Interprétation contextuelle et assistée de fonds d'archives numérisés : Application à des registres de ventes du XVIIIe siècle

Joseph Chazalon

Direction : Bertrand Coüasnon et Jean Camillerapp

Collaborateurs : Laurent Guichard, Aurélie Lemaitre, Alejandro Toselli

intuidoc

Équipe

INSA Rennes

UMR IRISA

Organismes de recherche

UEB Université Européenne de Bretagne

PRES

Yvelines Conseil général

Archives Yvelines

Financement

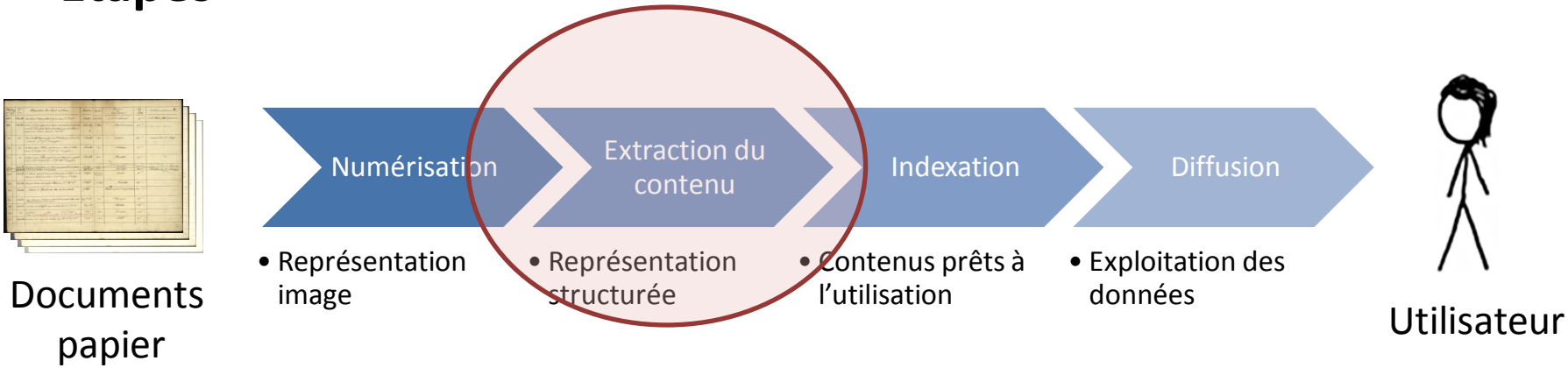
Contexte sociétal et scientifique

Dématérialisation de fonds d'archives

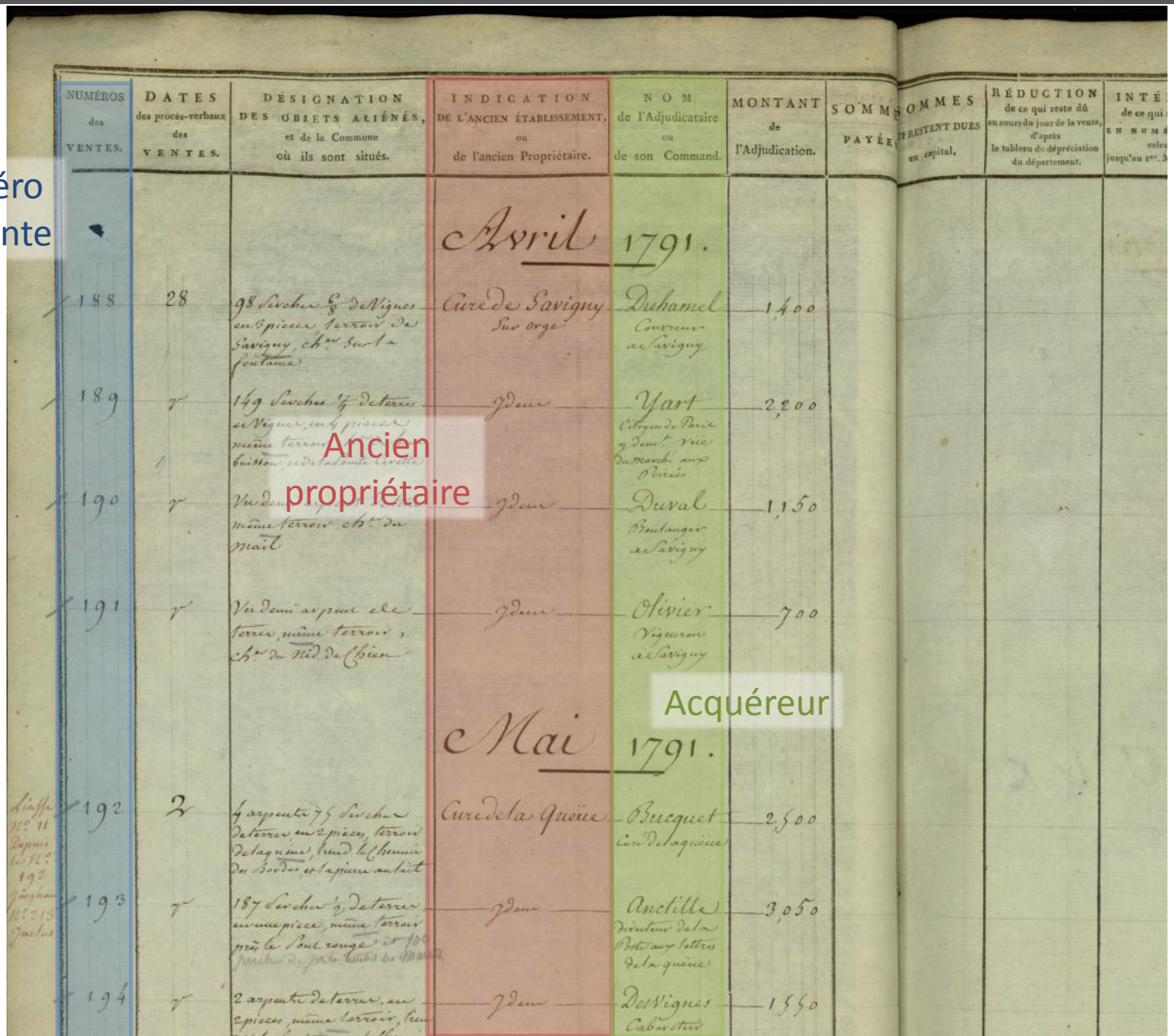
Avantages

- Protéger des contenus fragiles
- Diffuser des contenus rares
- Valoriser le patrimoine documentaire

