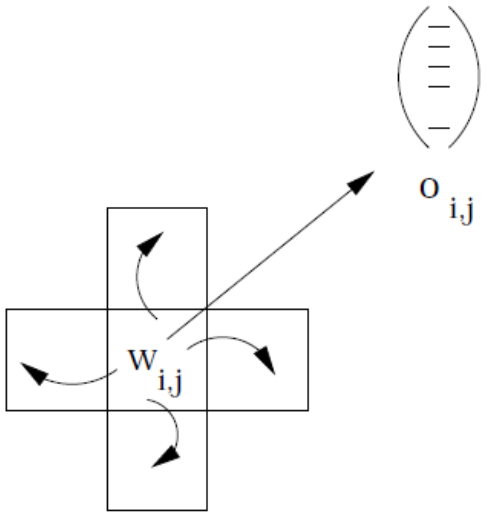
Nouveau type de documents $\Rightarrow$ base d'apprentissage étiquetée

- Coût de production
- Sélection des données

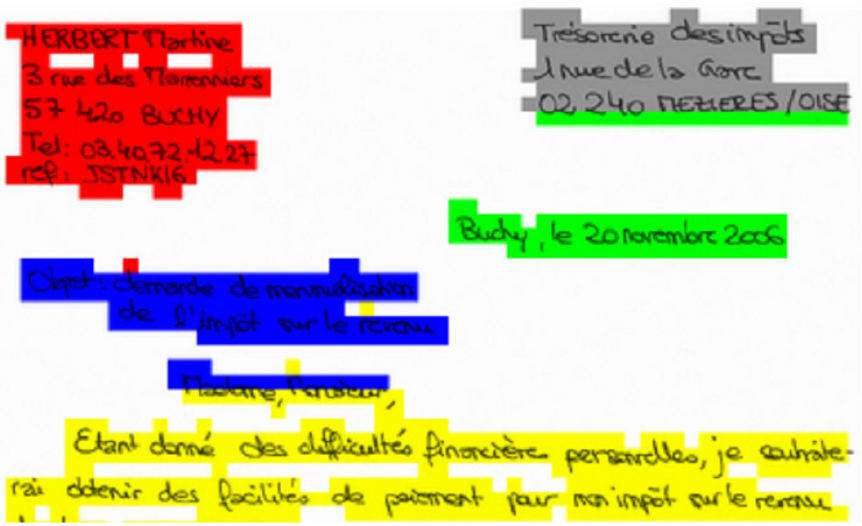
Méthodes statistiques - exemple

Apprentissage du modèle : pour chaque pixel

- Étiquette
- Caractéristiques
  - Position
  - Texture
- Voisinage



Résultat obtenu



Méthodes syntaxiques

Explicitations des connaissances par un expert dans un modèle

- À base de règles [Fisher, 1991]
  - Difficilement adaptables
  - Peu flexibles
- À base de grammaires [Conway, 1993] [Krishnmoorthy, 1993]
  - Adapté aux structures hiérarchiques
  - Bidimensionnelles [Coüasnon, 2006]
  - Stochastiques [Tateisi, 94] [Maroneze, 2011]

Nouveau type de documents

$\Rightarrow$ description manuelle de nouvelles règles

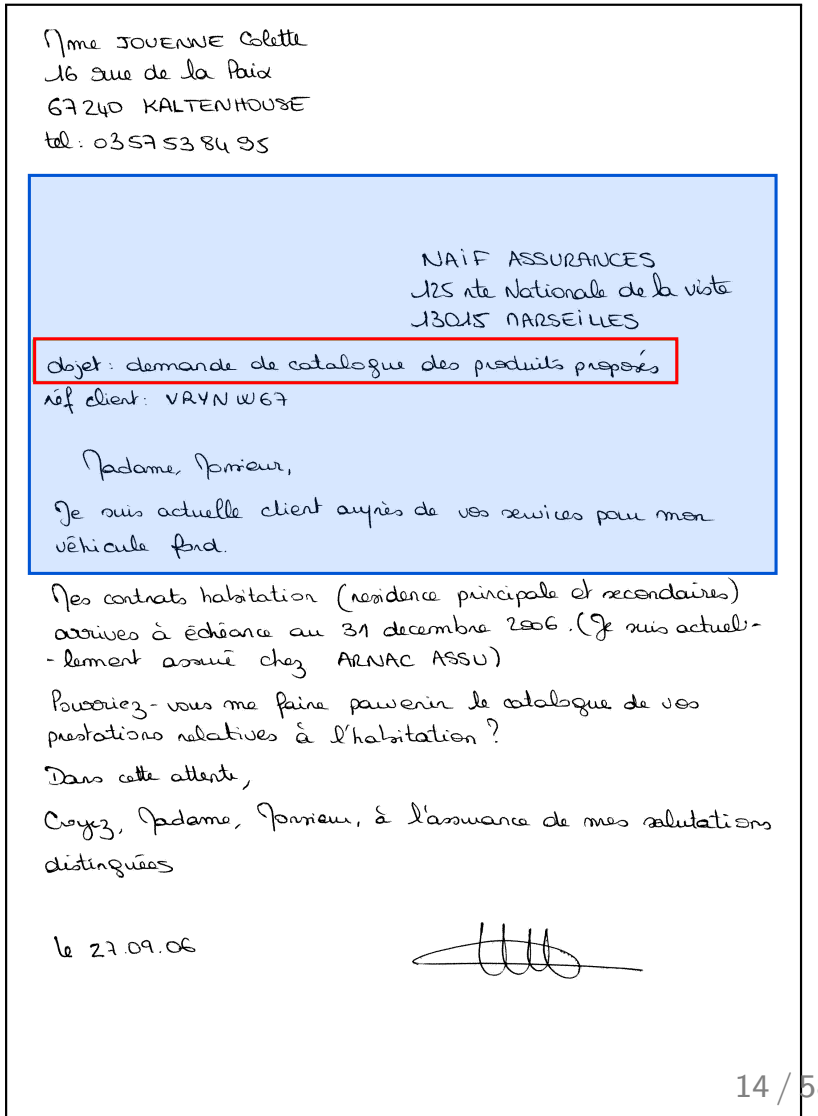
- Échantillon de taille limitée
- Choix des exemples

Méthodes syntaxiques - illustration

Description des connaissances syntaxiques

objetCourrier ::=
AT(milieuGauche) &&
ligneLongue L &&
AT(sousLigne L) &&
ligneDecalee.

objetCourrier ::=
AT(milieuGauche) &&
ligneLongue.


Mme JOUENNE Colette
16 rue de la Paix
67240 KALTENHOUSE
tel : 0357538495

NAIF ASSURANCES
125 rte Nationale de la Viste
13015 MARSEILLES

objet : demande de catalogue des produits proposés
réf client : VRYNW67

Madame, Monsieur,

Je suis actuelle client auprès de vos services pour mon véhicule Ford.

Mes contrats habitation (residence principale et secondaires) arrivés à échéance au 31 decembre 2006. (Je suis actuellement assuré chez ARNAC ASSU)

Pourriez-vous me faire parvenir le catalogue de vos prestations relatives à l'habitation ?

Dans cette attente,

Croyez, Madame, Monsieur, à l'assurance de mes salutations distinguées

le 27.09.06

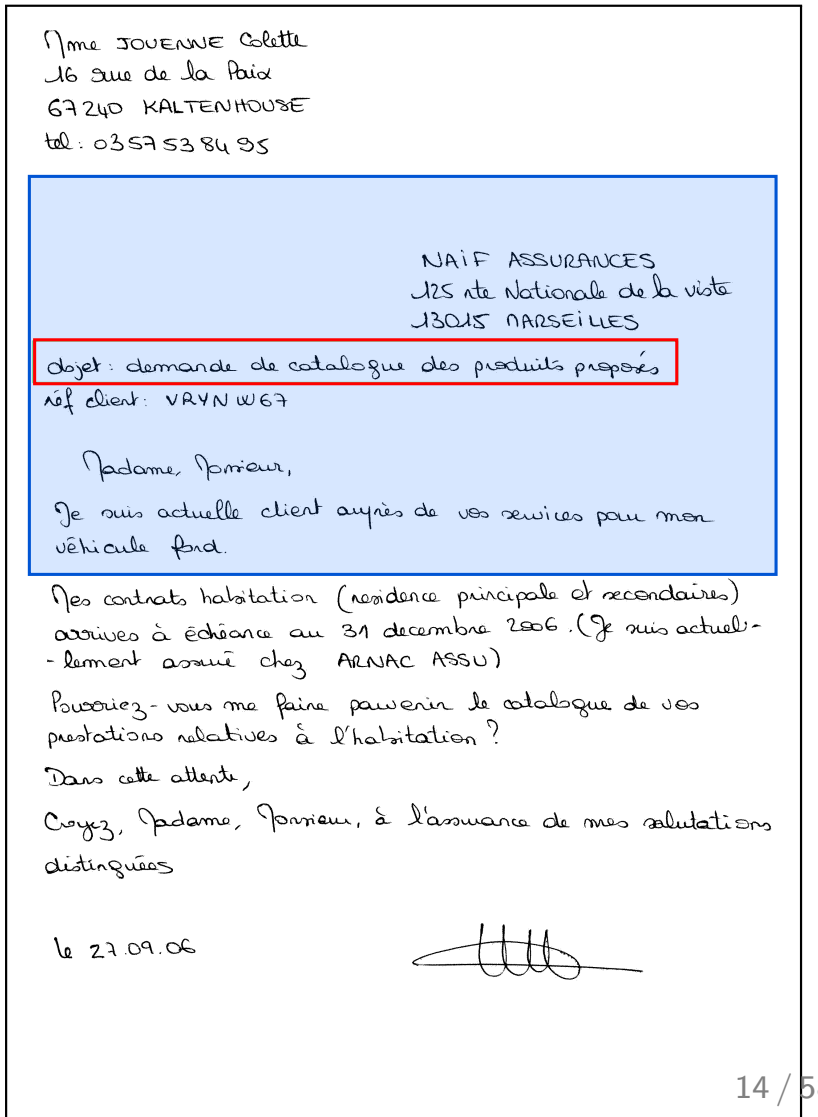
→ Type de règles à inférer

Méthodes syntaxiques - illustration

Description des connaissances syntaxiques

```
objetCourrier ::=
    AT(milieuGauche) &&
    ligneLongue L &&
    AT(sousLigne L) &&
    ligneDecalee.

objetCourrier ::=
    AT(milieuGauche) &&
    ligneLongue.
```

variantes logiques


Mme JOUENNE Colette
16 rue de la Paix
67240 KALTENHOUSE
tel : 0357538495

NAIF ASSURANCES
125 rte Nationale de la Viste
13015 MARSEILLES

objet : demande de catalogue des produits proposés
réf client : VRYNW67

Madame, Monsieur,

Je suis actuelle client auprès de vos services pour mon véhicule Ford.

Mes contrats habitation (residence principale et secondaires) arrivent à échéance au 31 decembre 2006. (Je suis actuellement assuré chez ARNAC ASSU)

Pourriez-vous me faire parvenir le catalogue de vos prestations relatives à l'habitation ?

Dans cette attente,

Croyez, Madame, Monsieur, à l'assurance de mes salutations distinguées

le 27.09.06

→ Type de règles à inférer

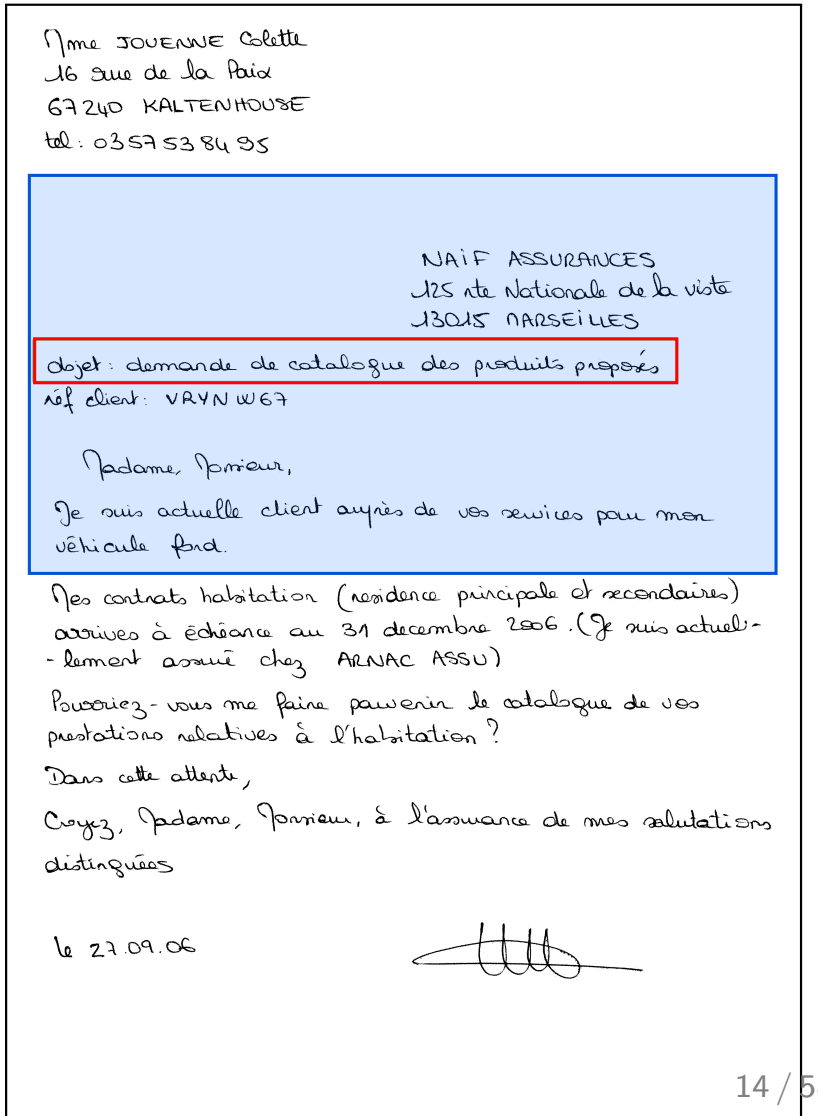
Méthodes syntaxiques - illustration

Description des connaissances syntaxiques

```
objetCourrier ::=
    AT(milieuGauche) &&
    ligneLongue L &&
    AT(sousLigne L) &&
    ligneDecalee.

objetCourrier ::=
    AT(milieuGauche) &&
    ligneLongue.

ligneLongue L :-
    segment L,
    largeur > 1000px.

milieuGauche :-
    Xd = 49% ImageWidth,
    Yd = 32% ImageHeight,
    X1 = 0,
    Y1 = 15% ImageHeight,
    X2 = 98% ImageWidth,
    Y2 = 50% ImageHeight.

...
```

variantes logiques

→ Type de règles à inférer


Mme JOUENNE Colette
16 rue de la Paix
67240 KALTENHOUSE
tel : 0357538495

NAIF ASSURANCES
125 rte Nationale de la Viste
13015 MARSEILLES

objet : demande de catalogue des produits proposés
réf client : VRYNW67

Madame, Monsieur,

Je suis actuelle client auprès de vos services pour mon véhicule Ford.

Mes contrats habitation (residence principale et secondaires) arrivés à échéance au 31 decembre 2006. (Je suis actuellement assuré chez ARNAC ASSU)

Pourriez-vous me faire parvenir le catalogue de vos prestations relatives à l'habitation ?

Dans cette attente,

Croyez, Madame, Monsieur, à l'assurance de mes salutations distinguées

le 27.09.06

Méthodes syntaxiques - illustration

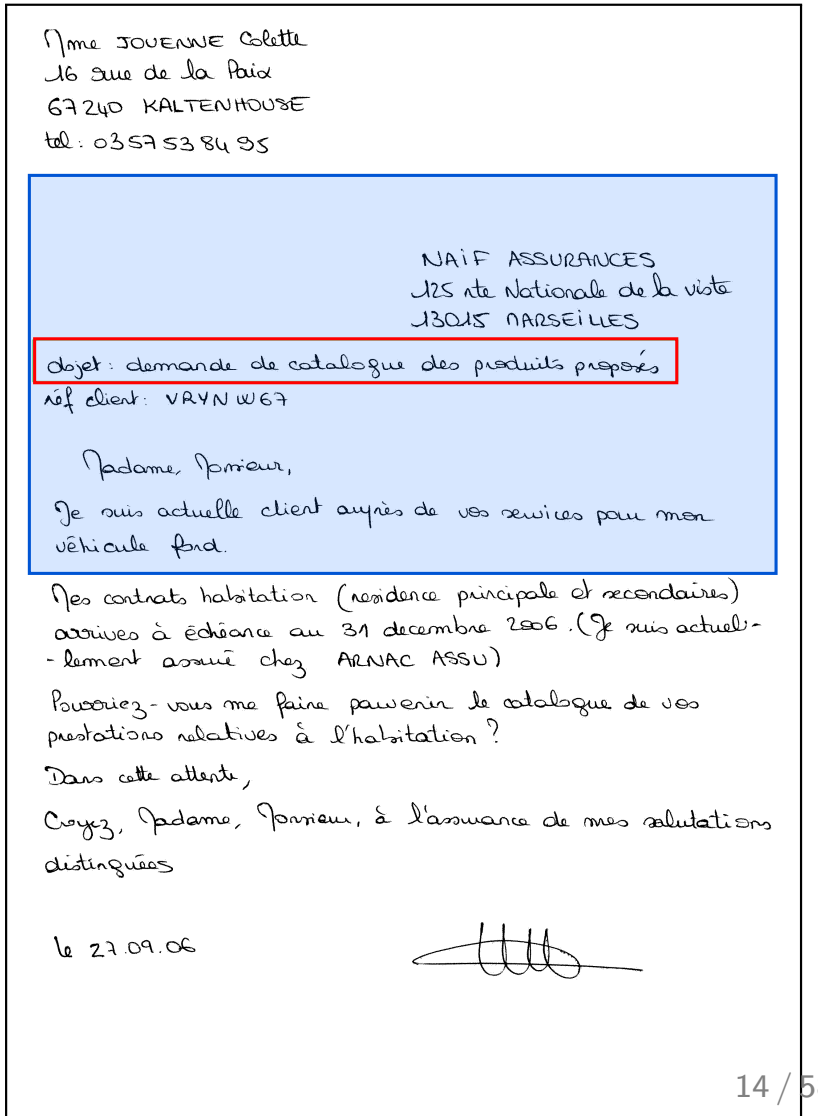
Description des connaissances syntaxiques

```
objetCourrier ::=
    AT(milieuGauche) &&
    ligneLongue L &&
    AT(sousLigne L) &&
    ligneDecalee.

objetCourrier ::=
    AT(milieuGauche) &&
    ligneLongue.

ligneLongue L :-
    segment L,
    largeur > 1000px.

milieuGauche :-
    Xd = 49% ImageWidth,
    Yd = 32% ImageHeight,
    X1 = 0,
    Y1 = 15% ImageHeight,
    X2 = 98% ImageWidth,
    Y2 = 50% ImageHeight.

...
```

variantes logiques

propriétés physiques

opérateur de position


Mme JOUENNE Colette
16 rue de la Paix
67240 KALTENHOUSE
tel : 03 57 53 84 95

NAIF ASSURANCES
125 rte Nationale de la Viste
13015 MARSEILLES

objet : demande de catalogue des produits proposés
réf client : VRYNW67

Madame, Monsieur,

Je suis actuelle client auprès de vos services pour mon véhicule Ford.

Mes contrats habitation (residence principale et secondaires) arrivés à échéance au 31 decembre 2006. (Je suis actuellement assuré chez ARNAC ASSU)

Pourriez-vous me faire parvenir le catalogue de vos prestations relatives à l'habitation ?

Dans cette attente,

Croyez, Madame, Monsieur, à l'assurance de mes salutations distinguées

le 27.09.06

→ Type de règles à inférer

Comparaison des méthodes

- Modélisation de connaissances complexes
- Adaptation facile et rapide à un nouveau type de documents

	Statistiques	Syntaxiques
Apprentissage	automatique	manuel
Données d'apprentissage	étiquetées	non étiquetées
		taille limité
Gestion de l'incertitude	oui	selon les méthodes
Gestion des cas rares	difficile	oui
Pouvoir d'expression	limité	élevé

Comparaison des méthodes

- Modélisation de connaissances complexes
- Adaptation facile et rapide à un nouveau type de documents

	Statistiques	Syntaxiques
Apprentissage	automatique	manuel
Données d'apprentissage	étiquetées	non étiquetées
		taille limité
Gestion de l'incertitude	oui	selon les méthodes
Gestion des cas rares	difficile	oui
Pouvoir d'expression	limité	élevé

Objectif : combiner les avantages

- Méthodes syntaxiques
  - Pouvoir d'expression et gestion des cas rares
- Méthodes statistiques
  - Apprentissage automatique $\longrightarrow$ sans le coût de l'annotation

Inférence dans les systèmes syntaxiques

Inférence grammaticale

- Nombreux domaines d'application
  - Bioinformatique, linguistique, reconnaissance de la parole
- Nombreux obstacles à l'inférence grammaticale [de la Higuera, 2005]
  - Bidimensionnalité
  - Données bruitées
  - Données d'apprentissage étiquetées nécessaires
  - Pas de combinaison avec des connaissances a priori

Inférence dans les systèmes syntaxiques

Inférence grammaticale

- Nombreux domaines d'application
  - Bioinformatique, linguistique, reconnaissance de la parole
- Nombreux obstacles à l'inférence grammaticale [de la Higuera, 2005]
  - Bidimensionnalité
  - Données bruitées
  - Données d'apprentissage étiquetées nécessaires
  - Pas de combinaison avec des connaissances a priori

Inférence grammaticale en reconnaissance de documents [Shilman, 2005]

Utilisateur fournit	Inférence
grammaire de pages	caractéristiques
pages annotées	paramètres

Inférence dans les systèmes syntaxiques

Inférence grammaticale

- Nombreux domaines d'application
  - Bioinformatique, linguistique, reconnaissance de la parole
- Nombreux obstacles à l'inférence grammaticale [de la Higuera, 2005]
  - Bidimensionnalité
  - Données bruitées
  - Données d'apprentissage étiquetées nécessaires
  - Pas de combinaison avec des connaissances a priori

Inférence grammaticale en reconnaissance de documents [Shilman, 2005]

Utilisateur fournit	Inférence
grammaire de pages	caractéristiques
pages annotées	paramètres

$\Rightarrow$ Pas d'inférence grammaticale automatique possible
Utilisation de l'expertise de l'utilisateur

Méthode Eyes Wide Open

Objectifs

- Combinaison statistique/structurelle
- Apprentissage sans vérité terrain
- Inférence de règles

Proposition : méthode Eyes Wide Open (EWO)

↪ Vue exhaustive sur les données

↪ Intégration de l'utilisateur

- Minimisation des interventions

Méthode	Données	Modèle
Statistiques	Utilisateur	Automatique
Syntaxiques	∅	Utilisateur
EWO	Utilisateur	

Construction progressive d'un système

- Premier niveau de la grammaire connu
  - $\longrightarrow$ éléments à reconnaître

Construction progressive d'un système

- Premier niveau de la grammaire connu
  - → éléments à reconnaître

courrier ::=
  coordonneesExpediteur &&
  coordonneesDestinataire &&
  dateLieu &&
  objet &&
  ouverture &&
  corpsDeTexte &&
  signature &&
  ps.

Construction progressive d'un système

- Premier niveau de la grammaire connu
  - → éléments à reconnaître

```
courrier ::=
    coordonneesExpediteur &&
    coordonneesDestinataire &&
    dateLieu &&
    objet &&
    ouverture &&
    corpsDeTexte &&
    signature &&
    ps.
```

- Construction progressive des règles
  - Réutilisation d'une approche classique pour les problèmes complexes
  - Décomposition en sous-problèmes

Résolution d'un sous-problème

Cas 1

- Forte connaissance a priori
- Peu de variabilité

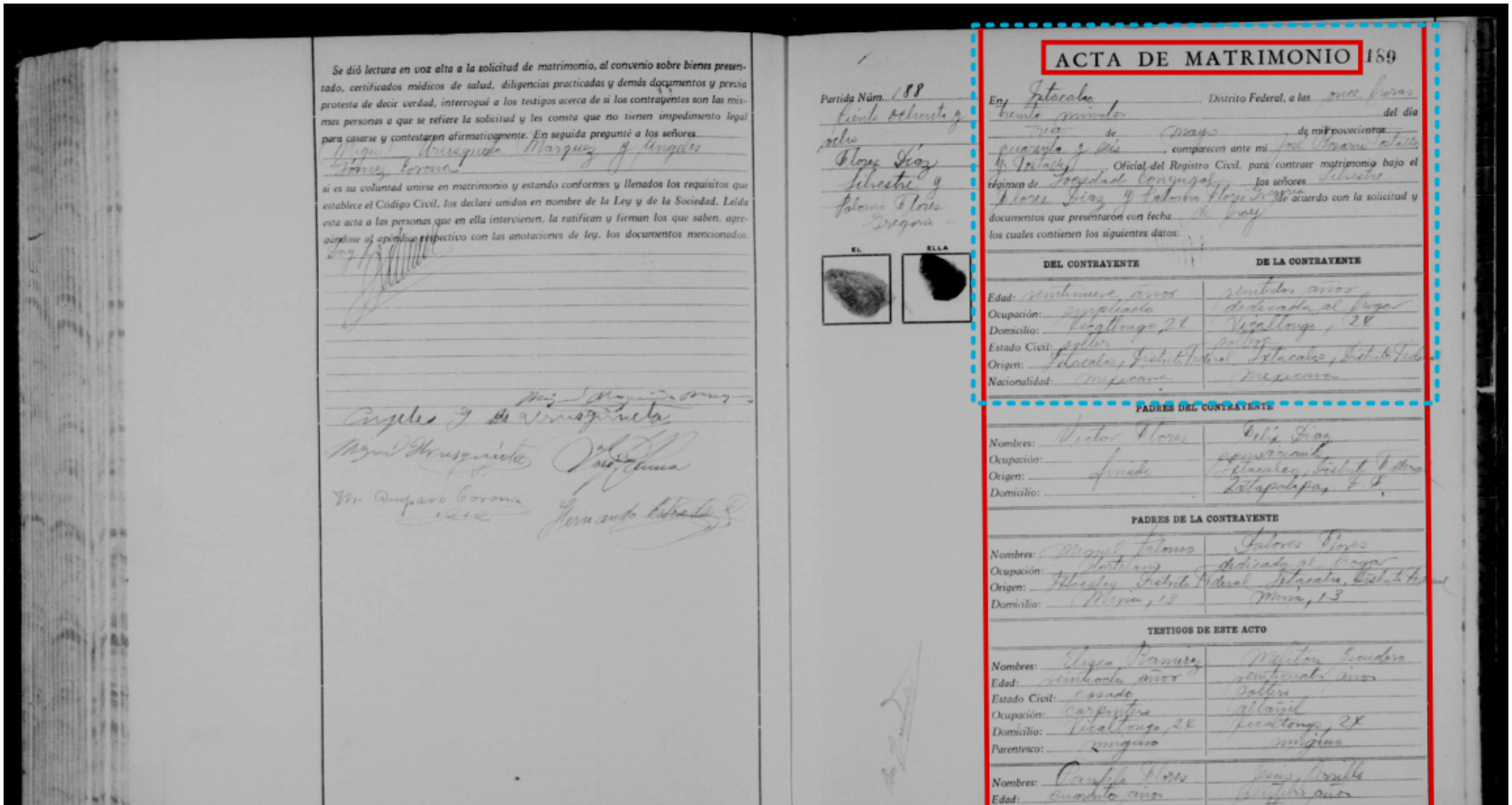
$\Rightarrow$ Description manuelle de la règle

Résolution d'un sous-problème

Cas 1

- Forte connaissance a priori
- Peu de variabilité

$\Rightarrow$ Description manuelle de la règle



Résolution d'un sous-problème

Cas 2

- Peu de connaissance a priori
- Forte variabilité

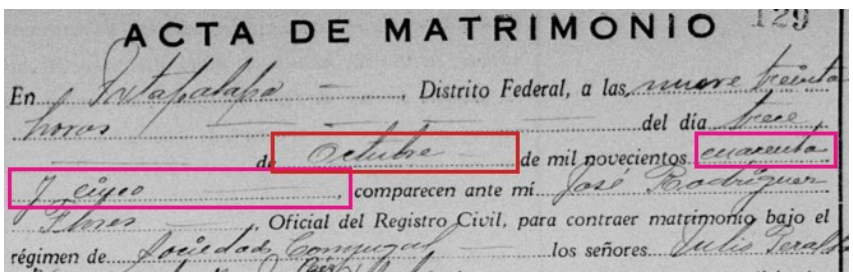
$\Rightarrow$ Utilisation de la méthode EWO
Inférence semi-automatique de la règle

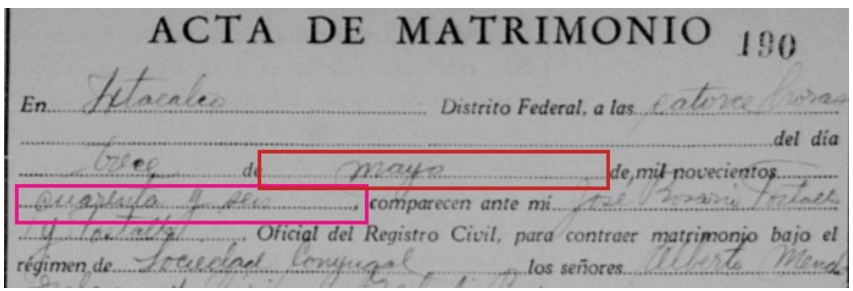
Résolution d'un sous-problème

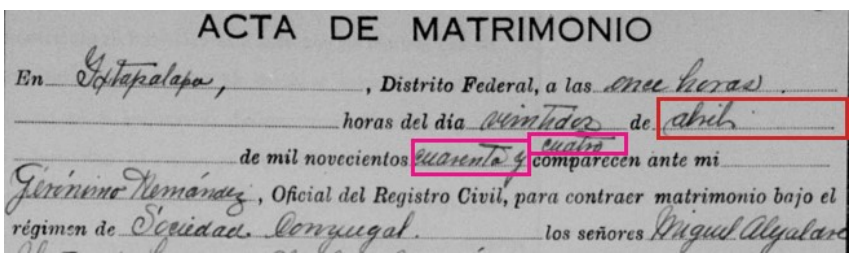
Cas 2

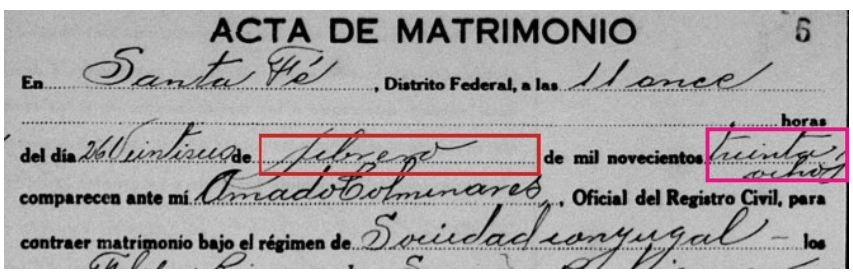
- Peu de connaissance a priori
- Forte variabilité