Étapes



Exemple de données : séquestres révolutionnaires



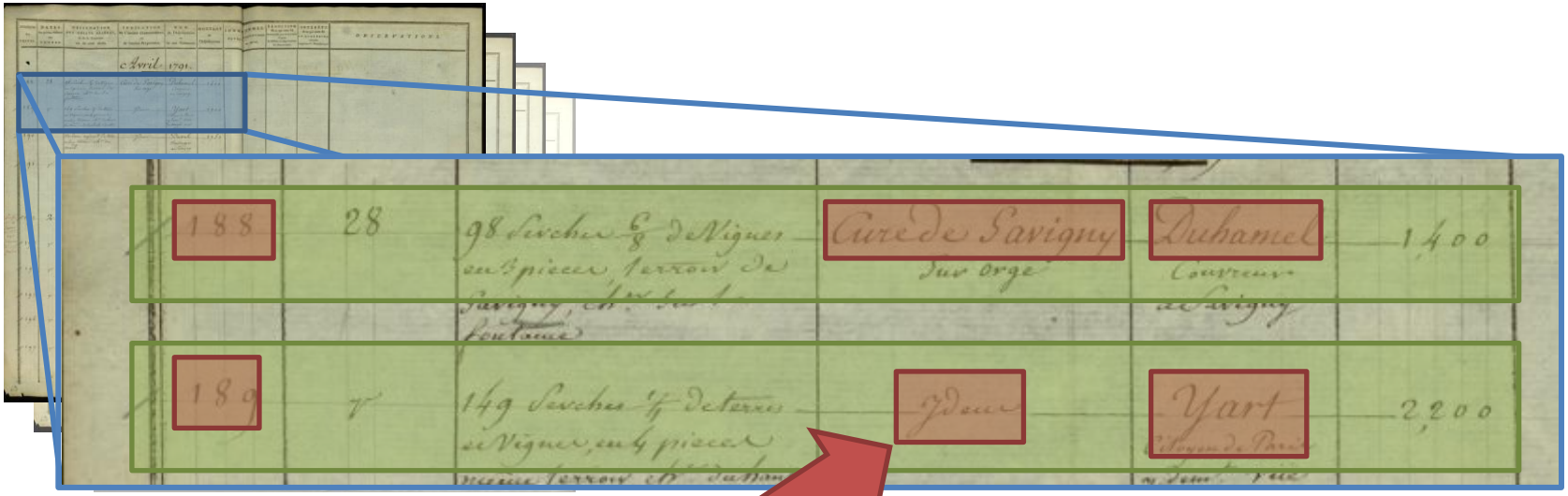
NUMÉROS des VENTES.	DATES des procès-verbaux des VENTES.	DÉSIGNATION DES OBJETS ALIÉNÉS, et de la Commune où ils sont situés.	INDICATION DE L'ANCIEN ÉTABLISSEMENT, ou de l'ancien Propriétaire.	NOM de l'Adjudicataire ou de son Command.	MONTANT de l'Adjudication.	SOMMES PAYÉES	SOMMES QUI RESTENT DUES en capital,	RÉDUCTION de ce qui reste dû au cours du jour de la vente, d'après le tableau de dépréciation du département.	INTÉ[illegible] de ce qui [illegible] EN NUM[illegible] calcu[illegible] jusqu'au 2[illegible]
			Avril	1791.					
188	28	98 perches 5/8 de Vignes en 3 pièces terroir de Savigny, ch.[illegible] sur la fontaine	Curé de Savigny sur orge	Duhamel Couvreur à Savigny	1,400				
189	7	149 perches 1/4 de terres à Vignes, en 4 pièces même terroir, buisson, et de la source verdée	7 deux	Yart Citoyen de Paris 7 deux rue du marché aux Poirées	2,200				
190	7	Un demi arpent [illegible] même terroir, ch.[illegible] du Mail	7 deux	Duval Boulanger à Savigny	1,150				
191	7	Un demi arpent de terres, même terroir, ch.[illegible] du Nid de Chien	7 deux	Olivier Vigneron à Savigny	700				
			Mai	1791.					
192	2	4 arpents 75 perches de terres en 2 pièces, terroir de la Queue, lieu le chemin des Bordes et la pièce en lait	Curé de la Queue	Bucquet Curé de la Queue	2,500				
193	7	187 perches 1/2 de terres en une pièce, même terroir près le Pont rouge	7 deux	Anclille directeur de la Poste aux lettres de la Queue	3,050				
194	7	2 arpents de terres en 2 pièces, même terroir, lieu	7 deux	Des Vignes Cabaretier	1,550				

Numéro de vente

Ancien propriétaire

Acquéreur

Exemple de besoin



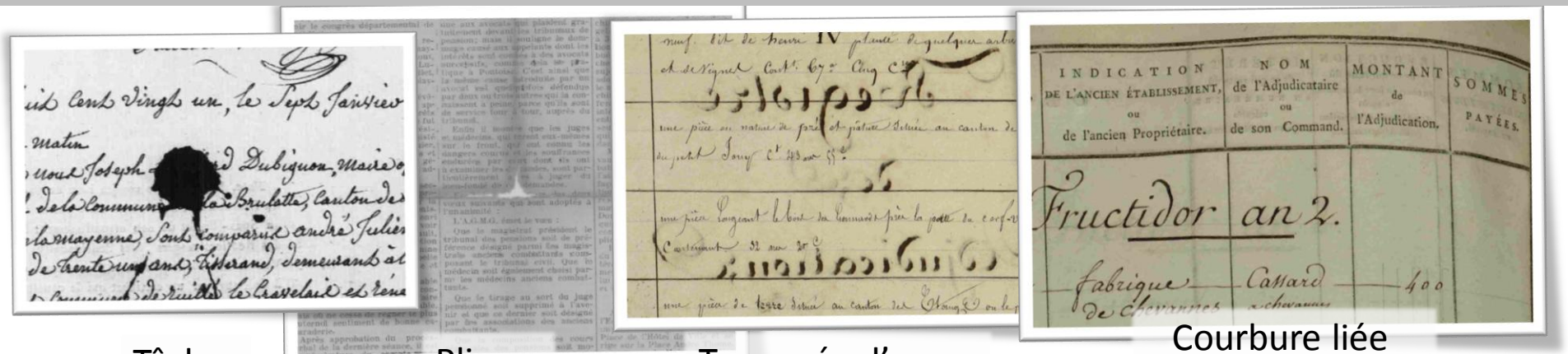
Extraction du contenu

- ➔ Localisation et reconnaissance
- ➔ Production de résultats prêts pour l'indexation
- ➔ Interprétation sémantique

Fichier image	Numéro vente	Ancien prop.	Acquéreur
00071.jpg	188	Cure de Savigny	Duhamel
00071.jpg	189	Cure de Savigny	Yart

Difficultés des fonds d'archives (1/2)

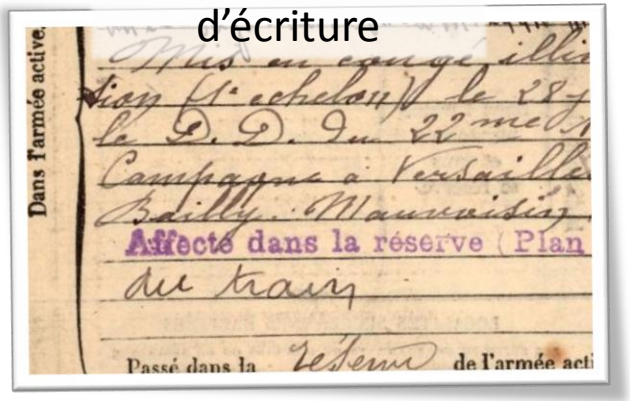
Dégradations : exemples



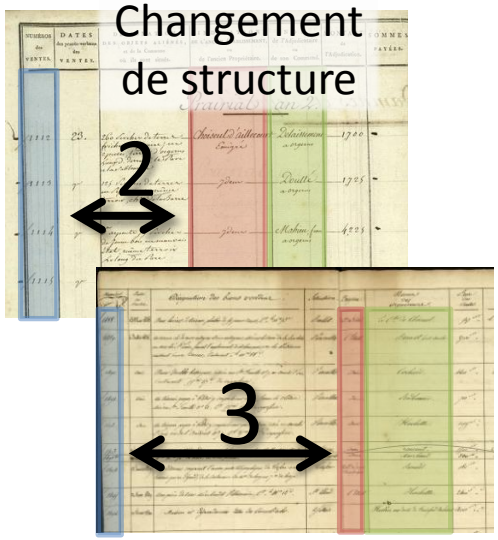
Tâches | Pliures | Traversée d'encre | Courbure liée à la numérisation

Variabilités : exemples

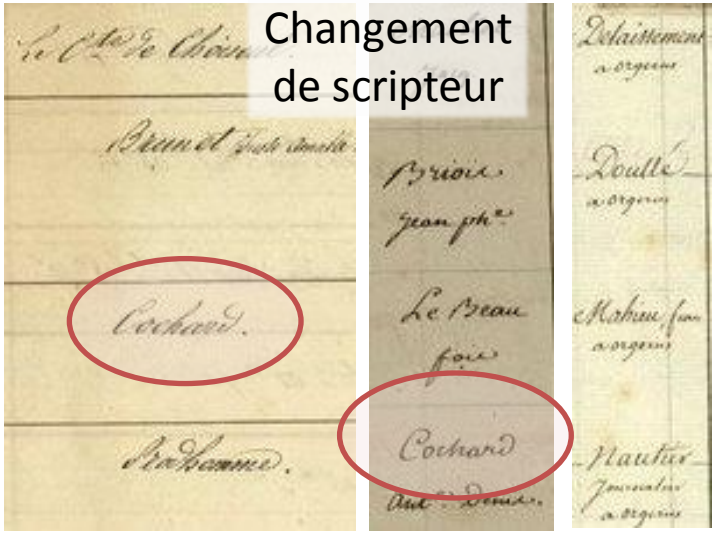
Mélange de styles d'écriture



Changement de structure



Changement de scripteur



Difficultés des fonds d'archives (2/2)