⇒ Utilisation de la méthode EWO
Inférence semi-automatique de la règle





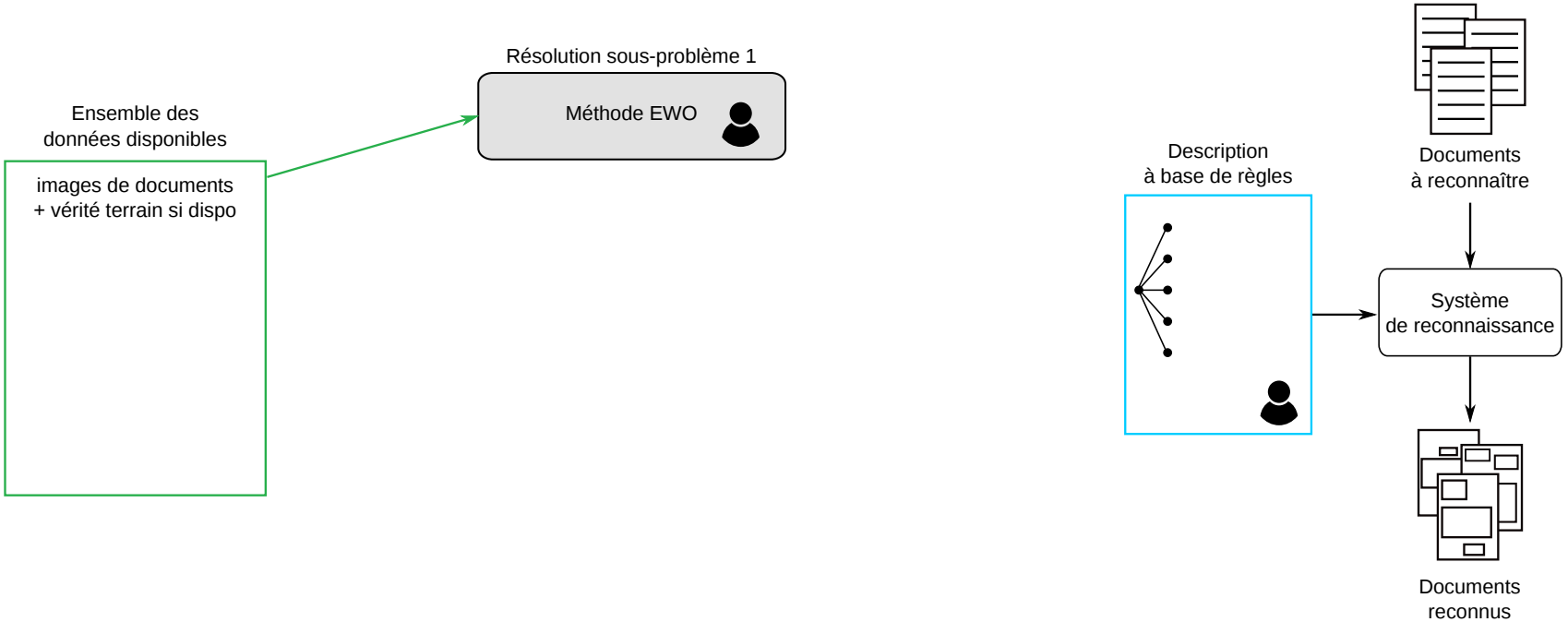




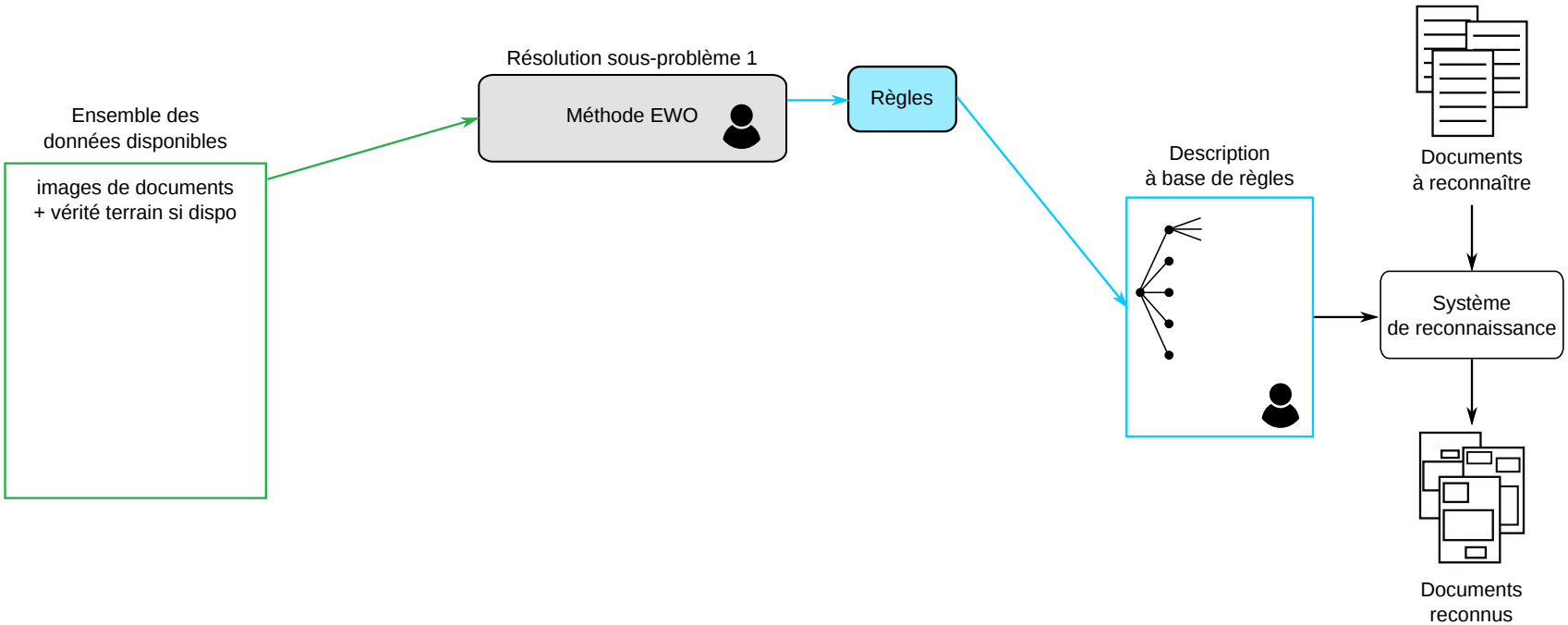
Construction progressive d'un système

Construction progressive d'un système

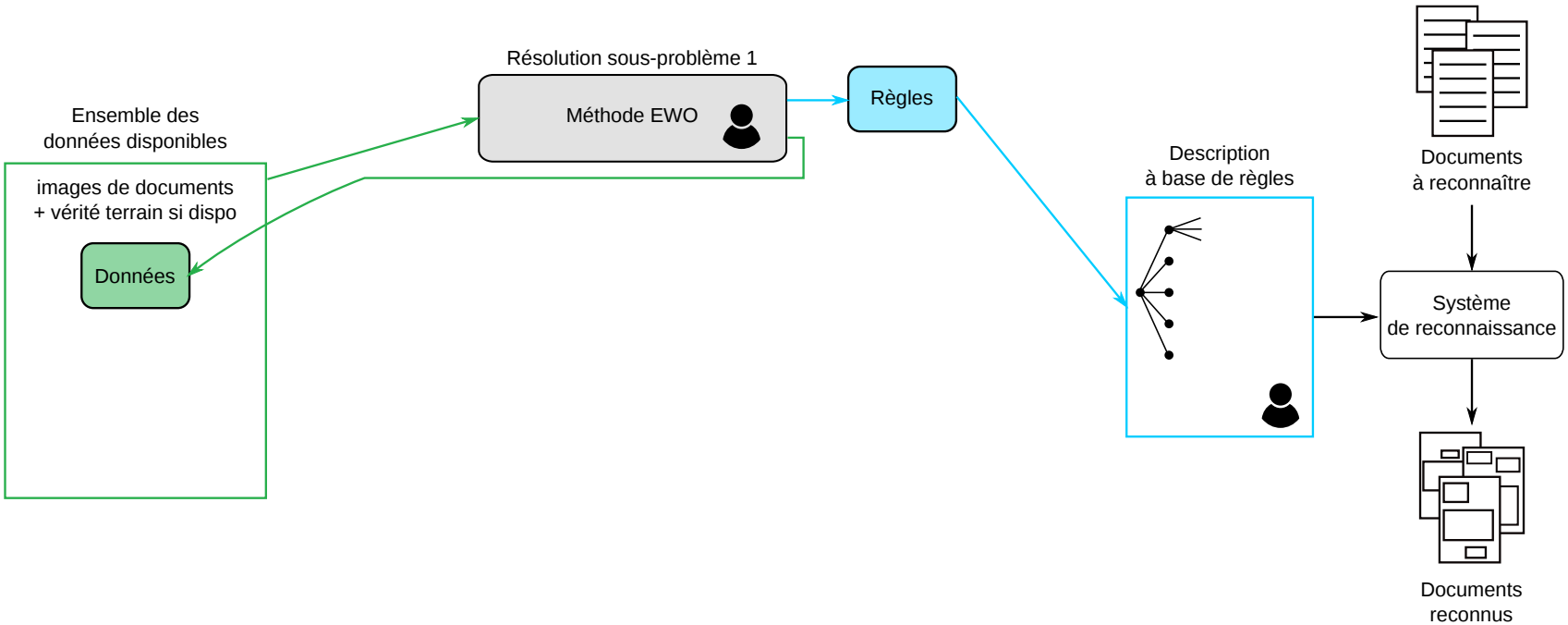
Construction progressive d'un système



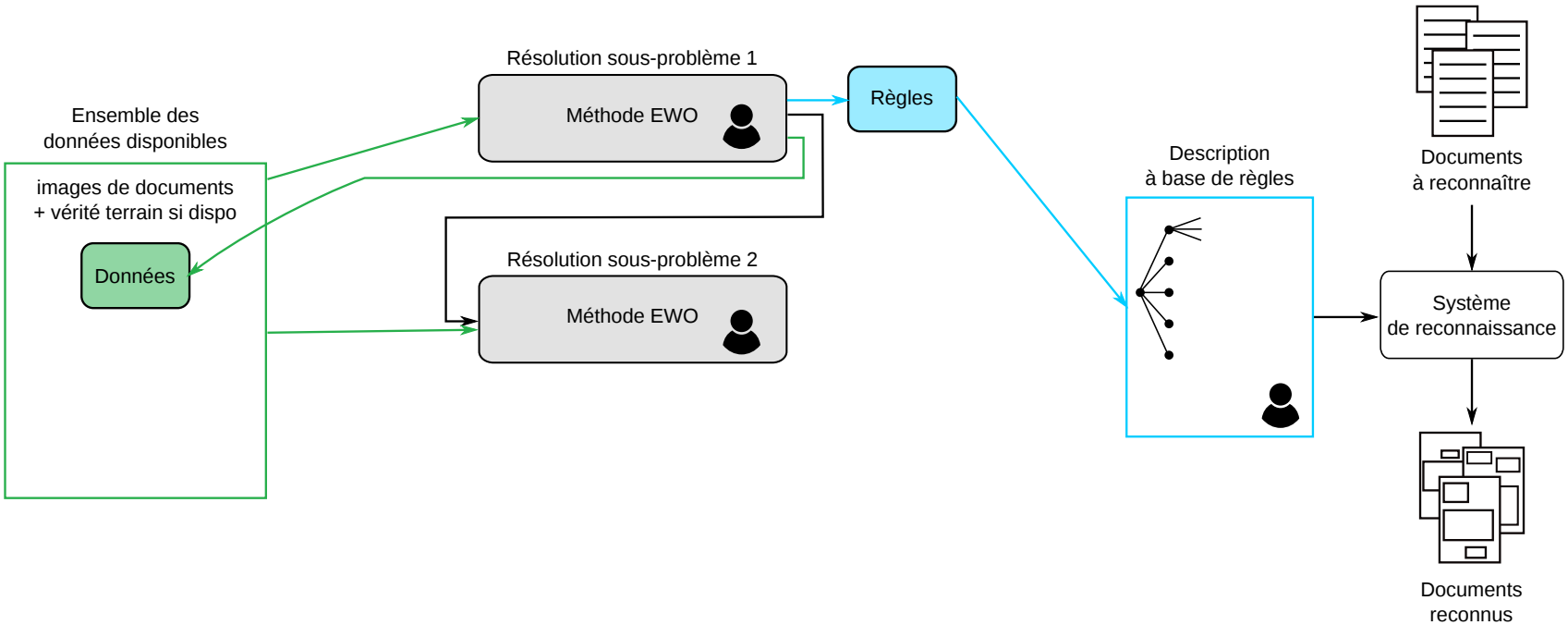
Construction progressive d'un système



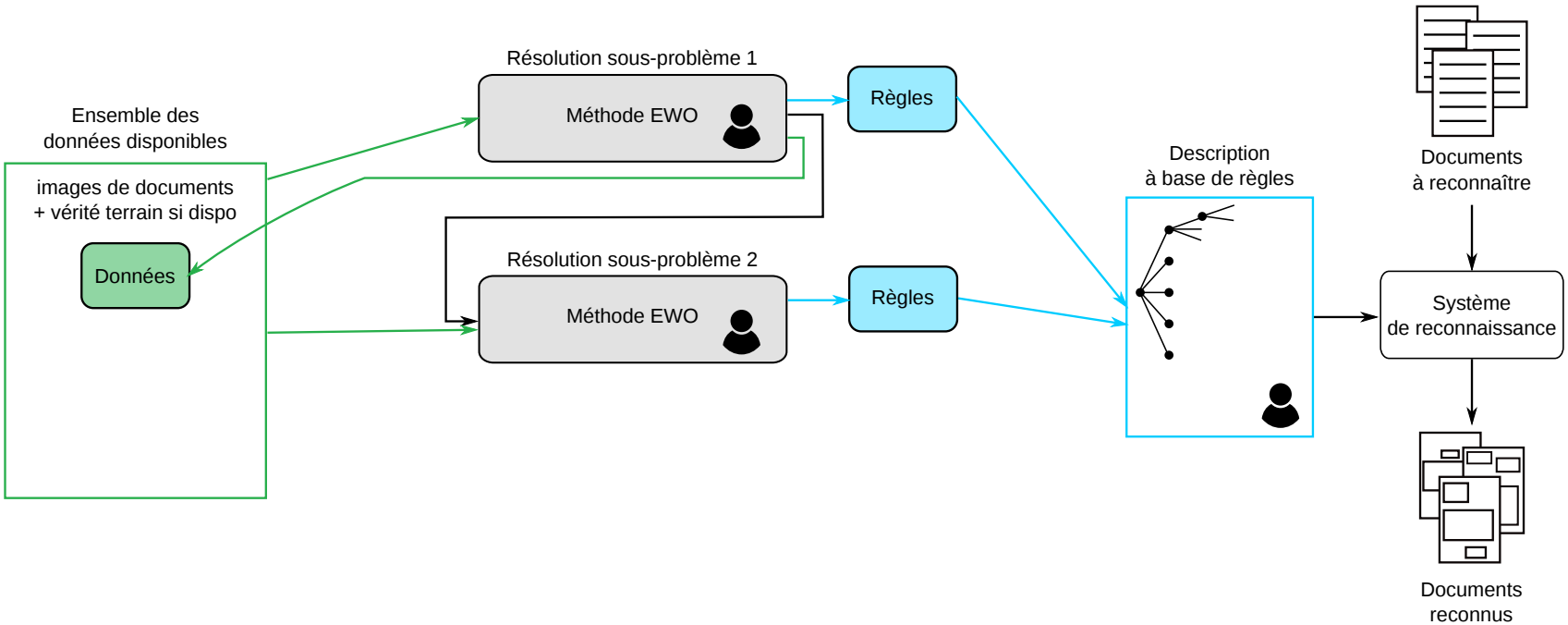
Construction progressive d'un système



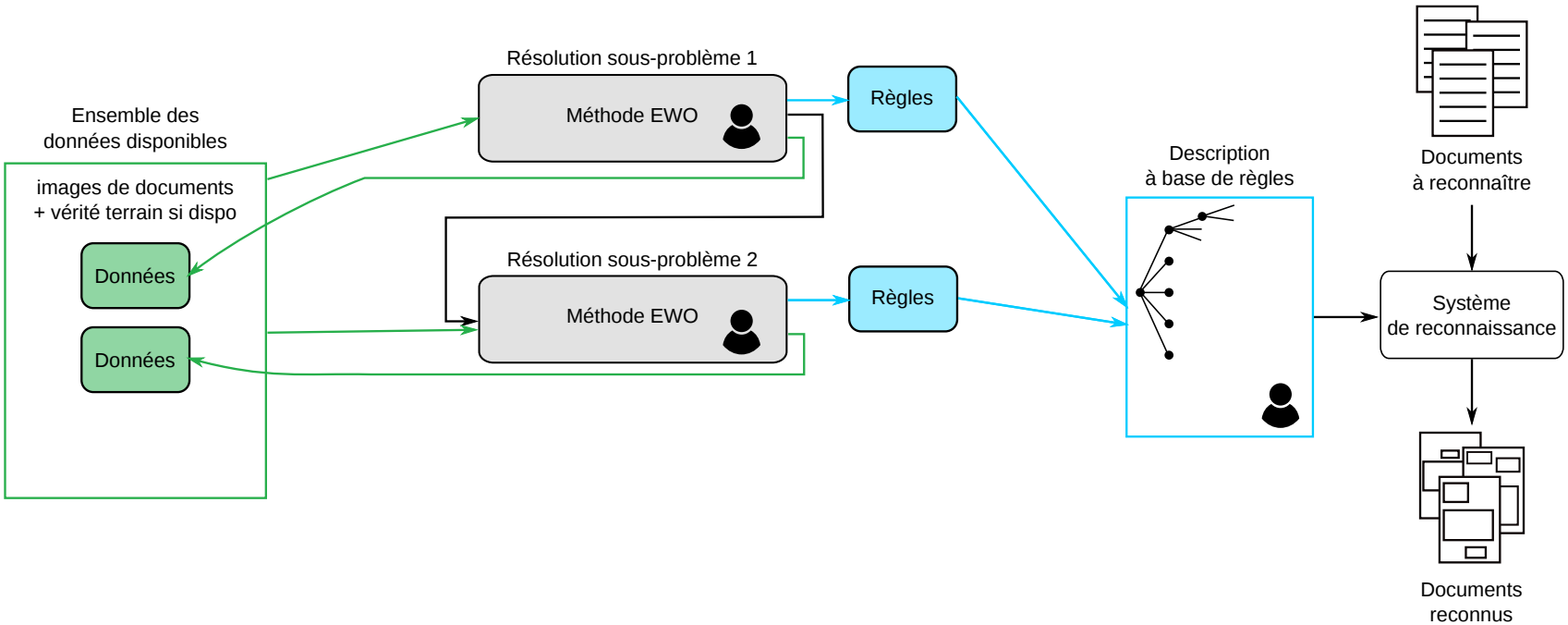
Construction progressive d'un système



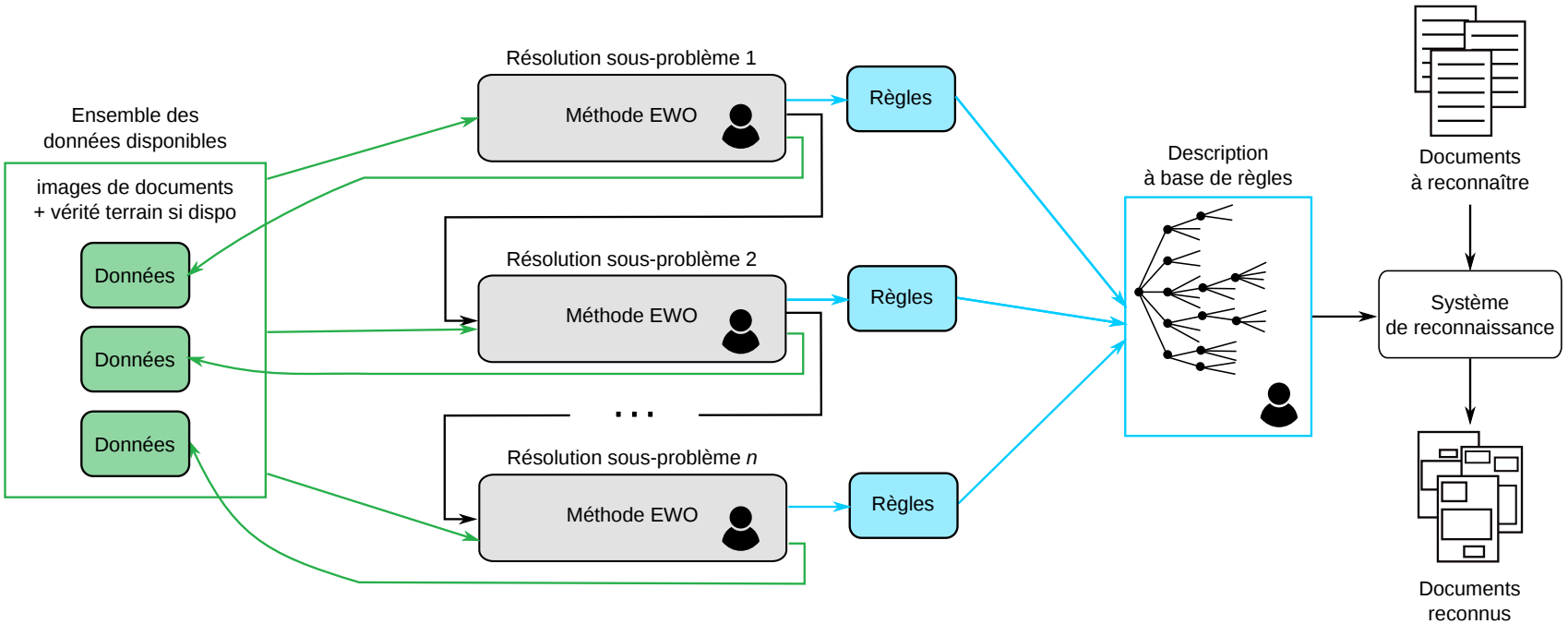
Construction progressive d'un système



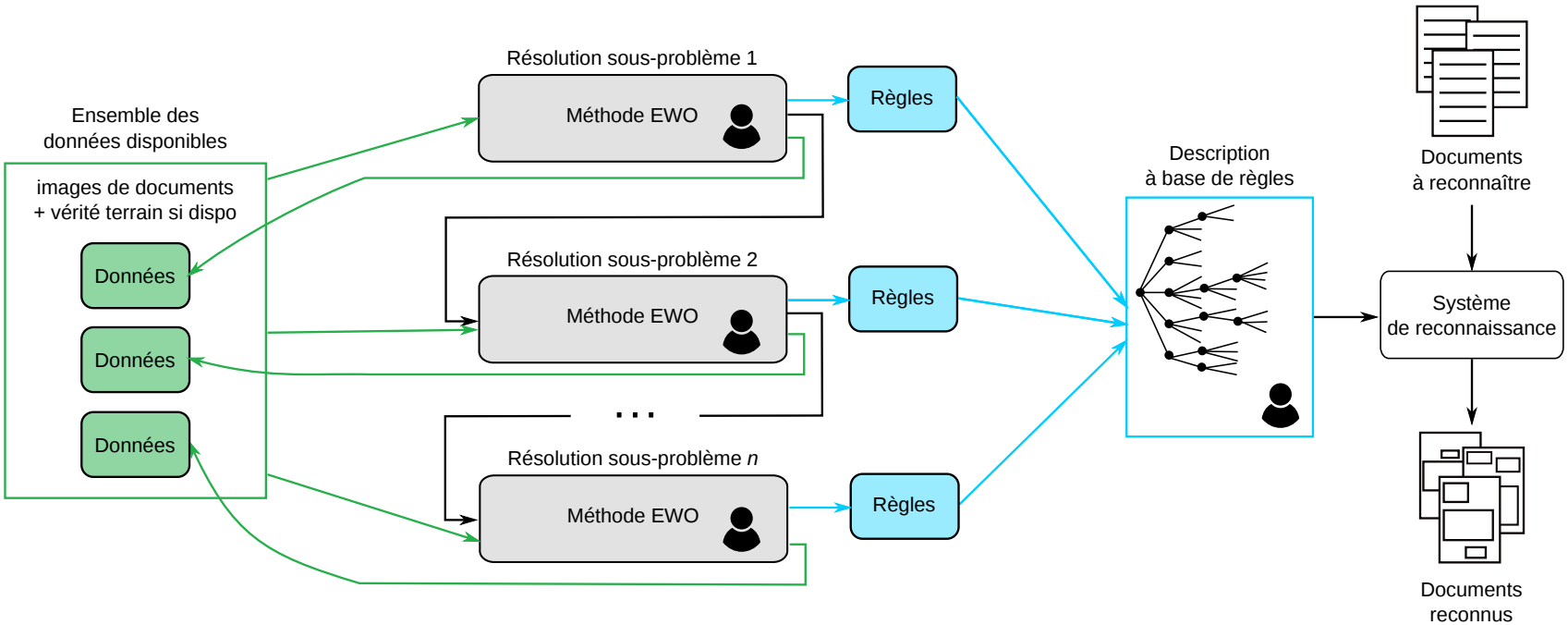
Construction progressive d'un système



Construction progressive d'un système



Construction progressive d'un système



- Enrichissement progressif
  - Description grammaticale
  - Ensemble des données disponibles
- Interaction avec l'utilisateur
  - Présentation synthétique des données
    - Prise de décision efficace
  - Intégré au cœur du système

Méthode EWO

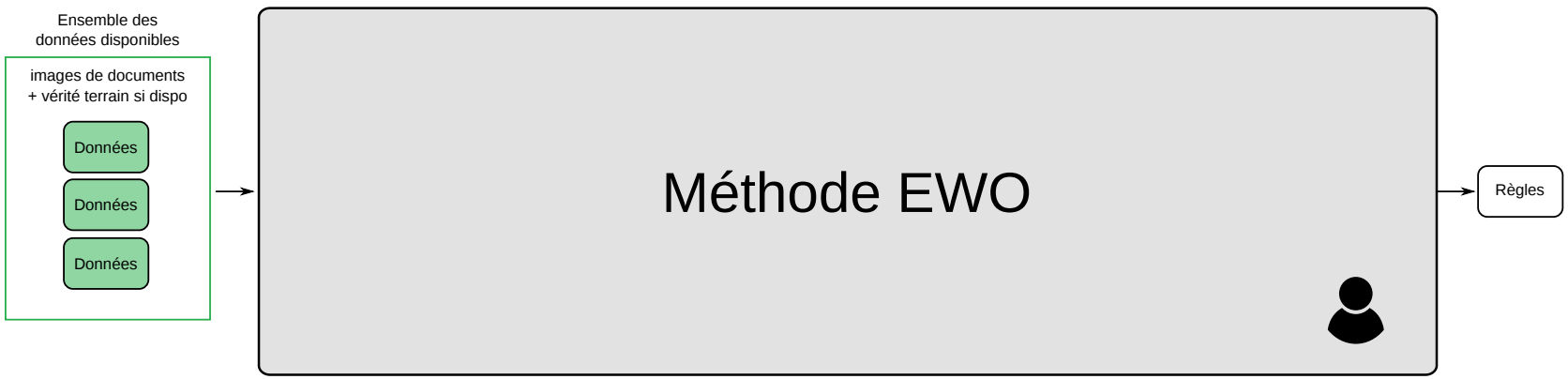
Propriétés de la méthode

- Généricité
- Utilisable sans vérité terrain
- Intégration de connaissances a priori
- Vue exhaustive sur les données

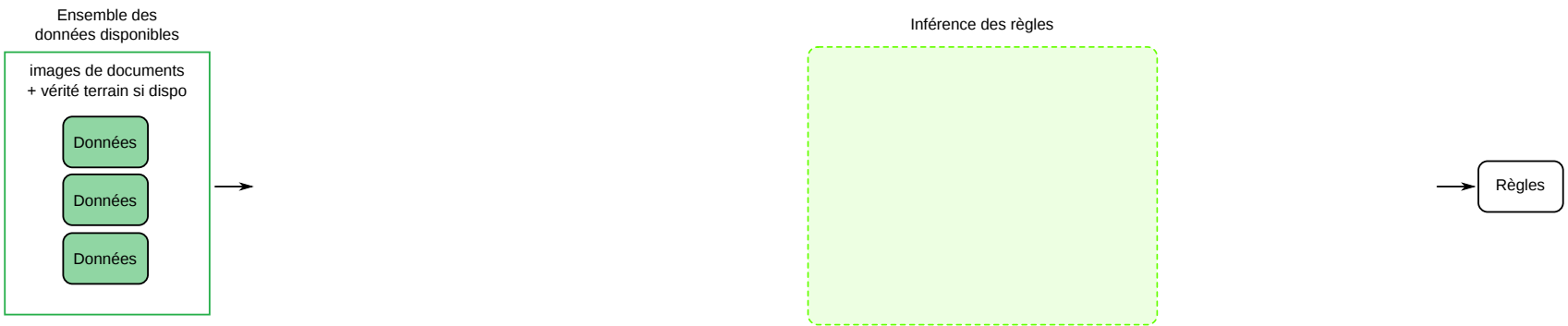
Moyens utilisés

- Clustering
  - ↪ Extraction de connaissances
  - ↪ Utilisation des redondances dans les données
  - ↪ Émergence automatique de structures
- Interaction utilisateur
  - ↪ Apport de sens aux données

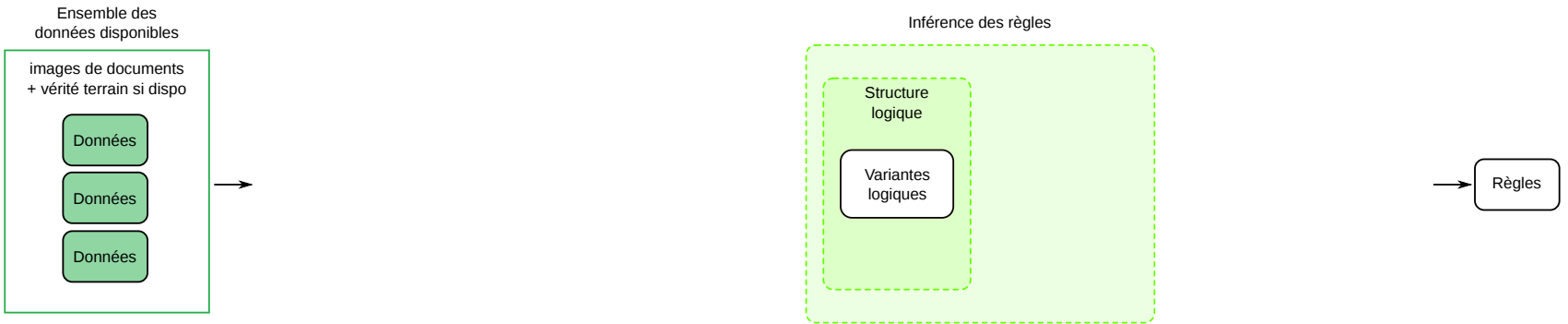
Inférence semi-automatique des règles d'un sous-problème



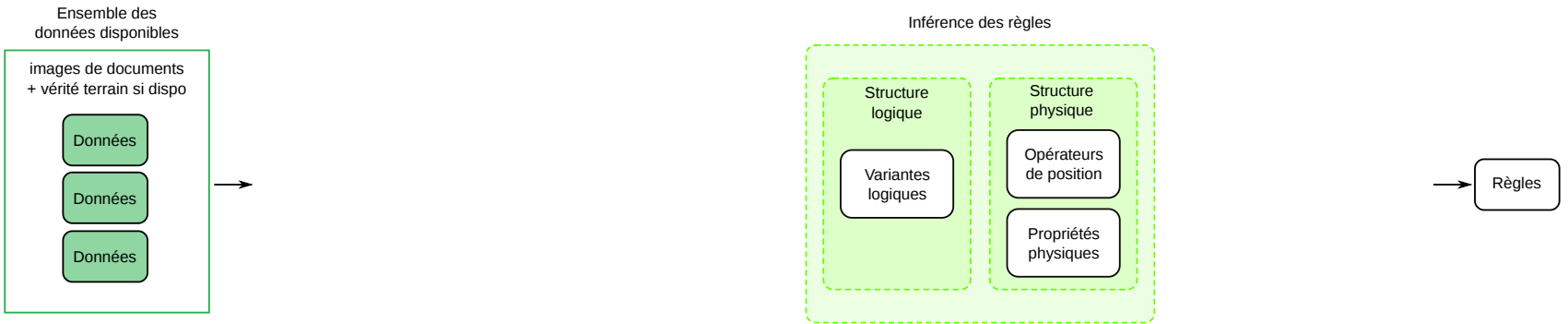
Inférence semi-automatique des règles d'un sous-problème



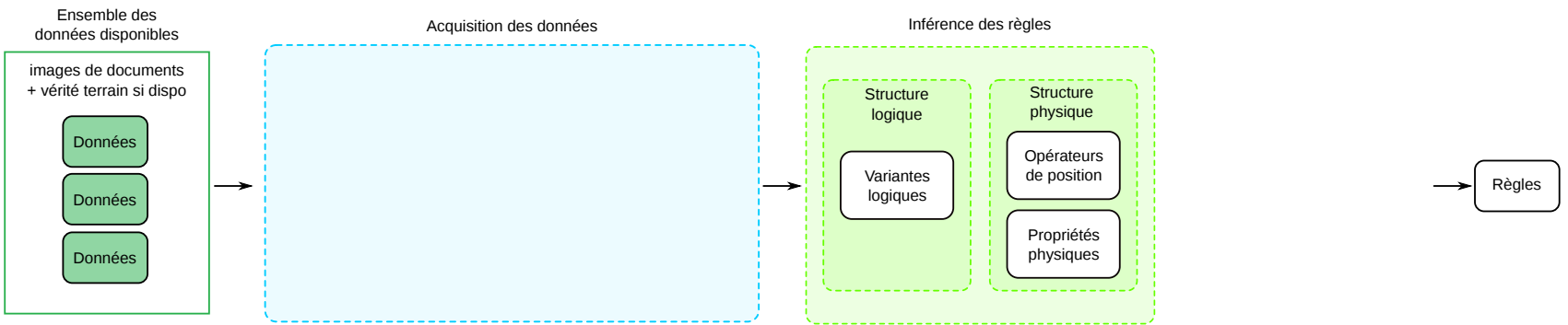
Inférence semi-automatique des règles d'un sous-problème



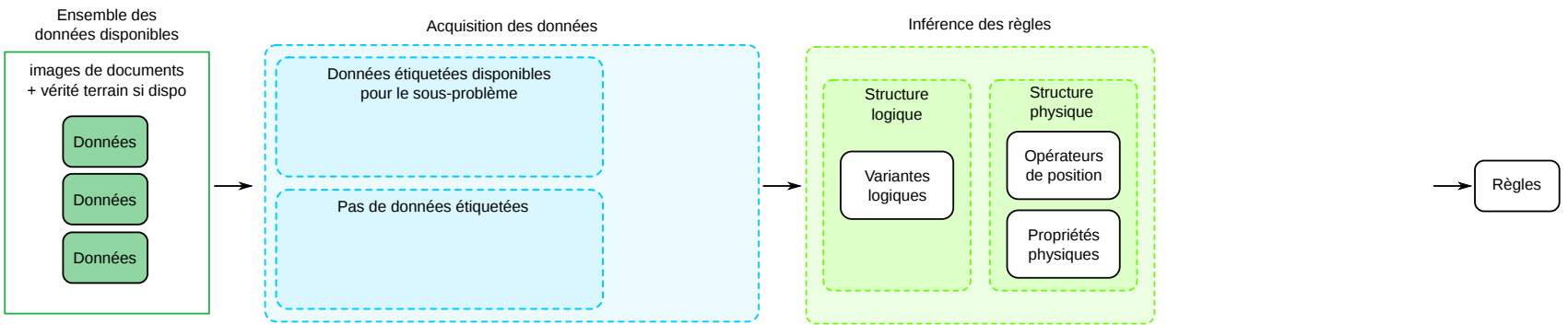
Inférence semi-automatique des règles d'un sous-problème



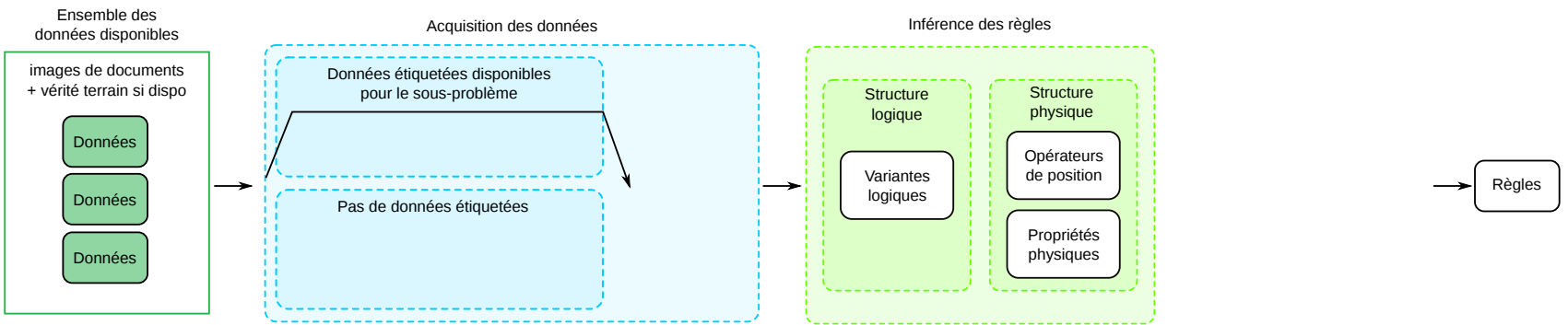
Inférence semi-automatique des règles d'un sous-problème



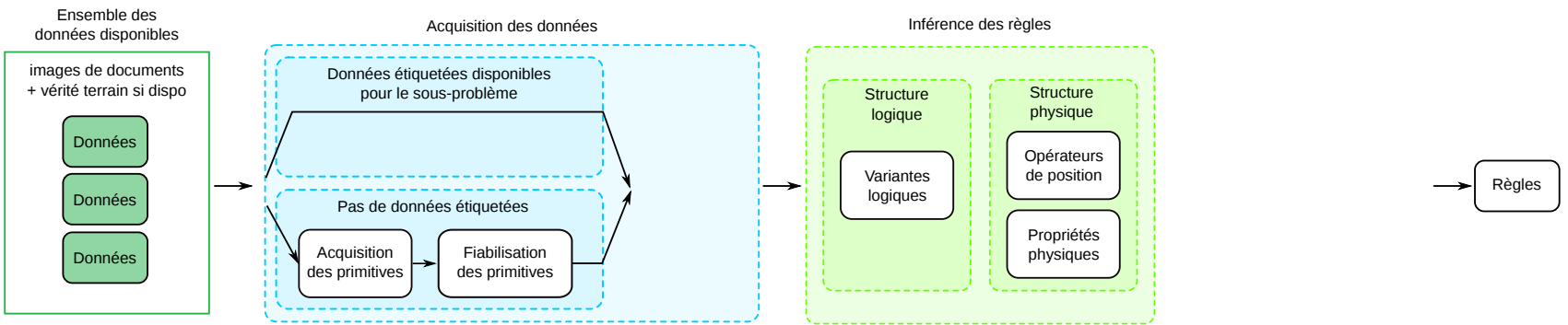
Inférence semi-automatique des règles d'un sous-problème



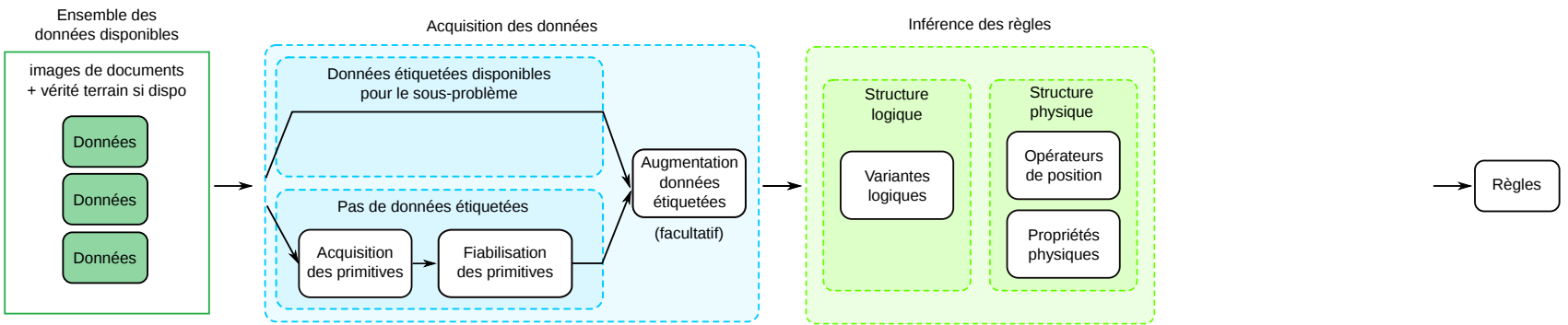
Inférence semi-automatique des règles d'un sous-problème



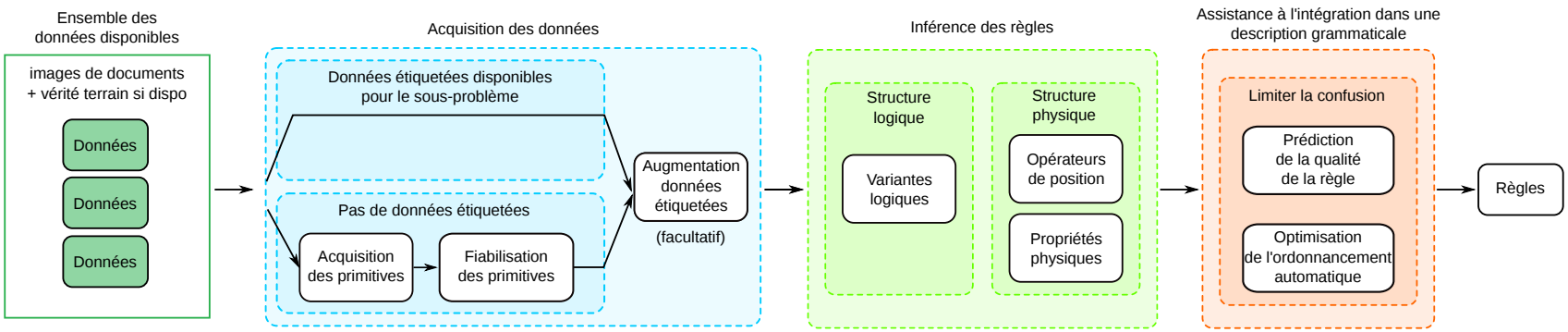
Inférence semi-automatique des règles d'un sous-problème



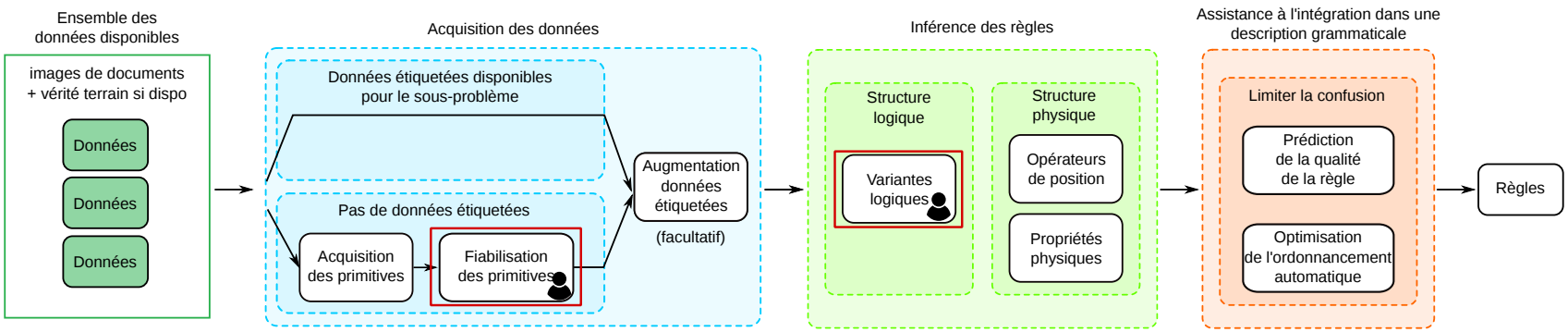
Inférence semi-automatique des règles d'un sous-problème



Inférence semi-automatique des règles d'un sous-problème



Inférence semi-automatique des règles d'un sous-problème

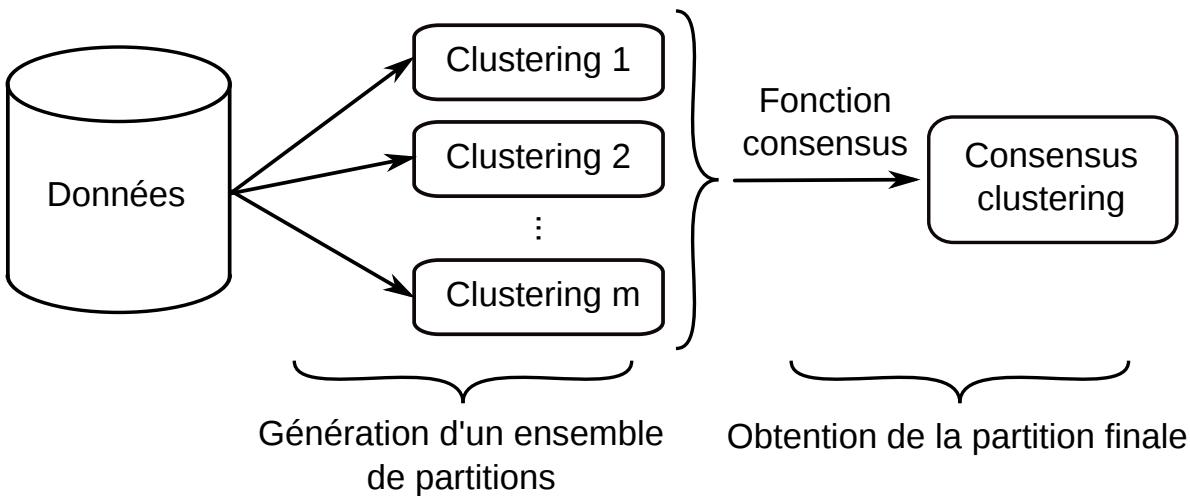


- Utilisateur intégré dans la méthode
- Vue exhaustive et synthétique
  - Prise de décision facilitée

Clustering - quel algorithme choisir ?

- Pas de connaissance utilisateur
  - Algorithme générique s'adaptant à n'importe quel jeu de données
  - Minimisation des paramètres à fixer par l'utilisateur
    $\hookrightarrow$ En particulier le nombre de clusters

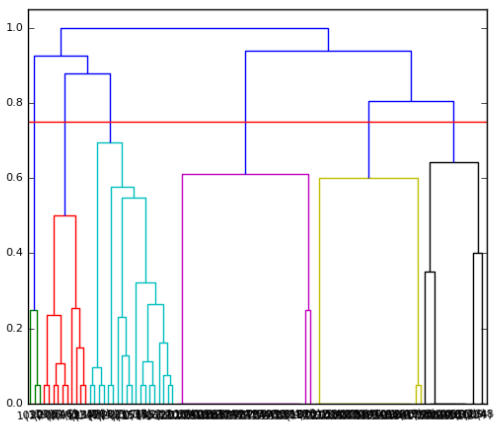
Proposition : utilisation d'une méthode de clustering ensembliste



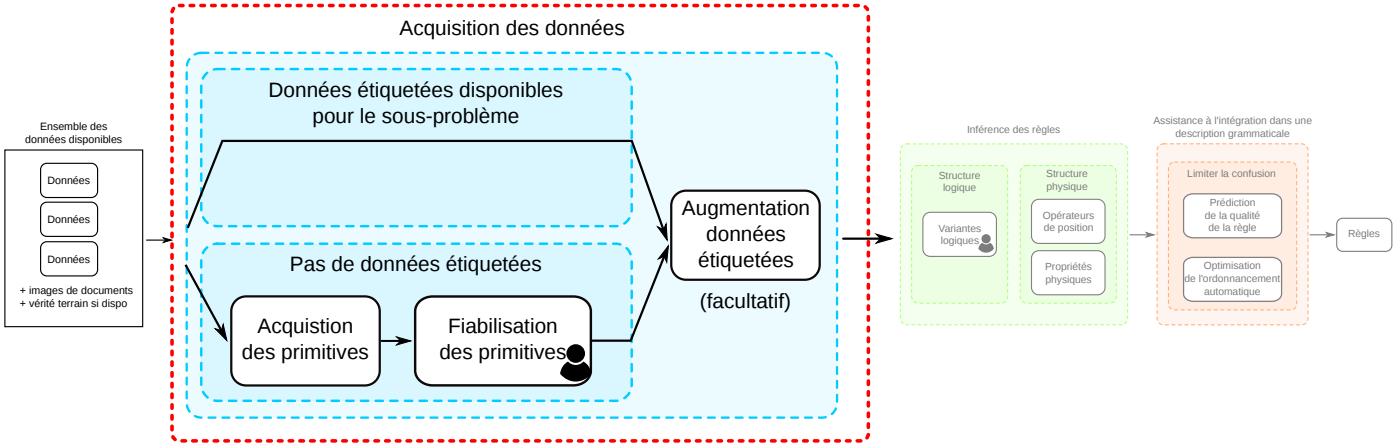
Evidence Accumulation Clustering [Fred and Jain, 2002]

1. Construction de N partitions avec différents
   - Algorithmes
   - Choix de paramètres
2. Combinaison des différentes partitions pour générer une matrice de similarité
   - Si 2 instances souvent ensemble $\Rightarrow$ proches
3. Détermination de la partition finale : algorithme de clustering hiérarchique

   $\hookrightarrow$ *Maximum cluster lifetime* pour couper le dendrogramme

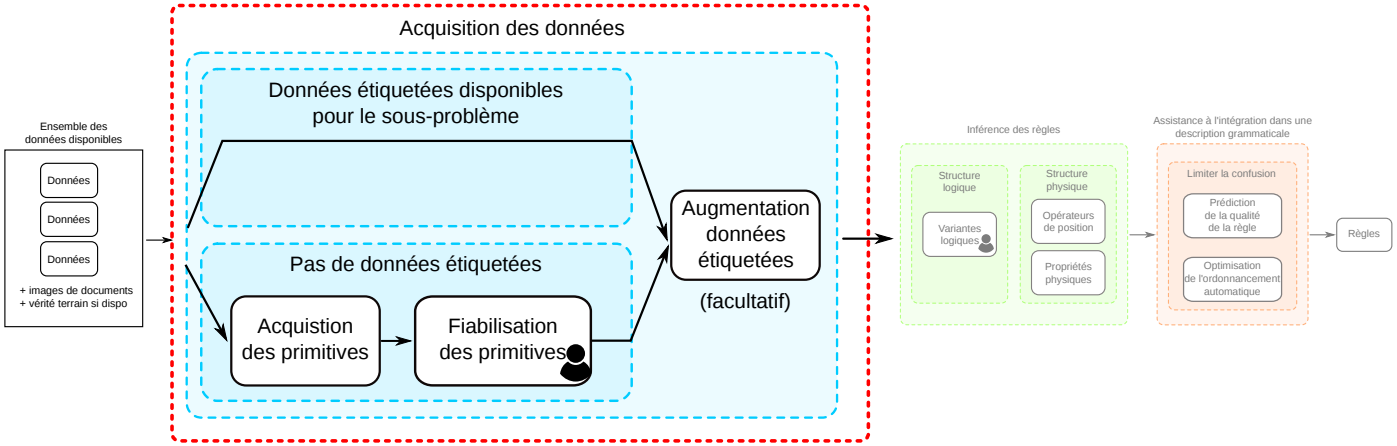


Acquisition des données



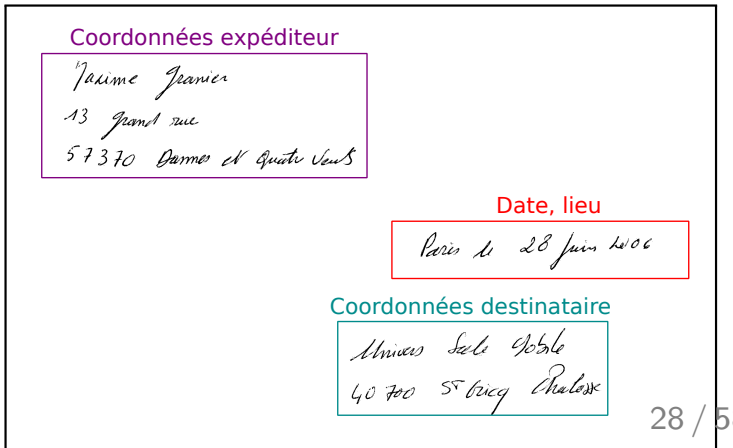
- Cas simple : les données étiquetées existent
- Sinon : construction d'une pseudo vérité terrain
  - Redondances
  - Grands volumes de documents

Acquisition des données



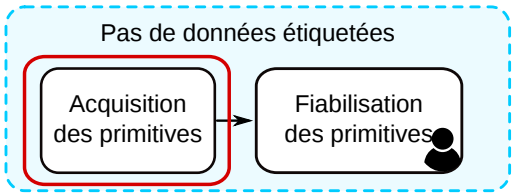
- Cas simple : les données étiquetées existent
- Sinon : construction d'une pseudo vérité terrain
  - Redondances
  - Grands volumes de documents

- Données utiles
  - Minimum : boîtes englobantes des éléments
  - Mais aussi : transcription, langue, ordre de lecture, etc.