➔ Conséquences

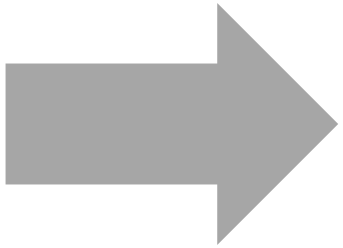
Dégradations

+

Variabilités

+

Cas particuliers

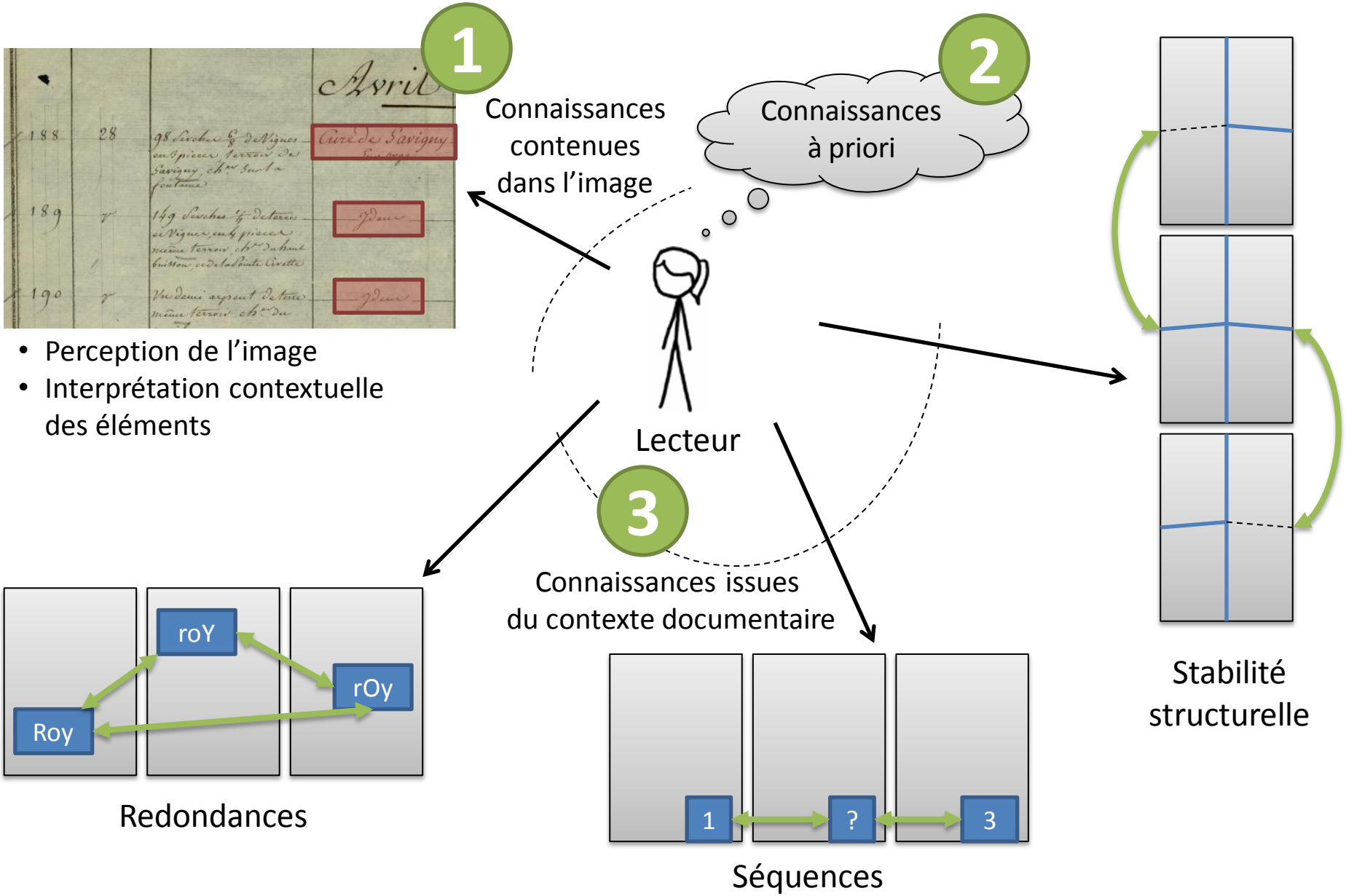


Ambiguïté

+

Impossibilité d'anticiper tous les cas

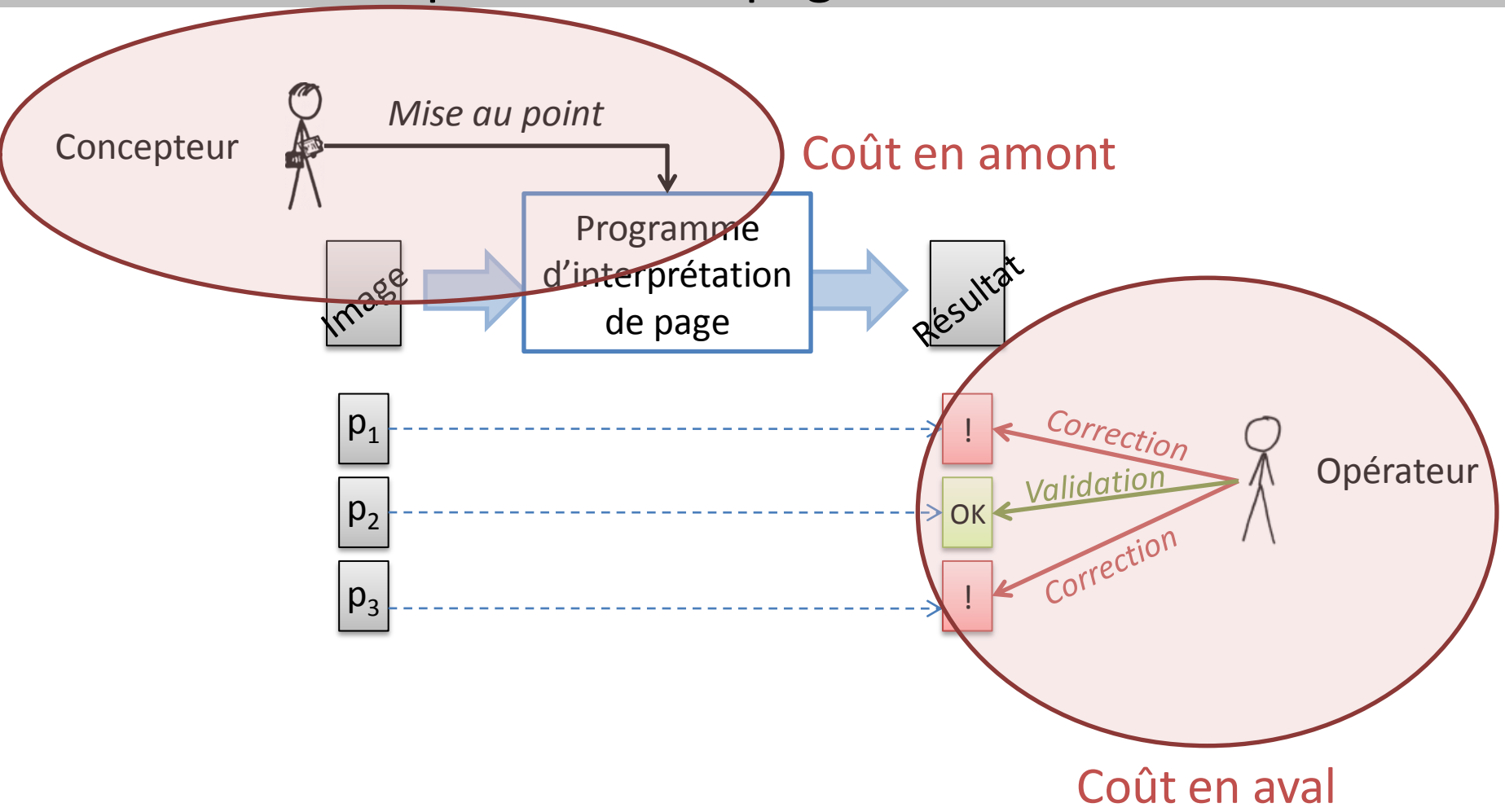
Quelles connaissances les humains utilisent-ils ?