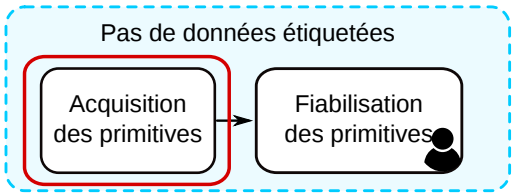
Pseudo vérité terrain : acquisition des primitives

- Extraction automatique
- Sources variées : OCR, word spotting, systèmes de reconnaissance (lignes de texte, segments, etc.)
- Déterminées par l'utilisateur



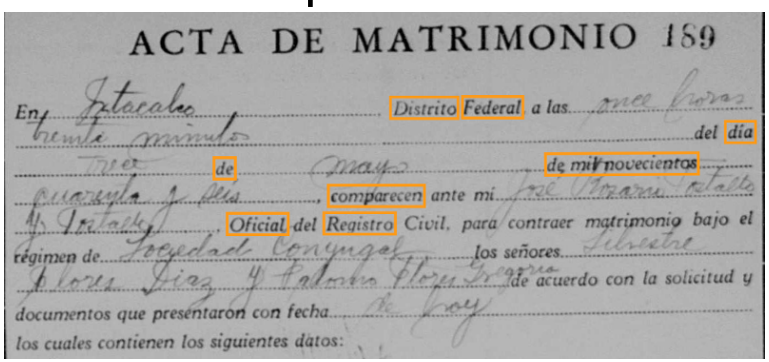
Pseudo vérité terrain : acquisition des primitives

- Extraction automatique
- Sources variées : OCR, word spotting, systèmes de reconnaissance (lignes de texte, segments, etc.)
- Déterminées par l'utilisateur



Extraction de la position de 8 mots-clés

- POI : descripteurs locaux sur la variation du gradient



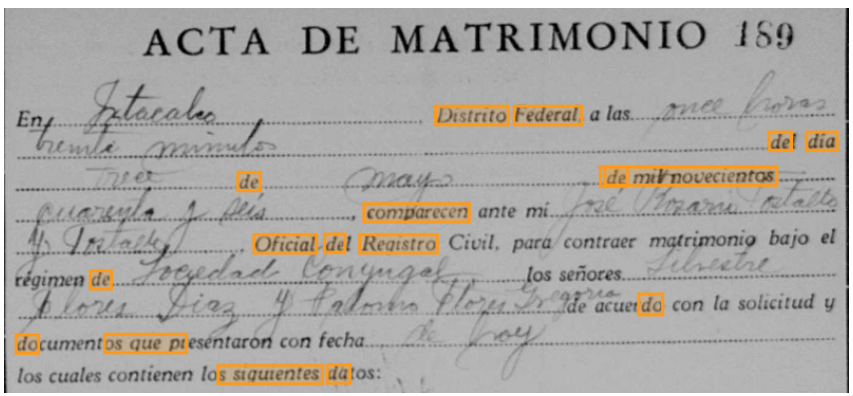
- Autre méthode possible : OCR
  - Faibles performances
  - Mauvaise qualité + interaction manuscrit/imprimé

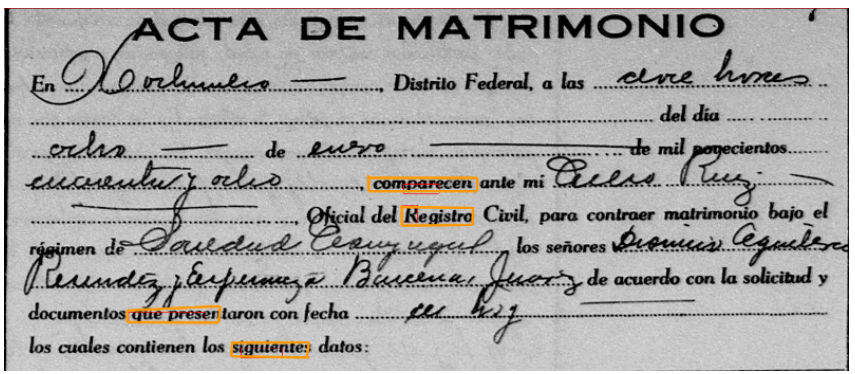
Pseudo vérité terrain : extraction des primitives

Cas idéal : 8 mots clés par document

Cas réel : erreurs d'extraction

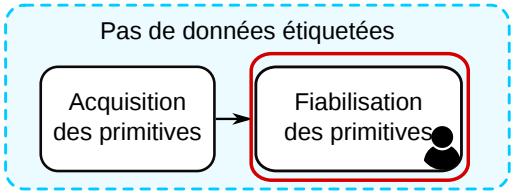
- Oublis, fausses reconnaissances





⇒ Données bruitées

Pseudo vérité terrain : fiabilisation des primitives



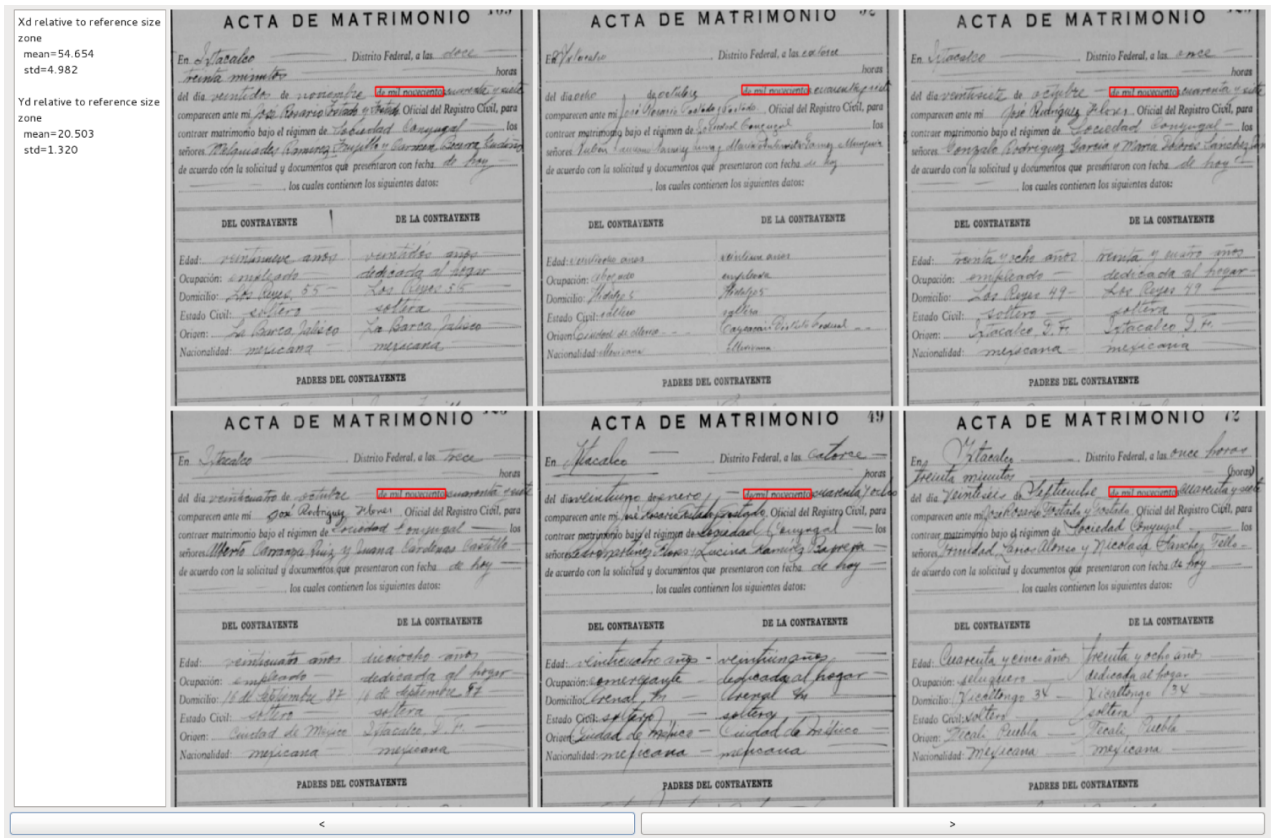
But : supprimer le bruit à moindre coût

1. Clustering des occurrences
   - EAC clustering
   - Étape automatique
2. Suppression des occurrences par cluster
   - Étape interactive
   - Visualisation de quelques exemples par cluster
     - Sélectionnés automatiquement
     - Proches du centroïde
   - Décision de supprimer ou conserver le cluster par l'utilisateur
     - Basée sur les exemples
     - Indicateurs de dispersion pour aider l'utilisateur

Résultat : pseudo vérité terrain construite semi-automatiquement

Pseudo vérité terrain : fiabilisation des primitives

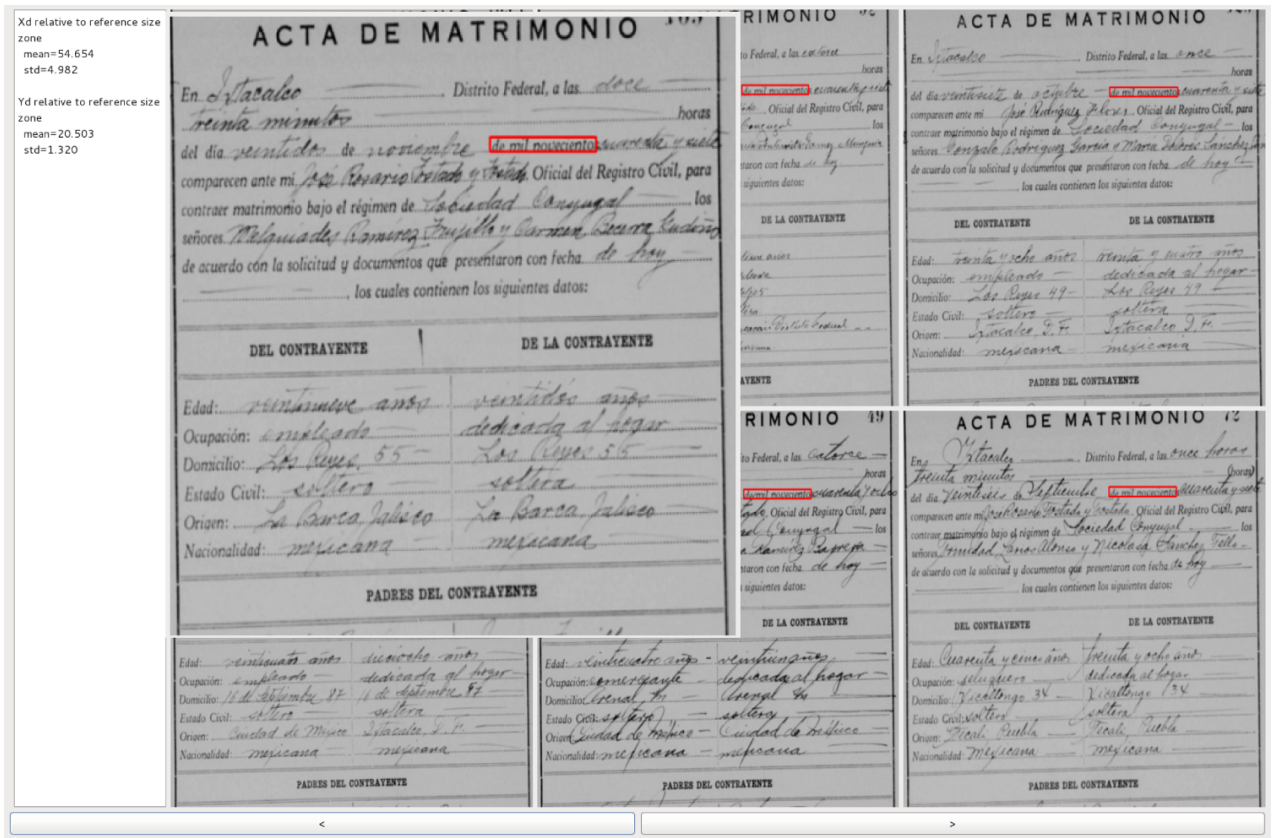
- Exemple : mot clé « de mil novecientos »
- Visualisation des clusters par l'utilisateur
  ↪ quelques exemples représentatifs



1192 occurrences

Pseudo vérité terrain : fiabilisation des primitives

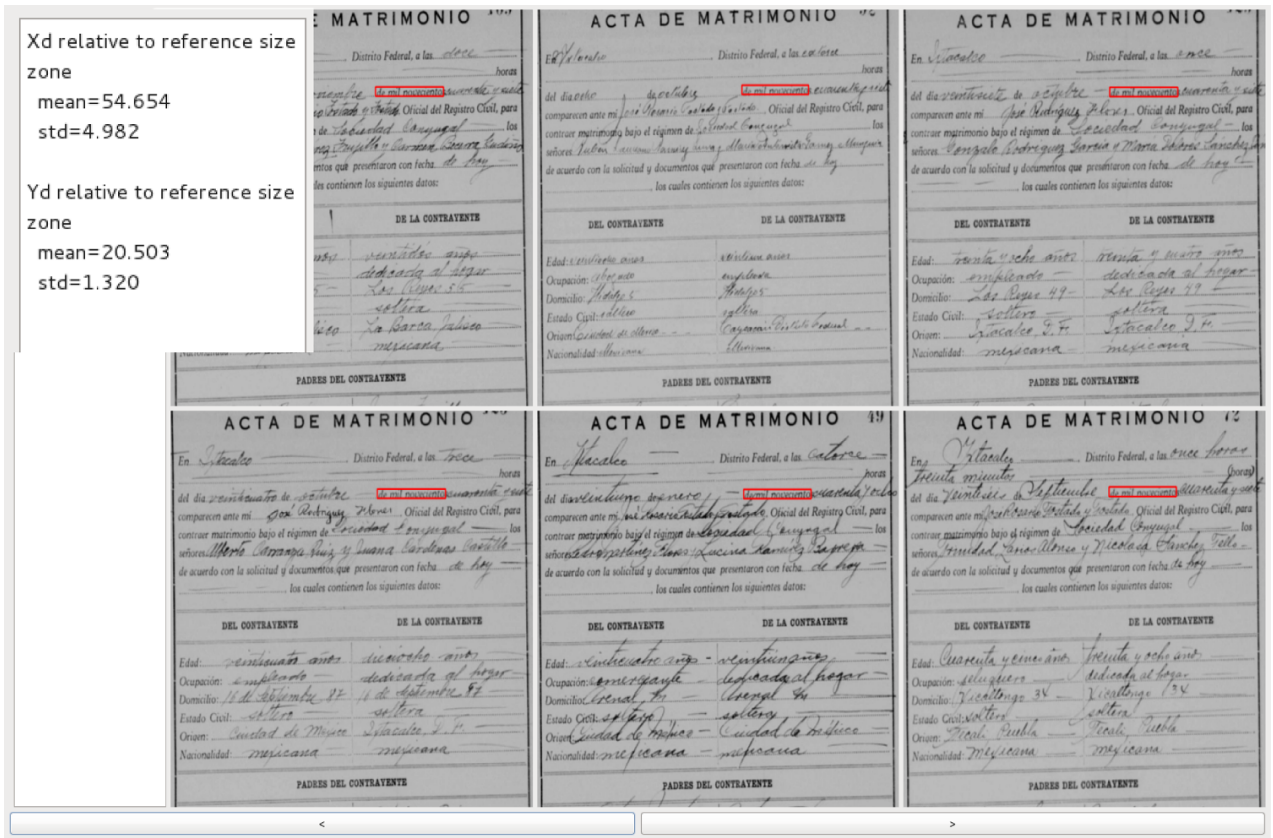
- Exemple : mot clé « de mil novecientos »
- Visualisation des clusters par l'utilisateur
  ↪ quelques exemples représentatifs



1192 occurrences

Pseudo vérité terrain : fiabilisation des primitives

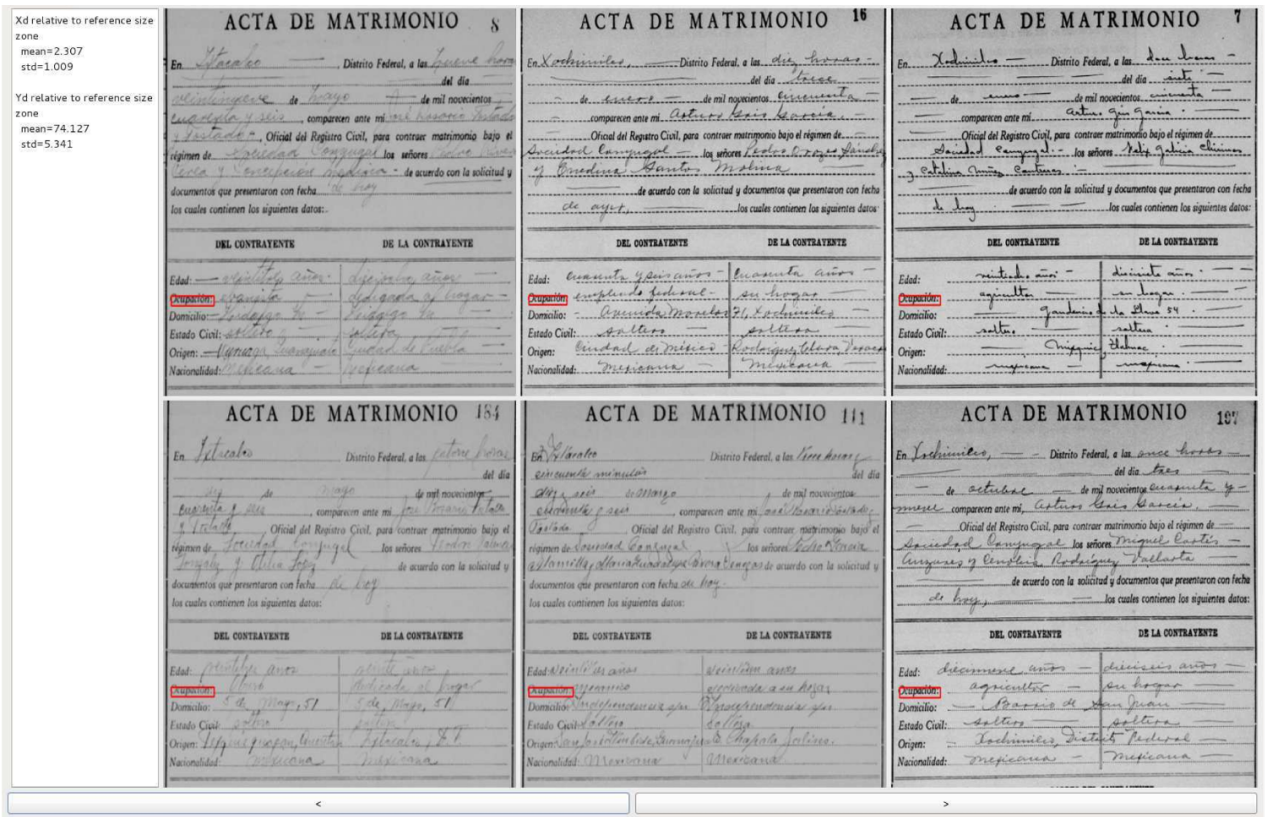
- Exemple : mot clé « de mil novecientos »
- Visualisation des clusters par l'utilisateur
  ↪ quelques exemples représentatifs



1192 occurrences → décision : cluster conservé

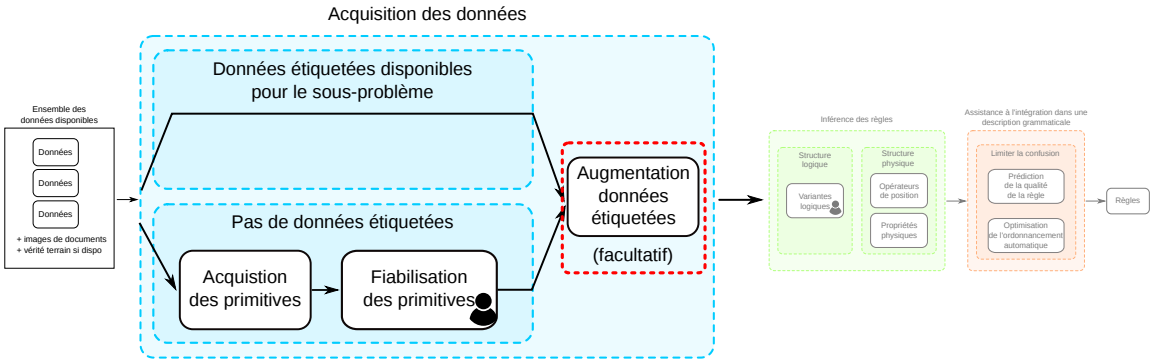
Pseudo vérité terrain : fiabilisation des primitives