- Perception de l'image
- Interprétation contextuelle des éléments

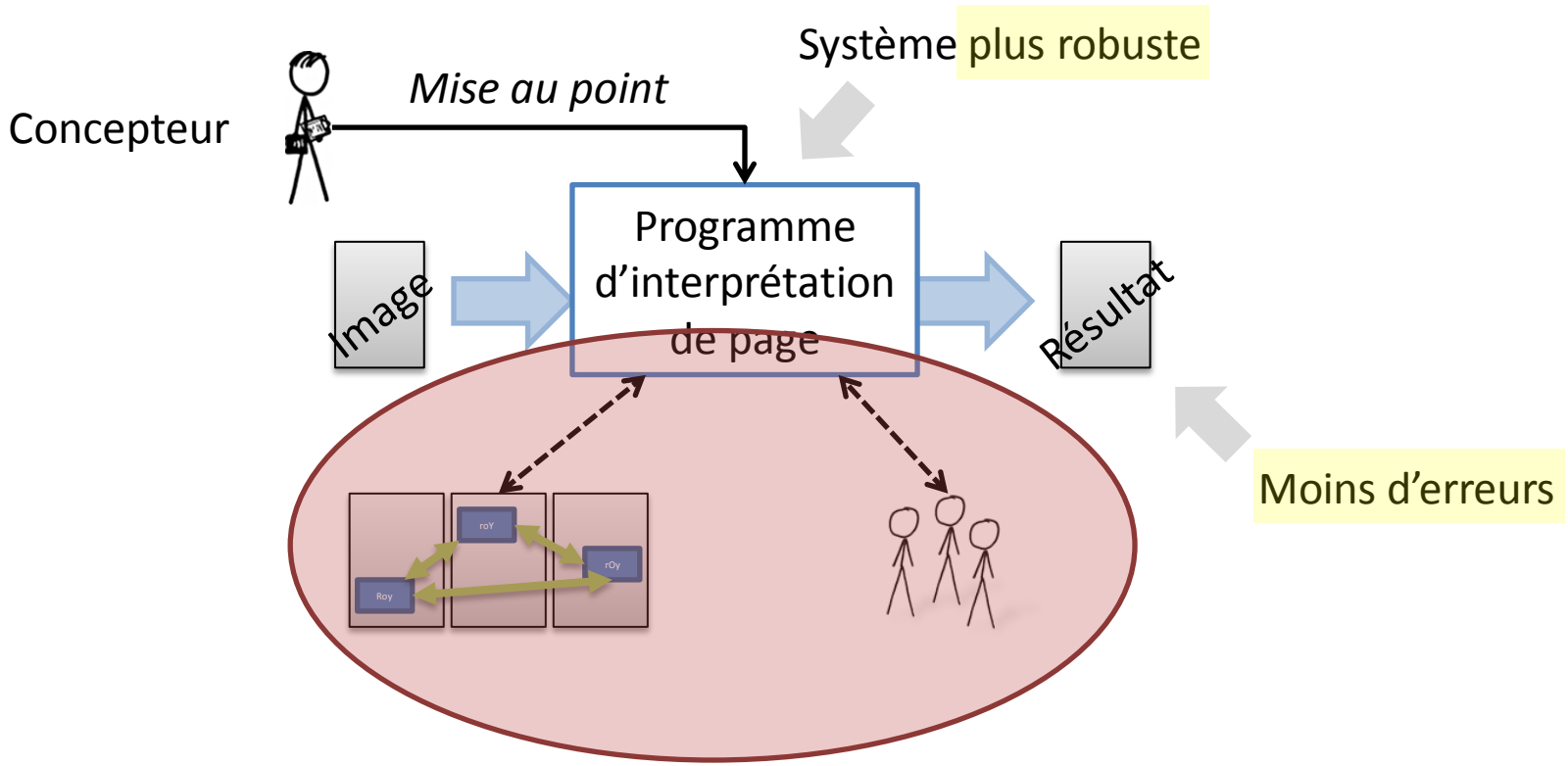
Interprétation automatique de fonds documentaires

Limites de l'interprétation de pages isolées



Notre objectif

Intégrer plus de connaissances lors de l'interprétation



Interdépendance + humains

➔

Apport asynchrone d'informations externes

Plan

Formalisation des connaissances (1/2)

Approches algorithmiques

[Tang94, O'Gorman95, Clavier04, Shafait06]

Concepteur ➔ programme

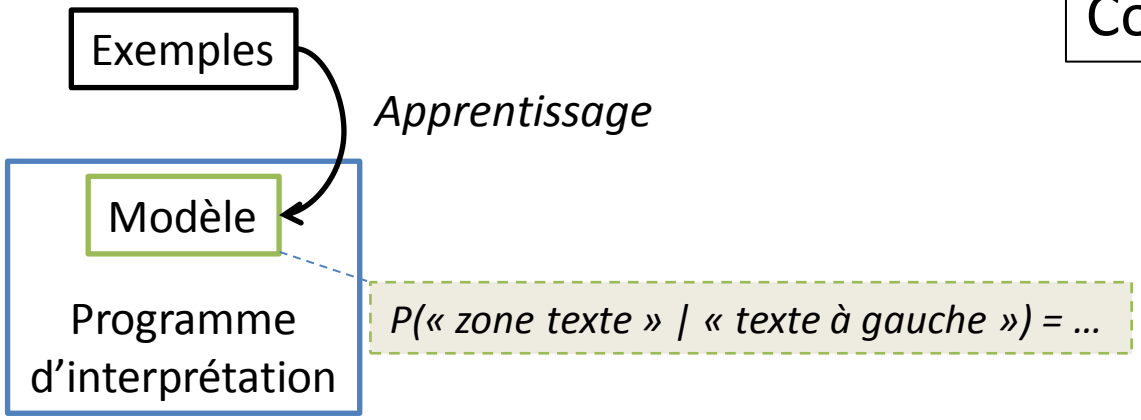
```
…
image.binariser(SEUIL);
couche1 = image.extraireComposantes();
couche1.appelOCR(LEXIQUE);
…
```

☹ Trop figées

Approches statistiques

[LeBourgeois01, Montreuil09]

Concepteur ➔ entraînement



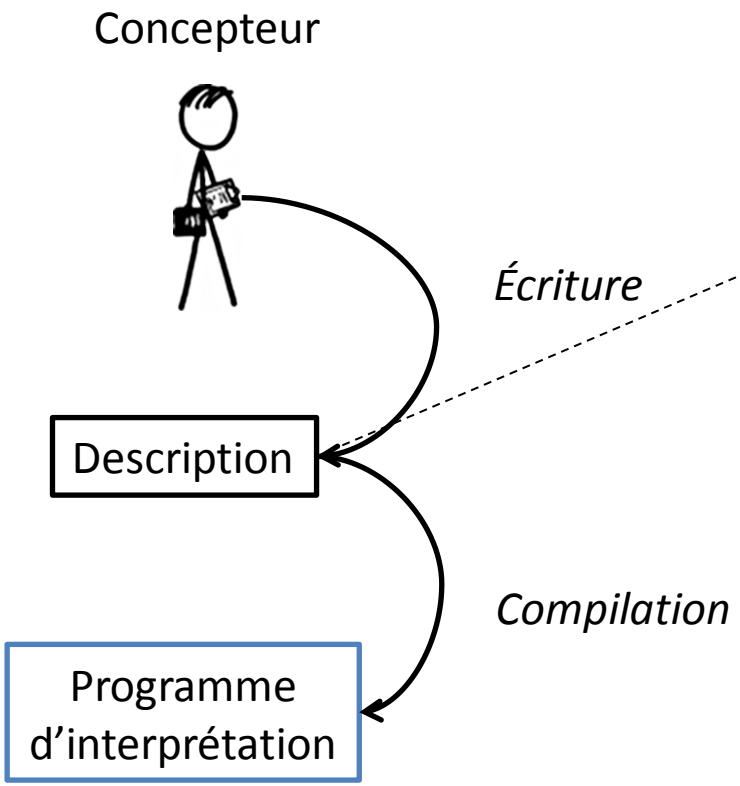
☹ Amorçage difficile

☹ Expressivité limitée

Formalisation des connaissances (2/2)

Approches déclaratives

[Vaxivere92, Tang95, Pasternak95, Couasnon96, Marriott98]




```
courrier → @(hautPage)
           expediteur
           @(sous expediteur)
           objet
           …
```

Concepteur ➔ description

☺ Extensibles

mais

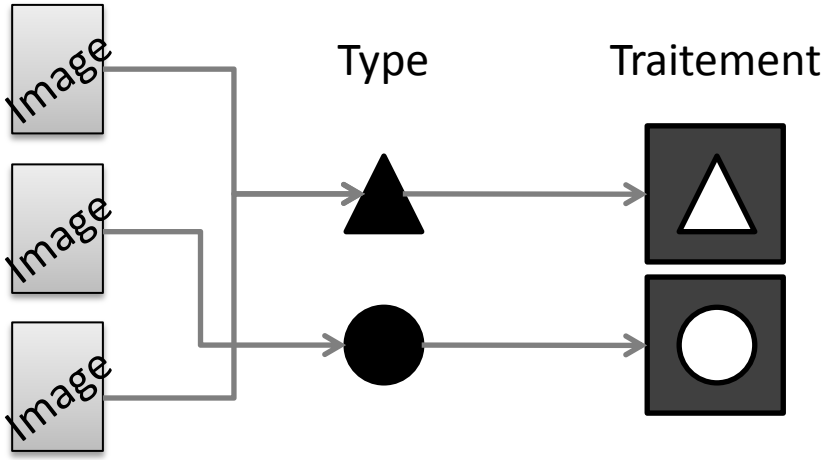
☹ Sensibles aux dégradations et à la variabilité

Plan

Exploitation du contexte documentaire

Approches avec classification du type de document

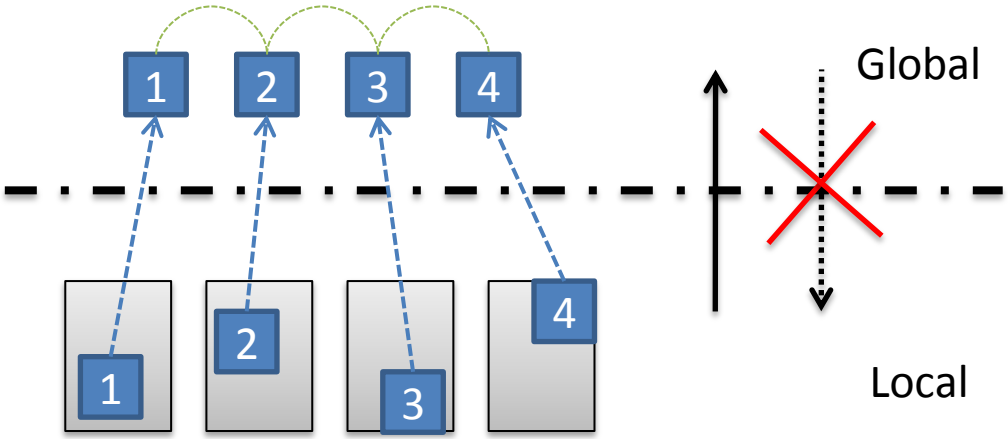
[Klein04, Lamiroy12]



☹ Pas d'exploitation globale des données

Approches à interprétation globale

[Lin97, Saund11, Xui12]



☺ Fiabilisation des résultats locaux

mais

☹ Pas de réinjection locale de l'information

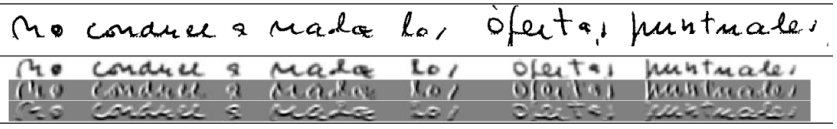
Plan

Interaction avec des opérateurs humains

Traitements initiés par l'humain

[Bapst96, Ramel07, Vidal08, Llados08]

Humain ➔ Détecte les erreurs

	Line Img	
	x	
INTER-0	(p)	()
INTER-1	$(\hat{s})$	*(no conduce o mala los afectos eventuales)*
	$(\hat{s}_p)$	(no conduce)
	(c)	**(a)**
	(p)	(no conduce a)
INTER-2	$(\hat{s})$	*(nada los afectos eventuales)*
	$(\hat{s}_p)$	(nada)
	(c)	**(las)**
	(p)	(no conduce a nada las)

☺ Résultat optimal en théorie

mais

☹ Interaction synchrone

☹ Détection manuelle des erreurs

Traitements initiés par la machine

[Ogier98, Robadey01, Klein04]

Humain ➔ Répond aux sollicitations

☺ Sollicitation asynchrone parcimonieuse

mais

☹ Détection d'erreur automatique faillible

Positionnement de notre approche

Formalisation des connaissances

➔ Extension des approches déclaratives

Exploitation du contexte documentaire

➔ Interprétation globale

➔ Ajout d'un mécanisme de réintégration (fusion)
global → local

Interaction avec des opérateurs humains

➔ Traitements initiés par la machine (questions)
mais possibilité d'accepter un guidage par l'humain

➔ Gestion automatique d'échanges asynchrones

Plan

Interprétation itérative (1/4)