- Exemple : mot clé « de mil novecientos »
- Visualisation des clusters par l'utilisateur
  ↪ quelques exemples représentatifs



2114 occurrences → décision : cluster rejeté (mot clé *ocupación*)

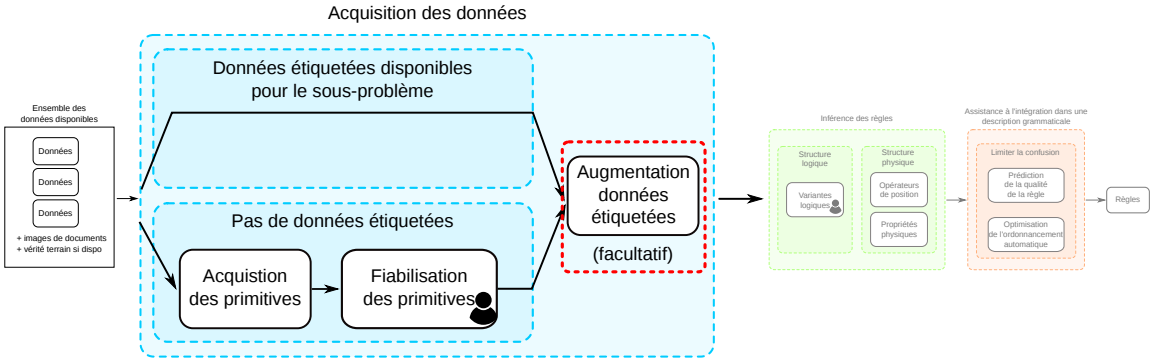
Augmentation de la vérité terrain



Ajout d'éléments

- Nécessaire à la reconnaissance
- Automatiquement
  - À partir de systèmes de reconnaissances existants
  - Utilisation de la position des éléments

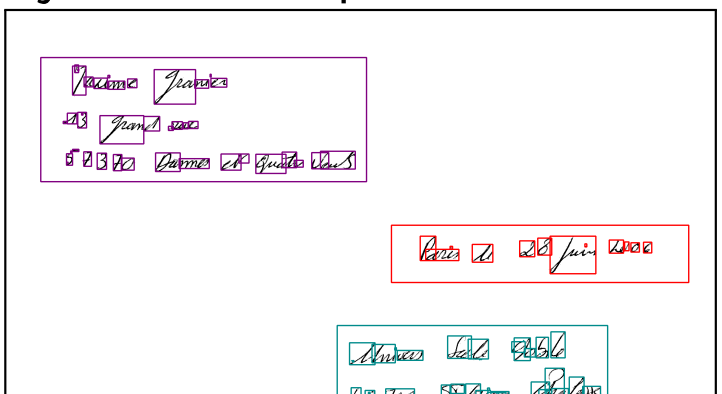
Augmentation de la vérité terrain



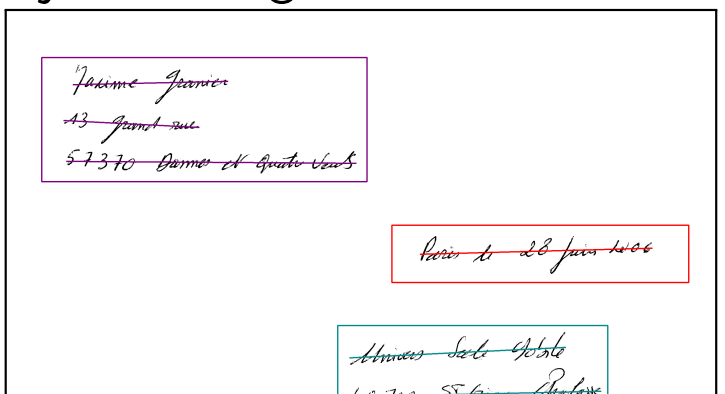
Ajout d'éléments

- Nécessaire à la reconnaissance
- Automatiquement
  - À partir de systèmes de reconnaissances existants
  - Utilisation de la position des éléments

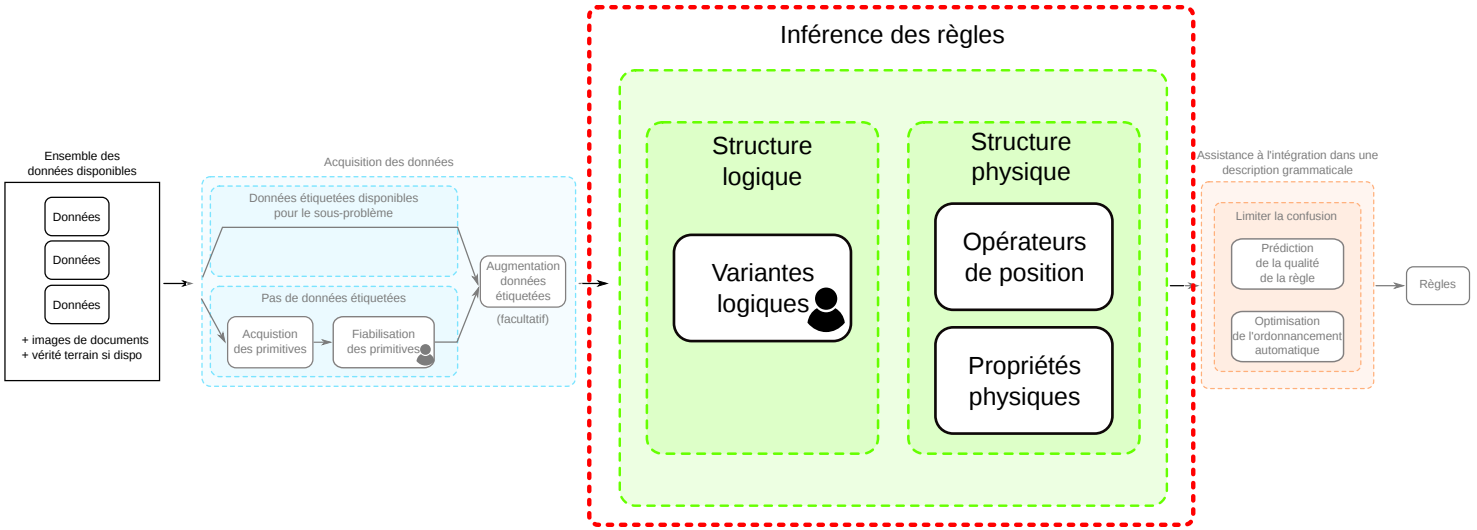
Ajout des composantes connexes



Ajout des lignes de texte



Inférence des règles



Grammaire de départ courrier manuscrit

courrier ::=

coordonneesExpediteur &&
coordonneesDestinataire &&
dateLieu &&
objet &&
ouverture &&
corpsDeTexte &&
signature &&
ps.

Inférence des règles - structure logique

Structure logique

- Apprentissage de la hiérarchie de la grammaire
- Grande variabilité possible par élément
- Nécessite une vision exhaustive des données

Inférence des règles - structure logique

Structure logique

- Apprentissage de la hiérarchie de la grammaire
- Grande variabilité possible par élément
- Nécessite une vision exhaustive des données

Apprentissage des variantes logiques des éléments

- Détection automatique des variantes
  - Clustering
- Apport de sens aux données
  - Interaction utilisateur
  - Visualisation de quelques exemples
  - Validation des variantes détectées
  - Attribution d'un nom significatif à chaque variante
    - Description grammaticale interprétable

Inférence des règles - structure logique

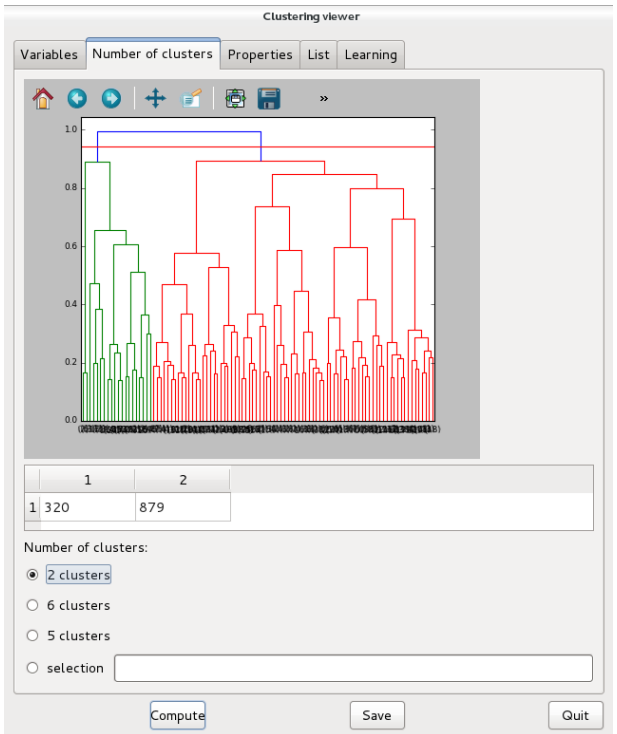
Coordonnées expéditeur

- Recherche des variantes logiques
- Variables : hauteur et largeur
- Partition en deux clusters proposée par EWO

Inférence des règles - structure logique

Coordonnées expéditeur

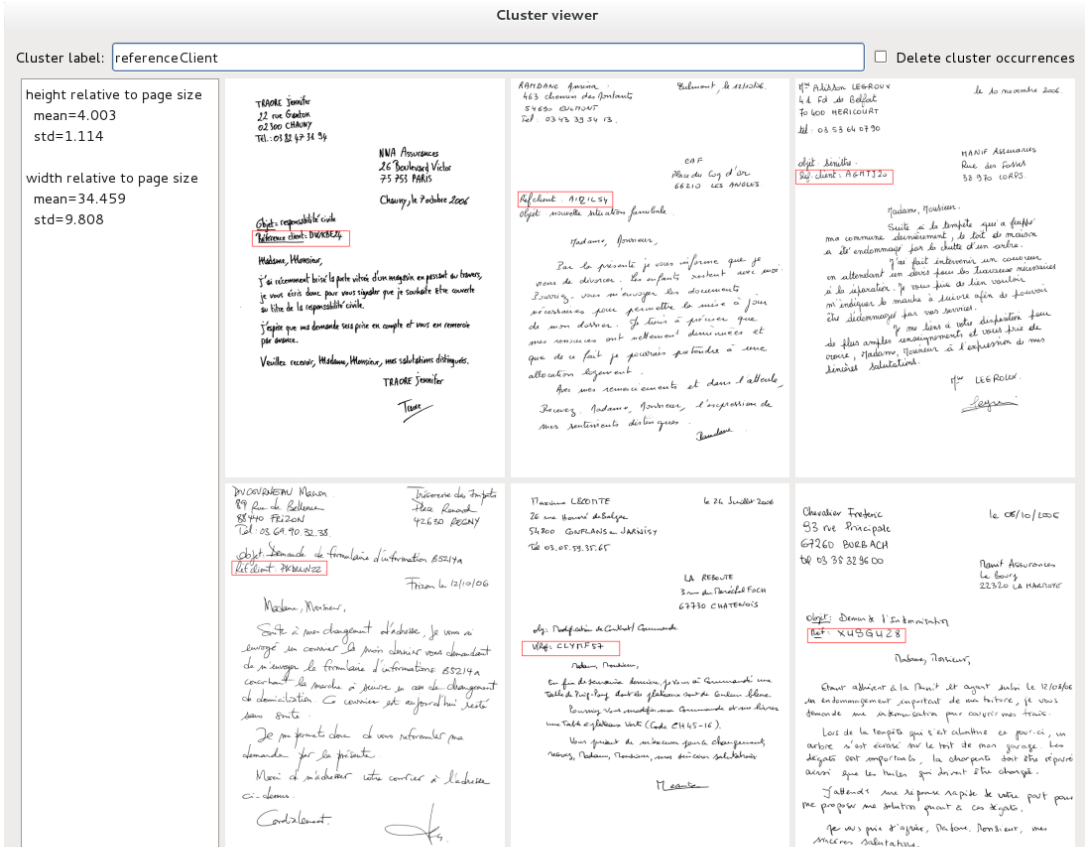
- Recherche des variantes logiques
- Variables : hauteur et largeur
- Partition en deux clusters proposée par EWO



Inférence des règles - structure logique

Coordonnées expéditeur

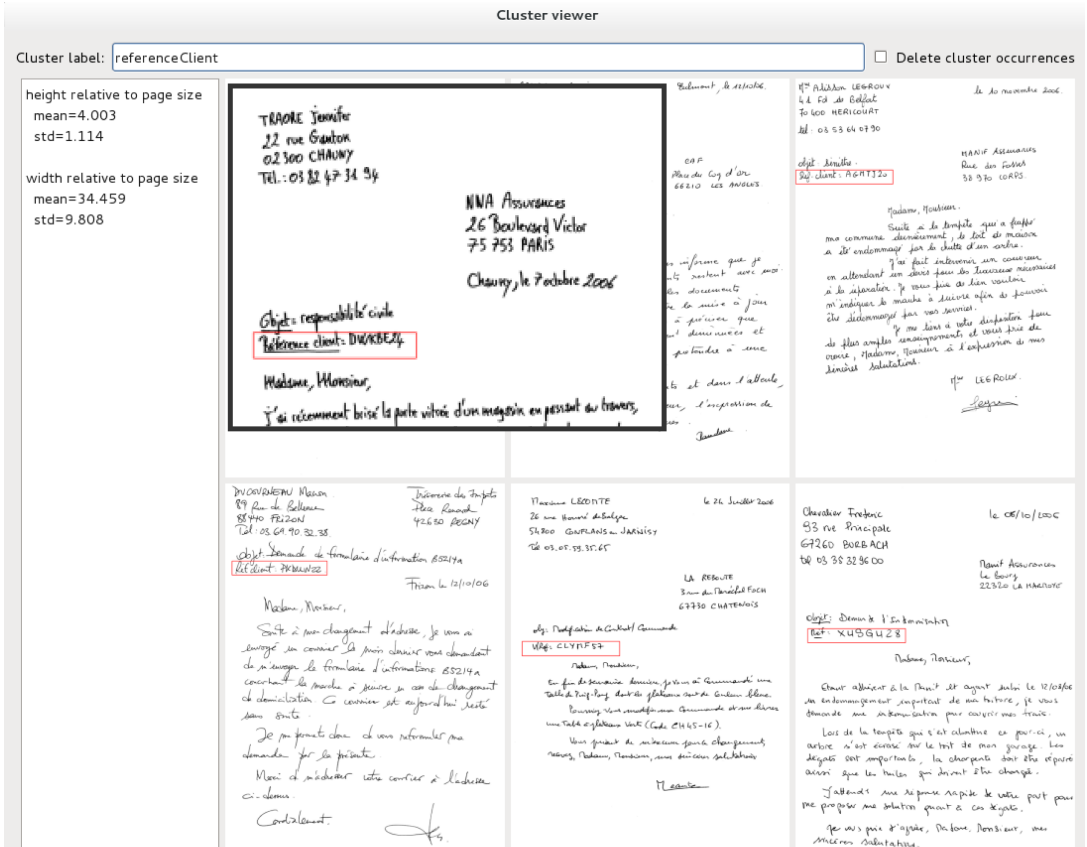
- Recherche des variantes logiques
- Variables : hauteur et largeur
- Partition en deux clusters proposée par EWO



Inférence des règles - structure logique

Coordonnées expéditeur

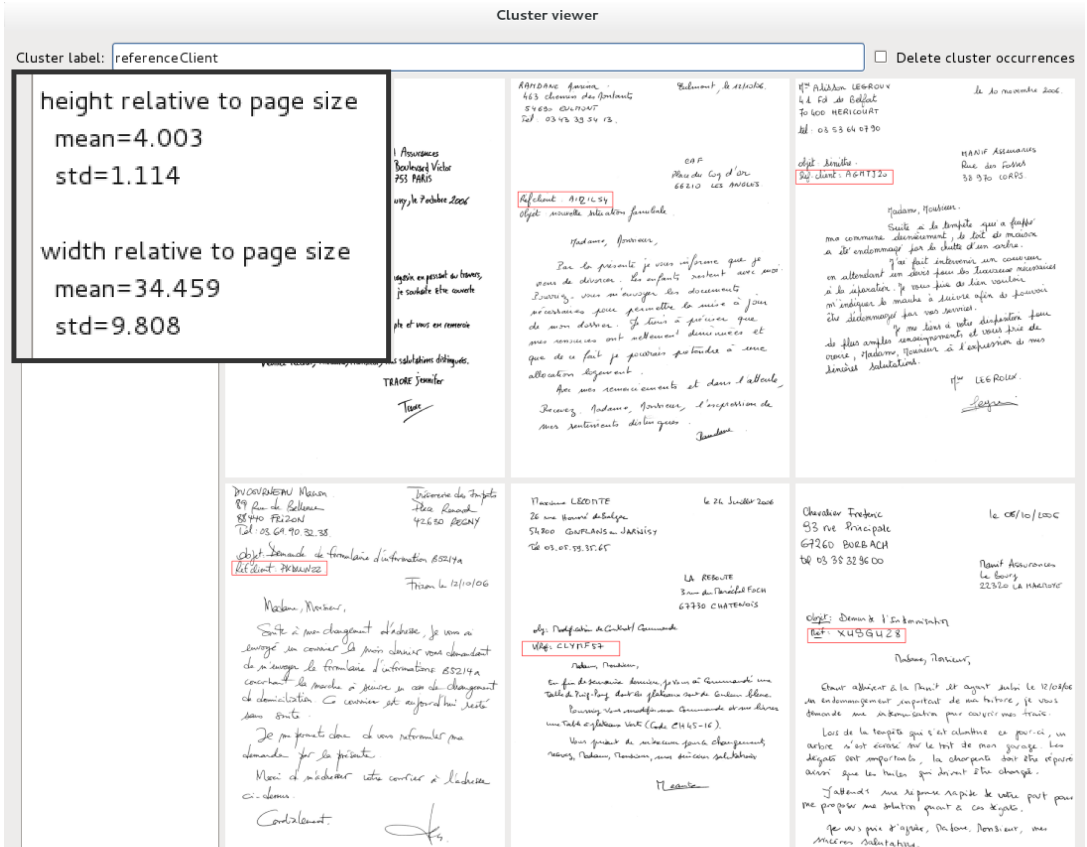
- Recherche des variantes logiques
- Variables : hauteur et largeur
- Partition en deux clusters proposée par EWO



Inférence des règles - structure logique

Coordonnées expéditeur

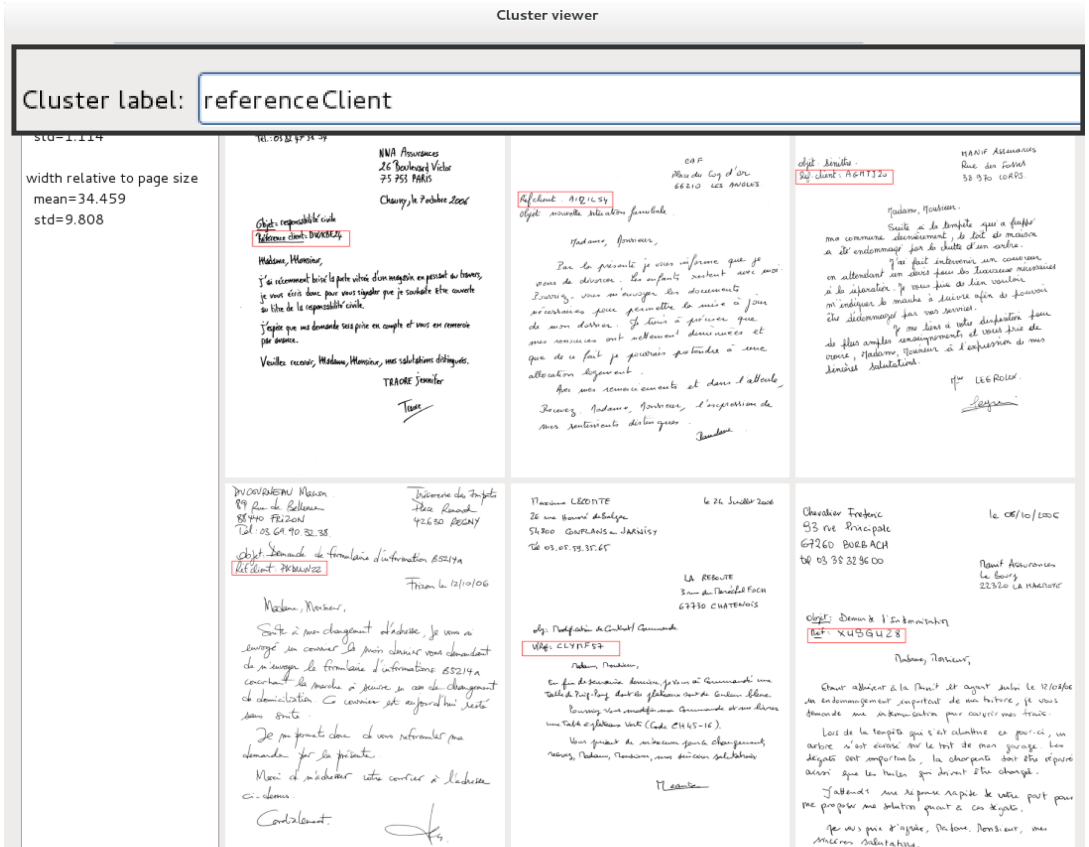
- Recherche des variantes logiques
- Variables : hauteur et largeur
- Partition en deux clusters proposée par EWO