Traitement page par page

Image

Module d'interprétation de page

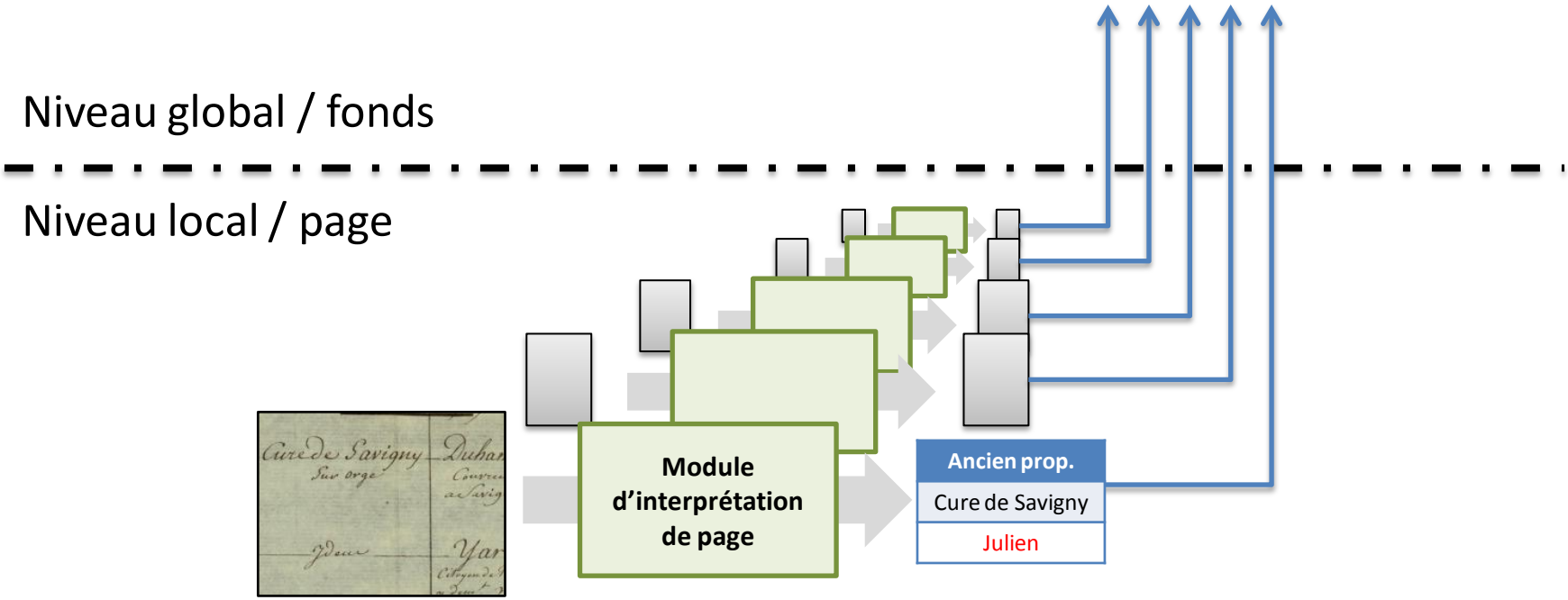
Ancien prop.
Cure de Savigny
Julien

Résultat

Interprétation itérative (2/4)

Collecte globale des résultats

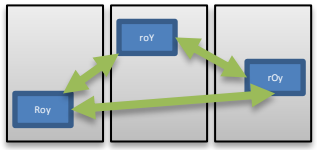
1. Traitement
2. Déchargement programme
3. Collecte des résultats



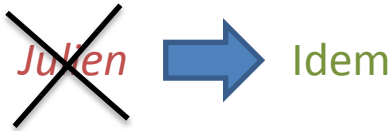
Interprétation itérative (3/4)

Interprétation globale

Exploitation du contexte documentaire



Interaction avec des opérateurs humains



Niveau global / fonds

Niveau local / page

Ancien prop.
Cure de Savigny
Julien

Résultat

Résultat

Résultat

Résultat

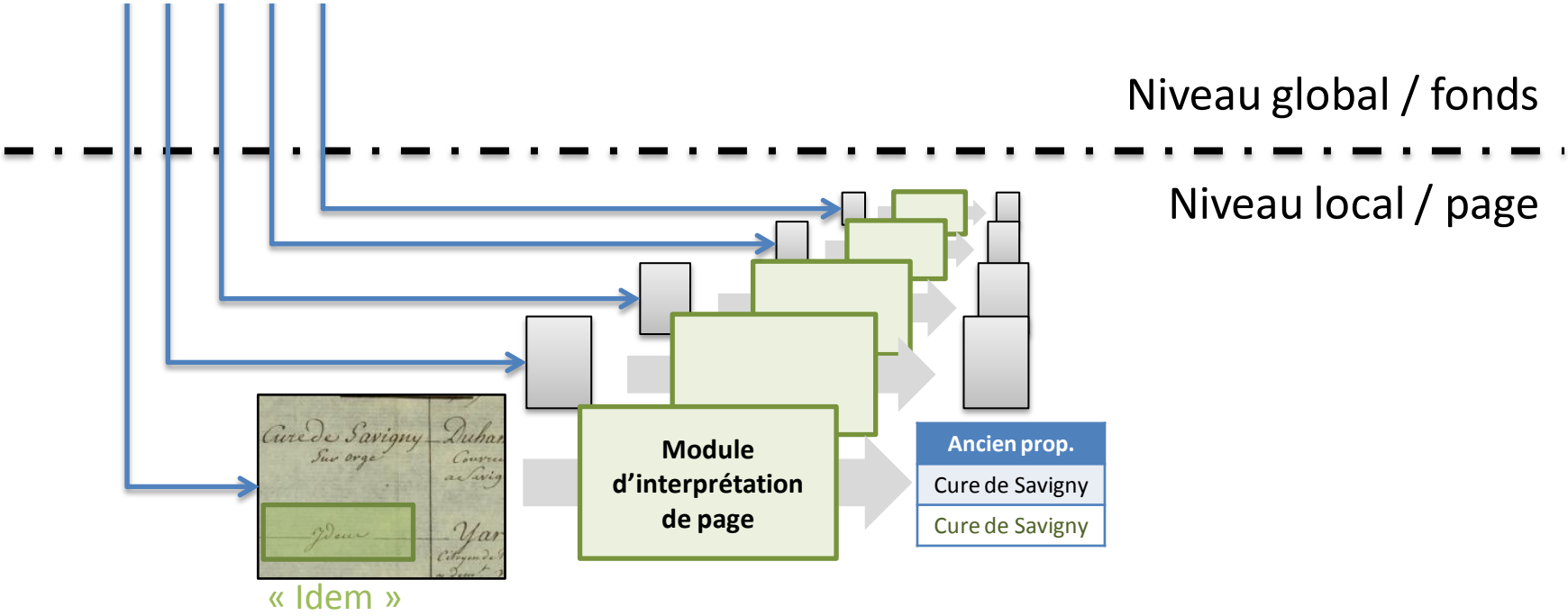
Résultat

Résultat

Interprétation itérative (4/4)

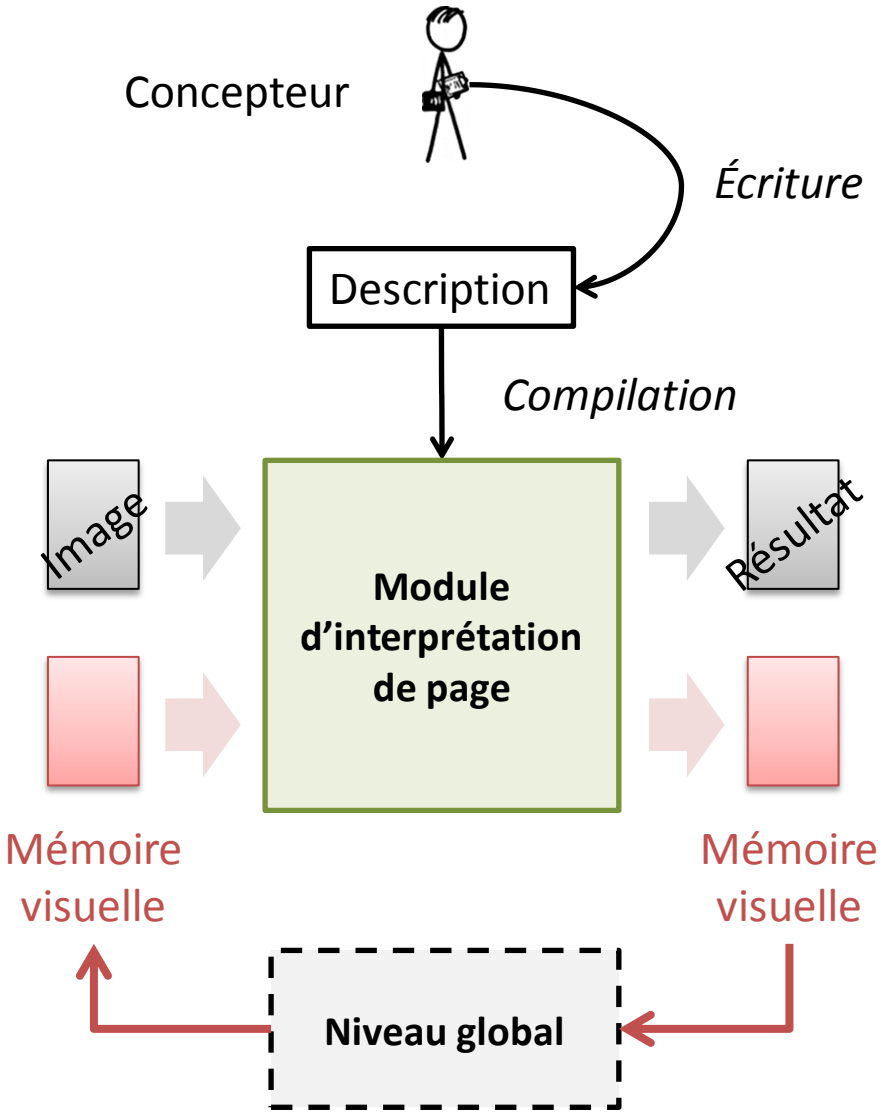
Réinterprétation locale avec des informations globales