Inférence des règles - structure logique

Coordonnées expéditeur

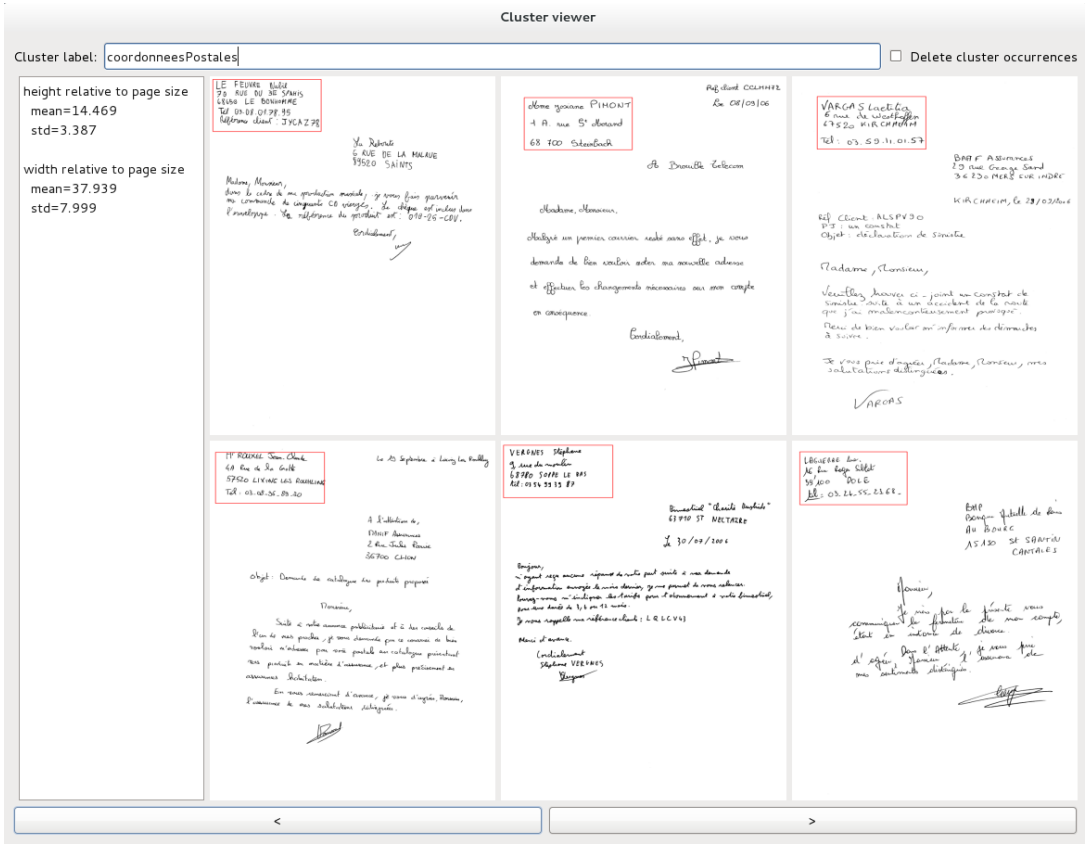
- Recherche des variantes logiques
- Variables : hauteur et largeur
- Partition en deux clusters proposée par EWO



Inférence des règles - structure logique

Coordonnées expéditeur

- Recherche des variantes logiques
- Variables : hauteur et largeur
- Partition en deux clusters proposée par EWO



Inférence des règles - structure logique

- Détection automatique de deux variantes logiques
- Production de nouvelles règles

courrier ::=
coordonneesExpediteur &&
coordonneesDestinataire &&
dateLieu &&
objet &&
ouverture &&
corpsDeTexte &&
signature &&
ps.

coordonneesExpediteur ::=
referenceClient.

coordonneesExpediteur ::=
coordonneesPostales.

Détection des valeurs extrêmes

- Cas limite de ce qui doit être reconnu
- Difficile à détecter par un humain
- Détectées automatiquement et présentées à l'utilisateur

1. Cas rares
   - Définition de variantes supplémentaires
2. Erreurs dans les données étiquetées
   - Suppression ou correction

$\Rightarrow$ Nécessaire pour la robustesse de l'analyse

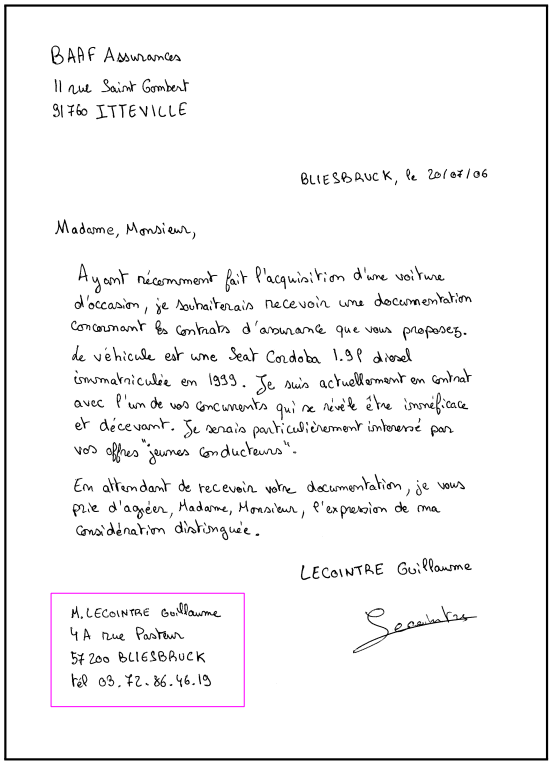
Détection des valeurs extrêmes

- Cas limite de ce qui doit être reconnu
- Difficile à détecter par un humain
- Détectées automatiquement et présentées à l'utilisateur

❶ Cas rares
  - Définition de variantes supplémentaires

❷ Erreurs dans les données étiquetées
  - Suppression ou correction

⇒ Nécessaire pour la robustesse de l'analyse


BAAF Assurances
11 rue Saint Gombert
91760 ITTEVILLE

BLIESBRUCK, le 20/07/06

Madame, Monsieur,

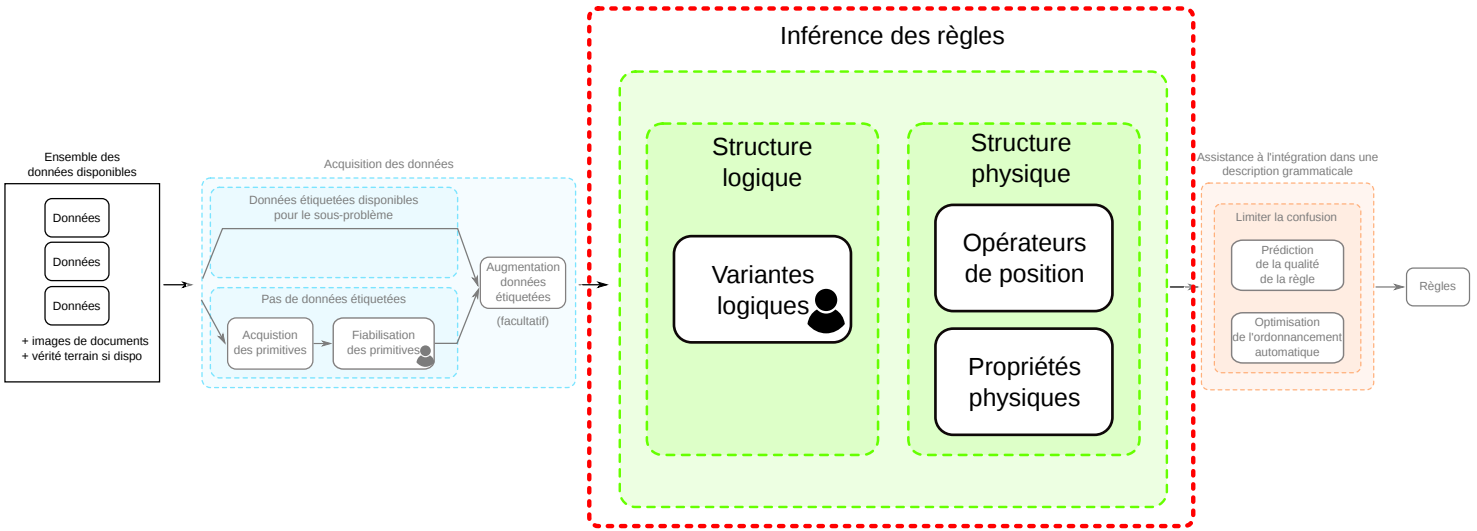
Ayant récemment fait l'acquisition d'une voiture d'occasion, je souhaiterais recevoir une documentation concernant les contrats d'assurance que vous proposez. Le véhicule est une Seat Cordoba 1.9l diesel immatriculée en 1999. Je suis actuellement en contrat avec l'un de vos concurrents qui se révèle être inefficace et décevant. Je serais particulièrement intéressé par vos offres "jeunes conducteurs".

En attendant de recevoir votre documentation, je vous prie d'agréer, Madame, Monsieur, l'expression de ma considération distinguée.

LECOINTRE Guillaume

M. LECOINTRE Guillaume
4A rue Pasteur
57200 BLIESBRUCK
tél 03.72.86.46.19

Inférence des règles - structure physique



Comment le système peut reconnaitre les éléments de la structure logique ?

- Positionnement des éléments
- Propriétés physiques des éléments

Nécessite une vue exhaustive sur les données

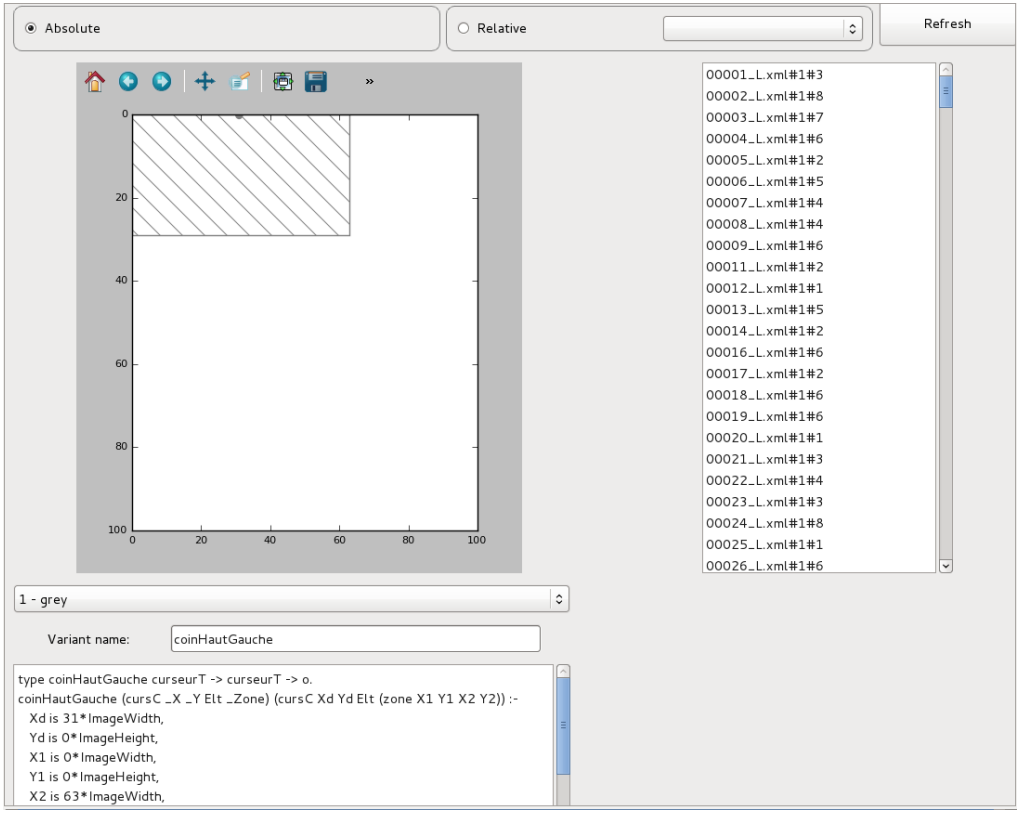
Positionnement des éléments

- Détermine la zone de recherche d'un élément
  - Gestion de la combinatoire
  - Maximisation du rappel

Positionnement des éléments

- Détermine la zone de recherche d'un élément
  - Gestion de la combinatoire
  - Maximisation du rappel

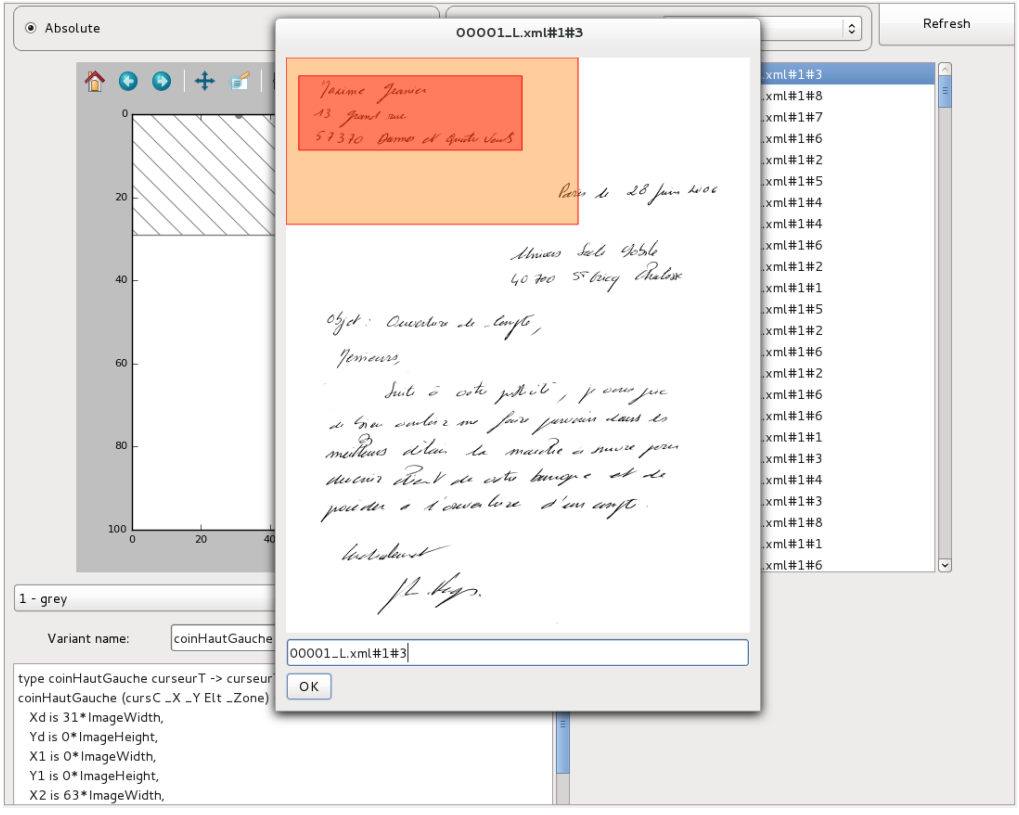
- Choix utilisateur
  - positionnement absolu ou relatif
- Détection automatique des variantes
- Frontières de la zone de recherche
- Ordre de parcours des éléments
- Nom choisi par l'utilisateur



Positionnement des éléments

- Détermine la zone de recherche d'un élément
  - Gestion de la combinatoire
  - Maximisation du rappel

- Choix utilisateur
  - positionnement absolu ou relatif
- Détection automatique des variantes
- Frontières de la zone de recherche
- Ordre de parcours des éléments
- Nom choisi par l'utilisateur



Positionnement des éléments

Production de nouvelles règles

courrier ::=
coordonneesExpediteur &&
coordonneesDestinataire &&
dateLieu &&
objet &&
ouverture &&
corpsDeTexte &&
signature &&
ps.

coordonneesExpediteur ::=
referenceClient.

coordonneesExpediteur ::=
coordonneesPostales.

coordonneesPostales ::=
AT(coinHautGauche) &&
...

coinHautGauche :-
Xp = 0,
Yp = -100,
Xa = 0,
Ya = 0,
Xb = 62% largeur page,
Yb = 27% hauteur page.

Propriétés physiques

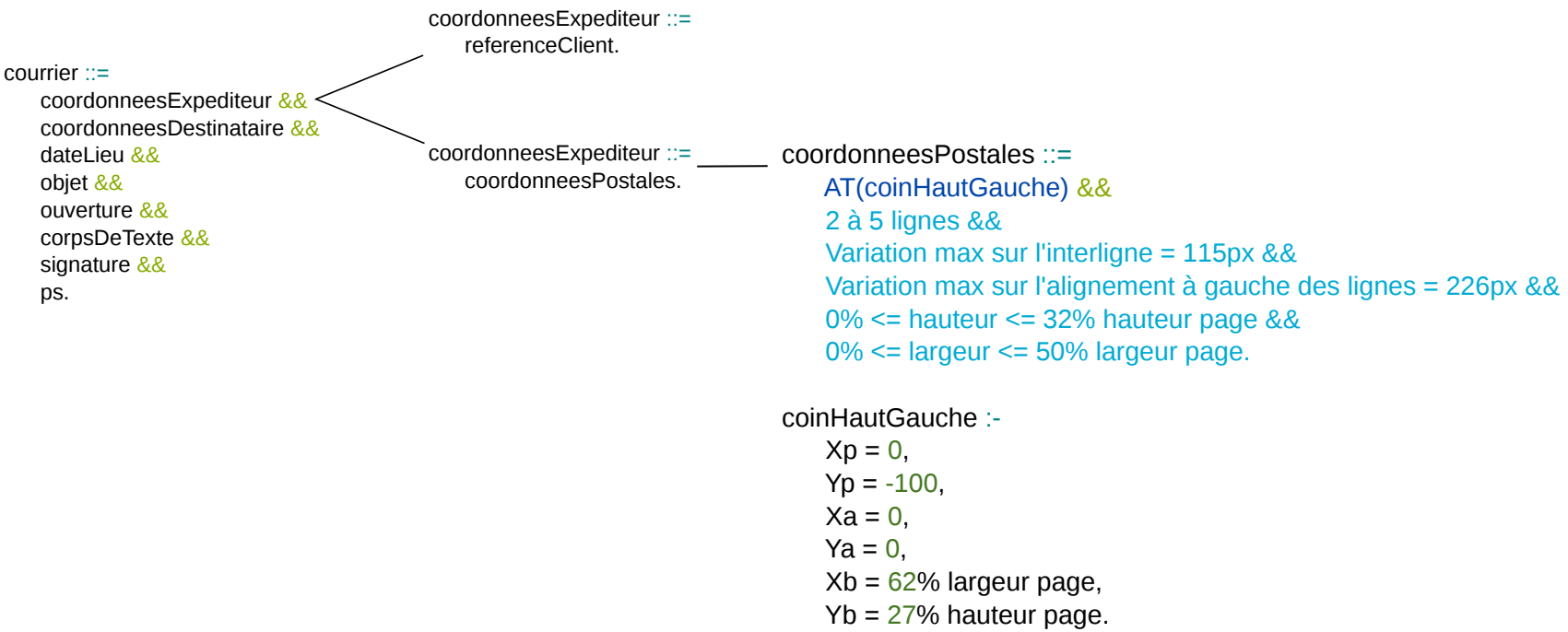
Détermination des caractéristiques propres de l'élément

- Nécessite une vision globale sur les données
- Permet la segmentation
- Statistiques descriptives