1. Nouvelle interprétation complète avec les données externes
2. Production de nouveaux résultats



Notre mise en œuvre d'une approche itérative (1/11)

Adapter une approche déclarative



Extension en 3 étapes

1. **Mémoire visuelle**
   ➔ Support de la communication avec l'environnement

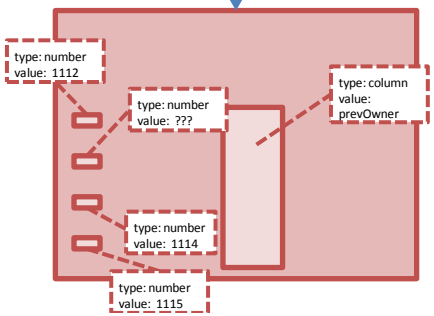
2. **Langage de description**
   ➔ Introduire de nouveaux opérateurs d'interaction

3. **Architecture à deux niveaux**
   ➔ Niveau global : collecte et distribue les données
   ➔ Niveau local : produit et exploite les données

Notre mise en œuvre d'une approche itérative (2/11)

Mémoire visuelle

Niveau global



Mémoire Visuelle

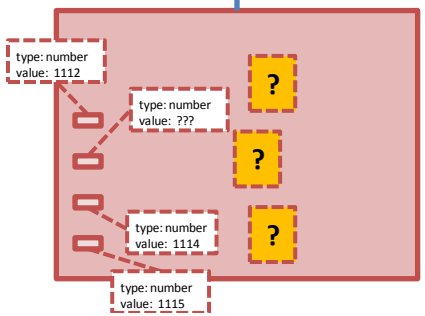
Propriétés

- Donnée ← Localisation
- 1 mémoire / page
- Modifiée localement et globalement

Modification externe

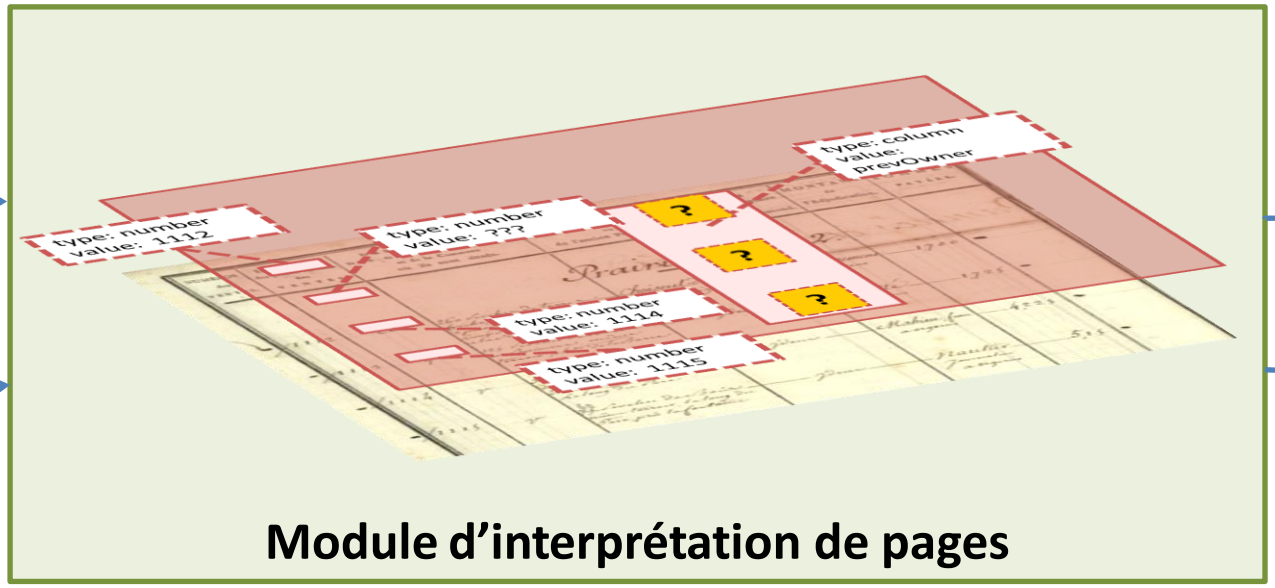
➔

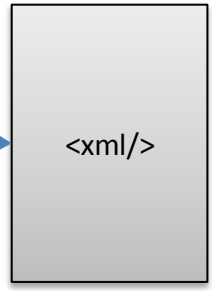
Changement du comportement local



Mémoire Visuelle

Image de page

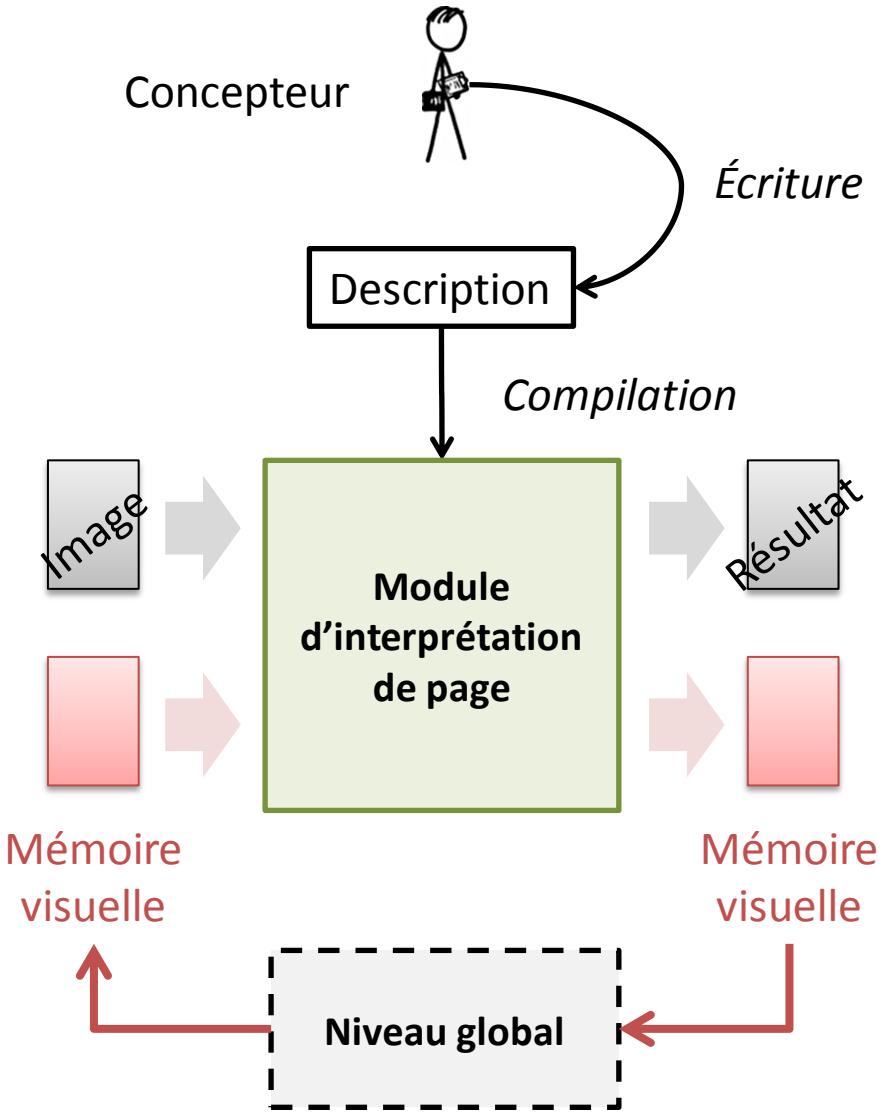




Résultat

Notre mise en œuvre d'une approche itérative (3/11)