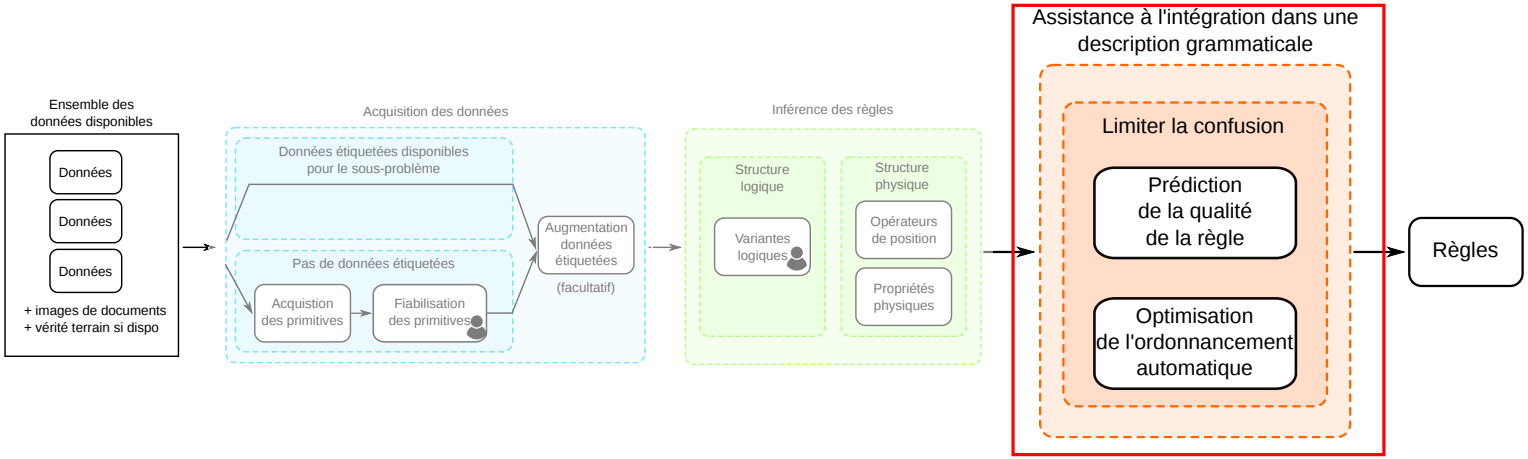
Propriétés physiques

Détermination des caractéristiques propres de l'élément

- Nécessite une vision globale sur les données
- Permet la segmentation
- Statistiques descriptives



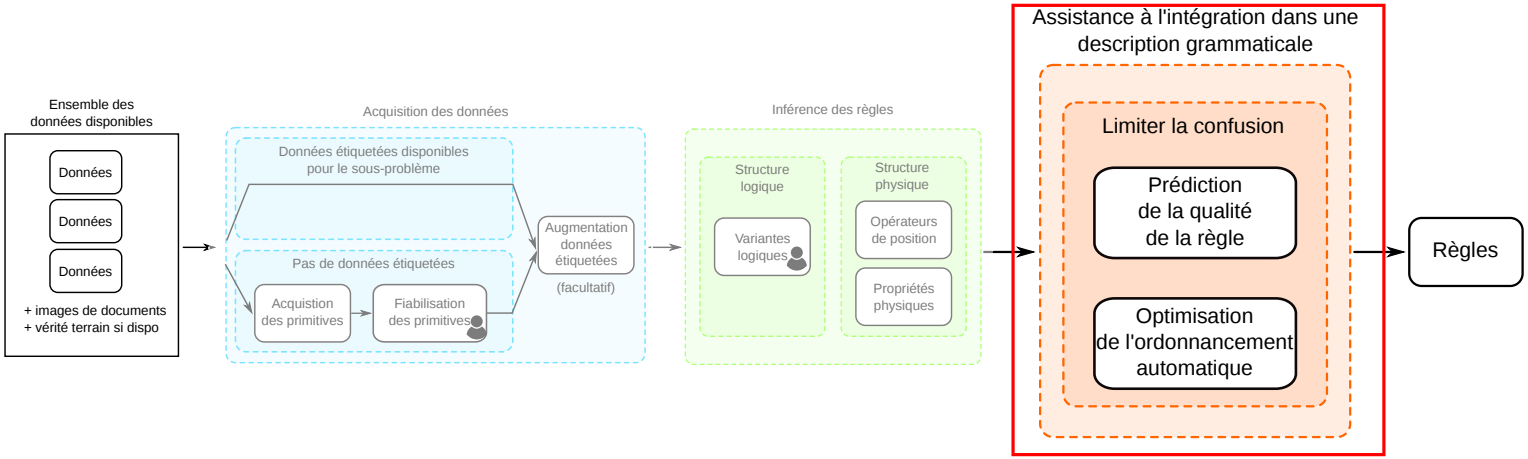
Intégration de la règle inférée dans la description complète



But : minimiser la confusion entre les éléments

- Vérifier que la règle est suffisamment spécifiée
- Ordonner les règles
  - Recherche des éléments les plus fiables en premiers

Intégration de la règle inférée dans la description complète



But : minimiser la confusion entre les éléments

- Vérifier que la règle est suffisamment spécifiée
- Ordonner les règles
  - Recherche des éléments les plus fiables en premiers

Comment ? Fonctionnement sous forme de requêtes au sein de la méthode EWO

- Approximation du système de reconnaissance
- Réduction du temps de calcul par rapport au système complet
- Action en temps réel de l'utilisateur avec le système

Intégration dans la description complète

Grammaire finale, inférence de

- 34 règles et variantes
- 14 opérateurs de position
- 60 propriétés physiques

Grammaire de base

courrier ::=
coordonneesExpediteur &&
coordonneesDestinataire &&
dateLieu &&
objet &&
ouverture &&
corpsDeTexte &&
signature &&
ps.

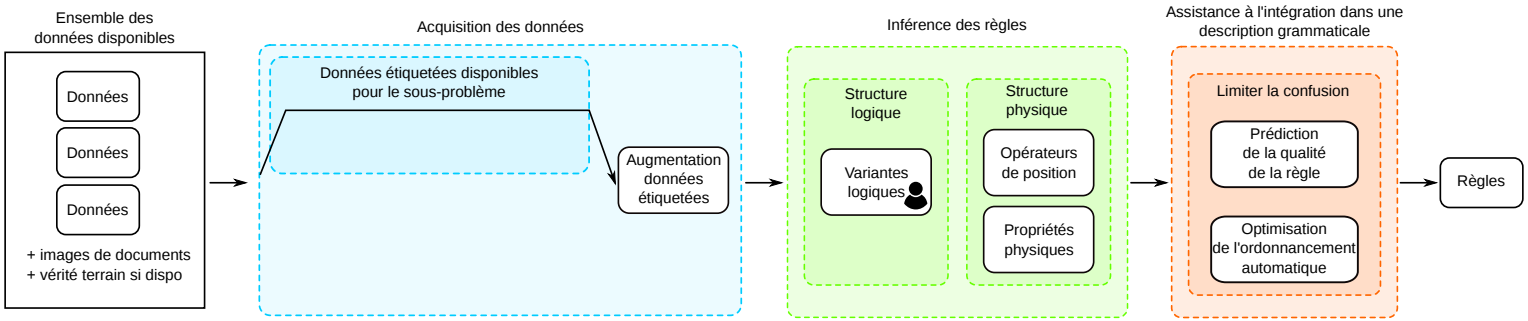
Grammaire finale - ordre optimisé

courrier ::=
ouverture &&
corpsDeTexte &&
signature &&
coordonneesPostales &&
coordonneesDestinataire &&
ps &&
dateLieu &&
objet &&
referenceClient.

Introduction

- Validation complète
  - Différentes parties séparément
  - Fonctionnement global
- Corpus variés
  - RIMES : corpus de courriers manuscrits avec vérité terrain
  - MAURDOR : corpus hétérogène avec vérité terrain
  - FamilySearch HIP 2013 : registres de mariages mexicains sans vérité terrain
- Système syntaxique utilisé : DMOS [Coüasnon, 2006]

RIMES



- Grammaire complète décrite avec EWO
- Métrique du concours
  - Taux de pixels noirs mal étiquetés
- Données
  - Apprentissage : 900 courriers
  - Test : base concours de 100 courriers

- Inférence
  - 34 règles et variantes
  - 14 opérateurs de position
  - 60 paramètres pour les propriétés physiques

RIMES - Résultats

Système	Taux d'erreur
Syntaxique (DMOS) [Lemaitre, 2008]	8,97
Syntaxique (DMOS) localement stochastique [Maroneze, 2011]	5,53
Statistique basé sur un MRF [Lemaitre, 2007]	8,53
Statistique basé sur des CRF [Montreuil, 2010]	6,33
Méthode EWO	5,82

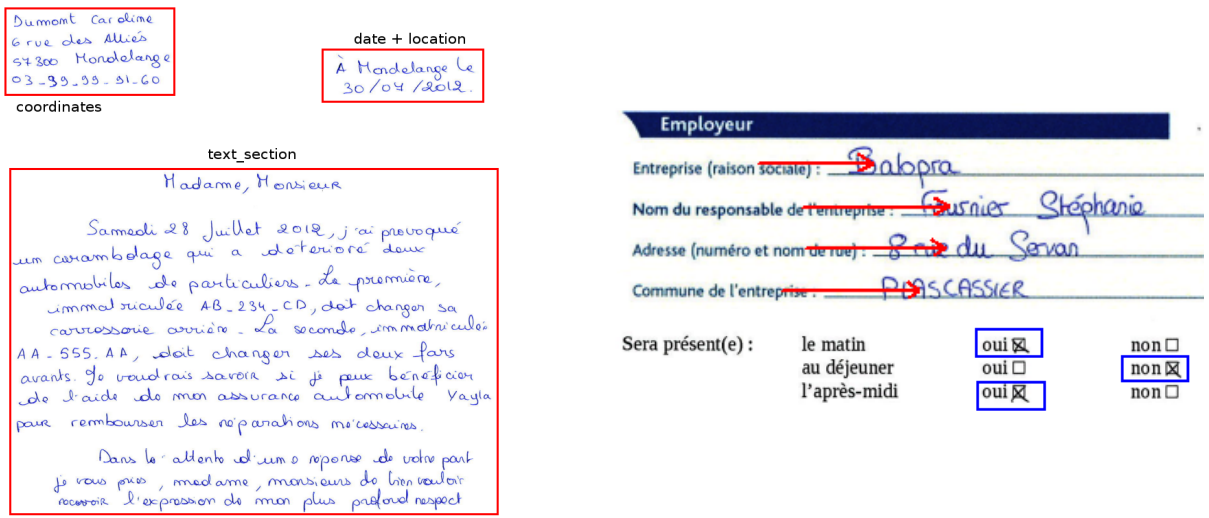
Résultats comparables

- Meilleur système syntaxique
  - Gain de temps avec la méthode EWO
- Meilleur système statistique
  - Meilleure gestion des cas rares avec EWO
  - Homogénéité dans les blocs construits

MAURDOR

Tâche 5 de la compétition sur zones segmentées

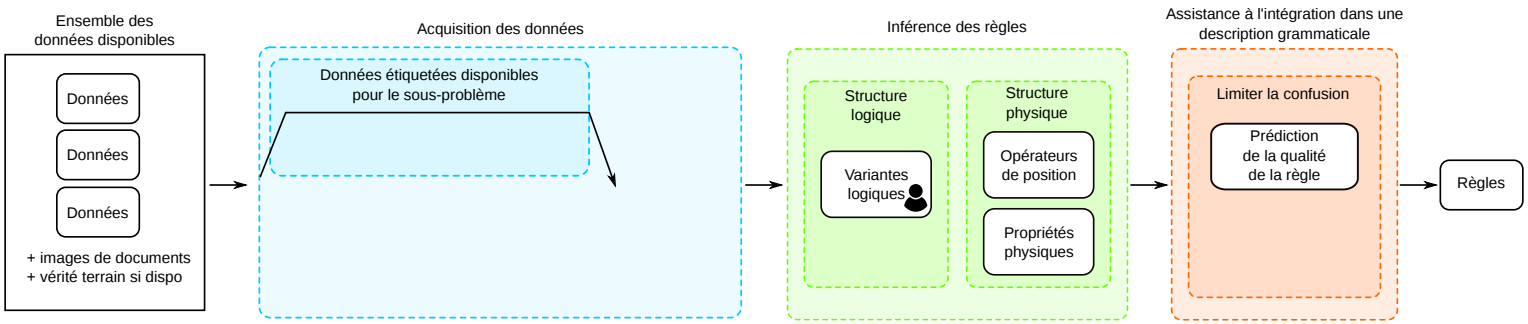
- Étiquetage logique
- Ordre de lecture
- Groupe



Documents variés

- Factures, formulaires, courriers, tableaux, plans, etc.
- Langues : français, anglais et arabe
- Manuscrit et dactylographié

MAURDOR - Résultats



- Données
  - Apprentissage : 2000 documents
  - Test : 1000 documents
- Deux systèmes
  - Syntaxique avec la méthode EWO
  - Second participant :
    - type : classifieur (SVM)
    - ordre + groupe : règles décrites manuellement

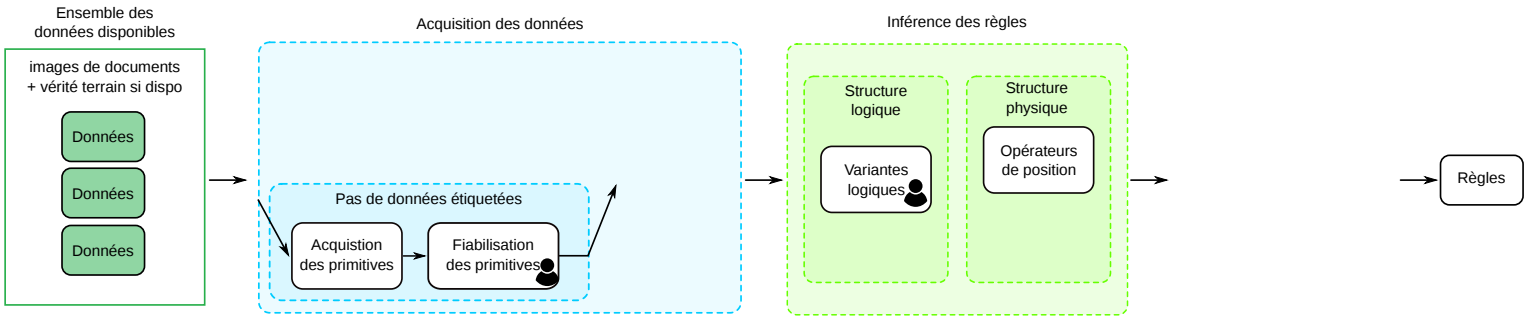
MAURDOR - Résultats

- Métrique du concours
  - Score $\in [-100; 100]$
  - Score $> 0 \Leftrightarrow$ apport de plus d'informations que d'erreurs

Système	Type	Ordre	Groupe
Méthode EWO	55	45	60
Participant 2	69	28	61

- Type : zones déjà segmentées $\Rightarrow$ tâche de classification
  - Meilleur résultat du SVM
- Ordre
  - Meilleurs résultats pour les règles définies avec EWO
  - Validation de l'apport de la vue exhaustive sur les données
- Détection de 164 erreurs dans la vérité terrain

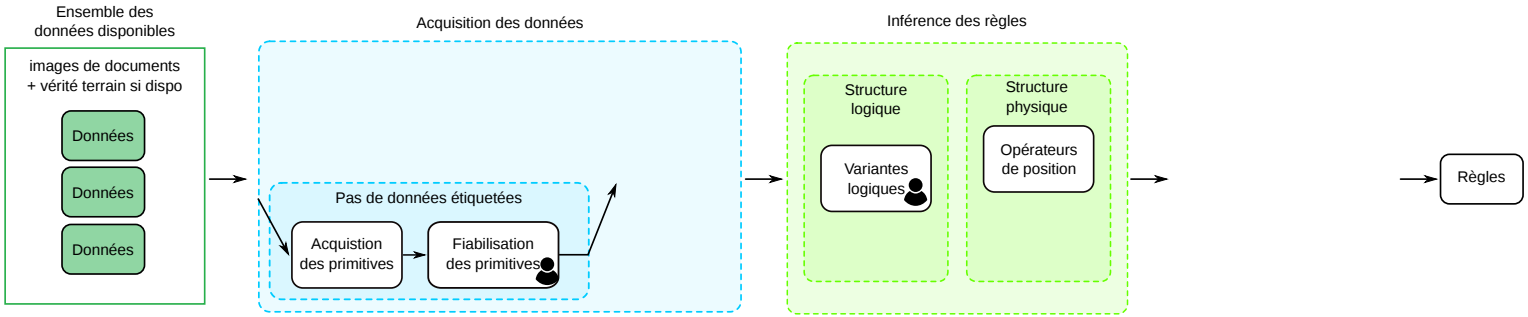
FamilySearch HIP2013



Jeu de données : registres de mariages de la compétition HIP2013 FamilySearch

- Apprentissage : 7000 documents
  - Sans vérité terrain
- Test : 2000 documents annotés manuellement

FamilySearch HIP2013



Jeu de données : registres de mariages de la compétition HIP2013 FamilySearch

- Apprentissage : 7000 documents
  - Sans vérité terrain
- Test : 2000 documents annotés manuellement

Évaluation de la correspondance spatiale de zones

- Métrique introduite par Garris (Garris, 1995)
  - Recouvrement entre zone attendue et zone obtenue
- Reconnaissance complète
  - Au moins 95% de la largeur et au moins 75% de la hauteur
- Reconnaissance partielle
  - 80 à 95% de la largeur et au moins 75% de la hauteur

FamilySearch HIP2013 - Constitution de la pseudo vérité terrain

- 7000 documents
- 54 141 zones
- 276 actions utilisateur
  - Annotation manuelle (8 actions par document)
    $\rightarrow$ 34,5 documents annotés
  - Coût divisé par 200


ACTA DE MATRIMONIO 189

En Distrito Federal a las del dia de de mil novecientos , comparecen ante mi Oficial del Registro Civil, para contraer matrimonio bajo el régimen de los señores de acuerdo con la solicitud y documentos que presentaron con fecha los cuales contienen los siguientes datos:

DEL CONTRAYENTE DE LA CONTRAYENTE

FamilySearch HIP2013 - Résultats

Deux systèmes évalués :

- 4 modèles définis manuellement [Lemaitre 2013]
- 11 modèles inférés automatiquement

Modèle	1	2	3	4	5	6	7	8	9	10	11
Nombre	1448	822	740	652	566	470	359	123	92	33	25

FamilySearch HIP2013 - Résultats

Deux systèmes évalués :

- 4 modèles définis manuellement [Lemaitre 2013]
- 11 modèles inférés automatiquement

Modèle	1	2	3	4	5	6	7	8	9	10	11
Nombre	1448	822	740	652	566	470	359	123	92	33	25

		Modèles	
		Inférés	Manuels
Zone	Reconnaissance complète	91.4%	89.7%
	Reconnaissance partielle	6.2%	4.0%
	Manquant	2.4%	6.3%
Taux de reconnaissance du document		89.8%	78.9%

FamilySearch HIP2013 - Résultats

Deux systèmes évalués :

- 4 modèles définis manuellement [Lemaitre 2013]
- 11 modèles inférés automatiquement

Modèle	1	2	3	4	5	6	7	8	9	10	11
Nombre	1448	822	740	652	566	470	359	123	92	33	25

		Modèles	
		Inférés	Manuels
Zone	Reconnaissance complète	91.4%	89.7%
	Reconnaissance partielle	6.2%	4.0%
	Manquant	2.4%	6.3%
Taux de reconnaissance du document		89.8%	78.9%

- Amélioration des taux de reconnaissance
  - Inférence semi-automatique des modèles
  - Meilleure couverture du corpus
    - Construction rapide d'une pseudo vérité terrain

Conclusion

Méthode facilitant l'adaptation d'un système de reconnaissance de structure combinant

- Apprentissage statistique
- Système syntaxique
  - Pouvoir d'expression et gestion des cas rares
- Interaction avec l'utilisateur
  - Au cœur du système

Méthode EWO permet

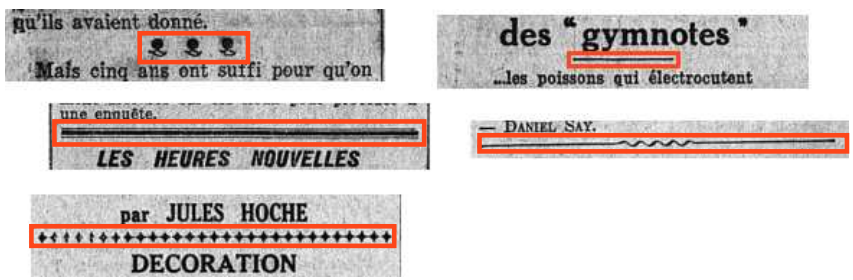
- Inférence des règles
  - Structures logique et physique
- Gestion de l'absence de vérité terrain
- Vision exhaustive et synthétique des données

Perspectives

- Extension des cadres applicatifs

Perspectives

- Extension des cadres applicatifs
  - Séparateurs d'articles d'archives de presse
  - Grammaire déjà existante



  - Grande variabilité
  - Création d'une pseudo vérité terrain
  - Utilisation d'éléments horizontaux combinant
    - OCR
    - Lignes de texte détectées avec une grammaire
    - Séparateurs déjà détectés

Perspectives

- Utilisation dans d'autres contextes
  - Autres méthodes de reconnaissance de structures
  - Construction rapide de bases étiquetées

Perspectives

- Utilisation dans d'autres contextes
  - Autres méthodes de reconnaissance de structures
  - Construction rapide de bases étiquetées
- Problématique de l'exhaustivité
  - Exploitation de très grandes quantités de données
  - Comportement de l'EAC clustering
  - Détection des cas rares
  - Minimisation des interventions utilisateur
    - Sélection des primitives
    - Sélection automatique des propriétés physiques