Adapter une approche déclarative



Extension en 3 étapes

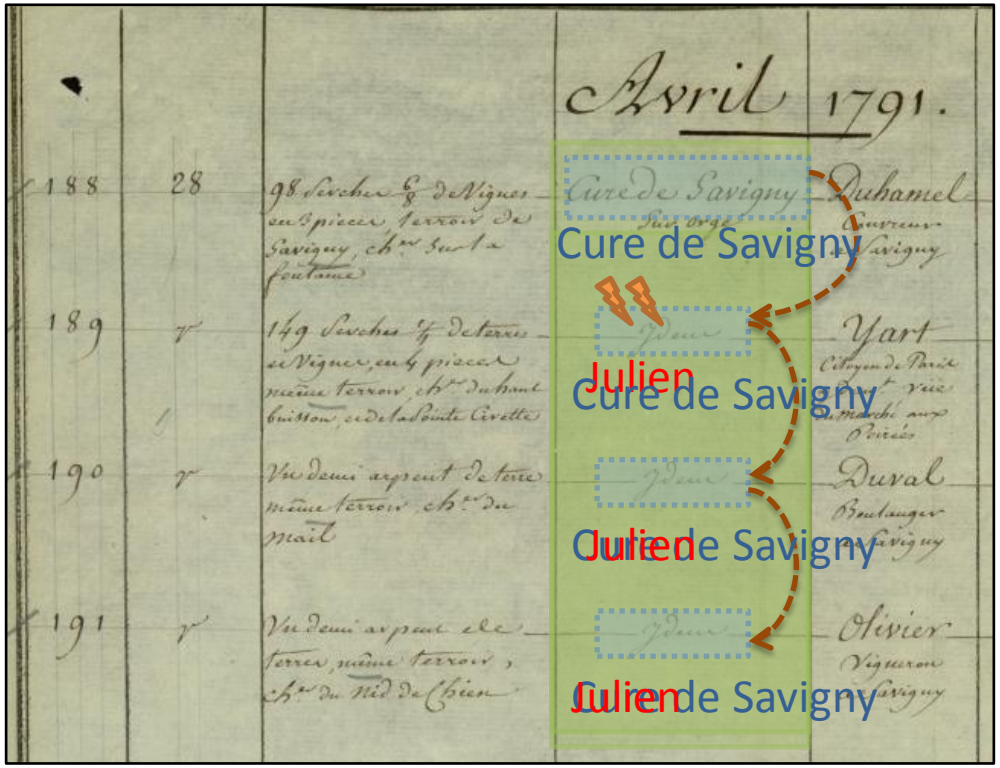
1. **Mémoire visuelle**
   ➔ Support de la communication avec l'environnement

2. **Langage de description**
   ➔ Introduire de nouveaux opérateurs d'interaction

3. **Architecture à deux niveaux**
   ➔ Niveau global : collecte et distribue les données
   ➔ Niveau local : produit et exploite les données

Notre mise en œuvre d'une approche itérative (4/11)

Exemple de description



Description SANS interaction

```
page →
  @(colonneAncienProp)
  lireTousLesNoms("inconnu")

lireTousLesNoms(precedent)→
  @(detecterNom)
  nom = lireNom(precedent)
  @(dessous)
  LOOP(lireTousLesNoms(nom))

lireNom(precedent) →
  reco = reconnaitNom
  res = (si reco=="IDEM"
         alors precedent
         sinon reco)
```

Sans interaction : résultat peu fiable ➔ propagation d'erreur

Notre mise en œuvre d'une approche itérative (5/11)

Nouveaux opérateurs de description

Détection des erreurs

RAISE_QUESTION(Question, Type) :
Ajouter position → (Question, Type)
en mémoire visuelle
Poursuivre (éventuellement) l'interprétation

Reprise sur erreur

CATCH(Règle, Type) :
Reprendre l'interprétation si une information de type Type est demandée dans Règle

Fusion d'informations externes

TRY(Règle, Type) :
Si ∃ donnée : Type ∈ mémoire visuelle,
alors l'utiliser directement
sinon appeler Règle
et conserver le résultat en mémoire visuelle

Mécanisme proche de la gestion d'exceptions

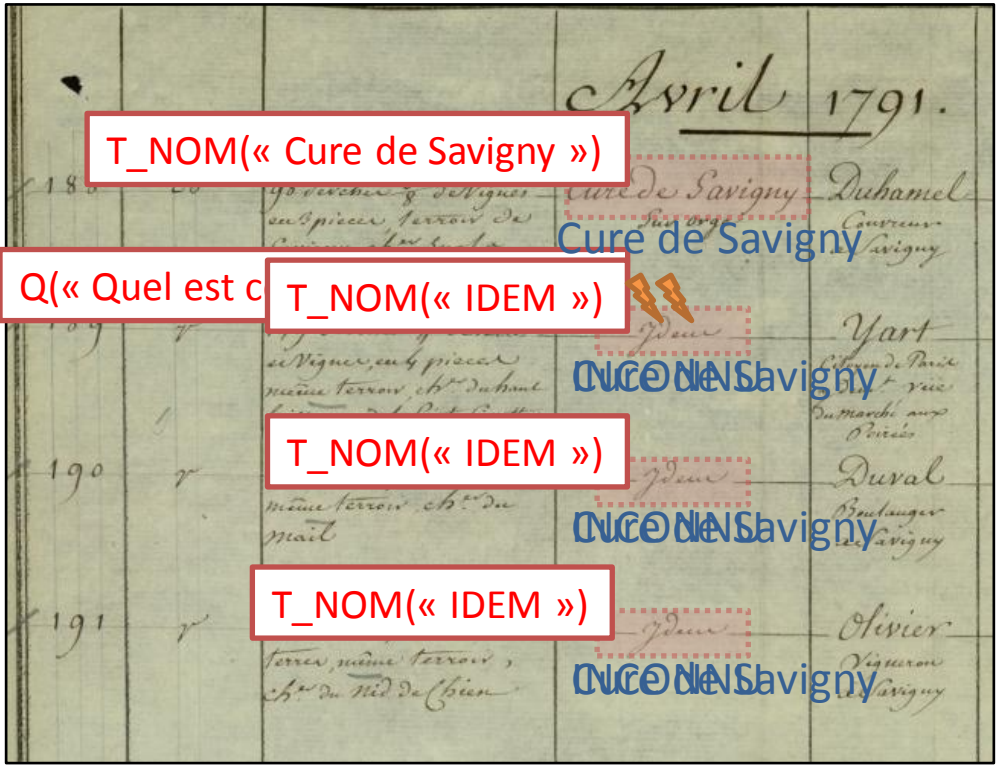
Description AVEC interaction

```
page →
  @(colonneAncienProp)
  lireTousLesNoms("inconnu")

lireTousLesNoms(precedent)→
  @(detecterNom)
  nom = CATCH(lireNom(precedent),
              T_NOM)
  @(dessous)
  LOOP(lireTousLesNoms(nom))

lireNom(precedent) →
  reco = TRY(reconnaitNomINT, T_NOM)
  res = (si reco=="IDEM"
         alors precedent
         sinon reco)

reconnaitNomINT →
  reconnaitNom
  ou RAISE_QUESTION("Quel est le nom
    de cet ancien propriétaire ?",
    T_NOM)
```

Notre mise en œuvre d'une approche itérative (6/11)

Progression au cours des itérations

Itération 1 :	Traitement	Interaction asynchrone
Itération 2 :	Traitement	…



Avec interaction :
correction des erreurs au plus tôt
➔ propagation limitée

Description AVEC interaction

```
page →
  @(colonneAncienProp)
  lireTousLesNoms("inconnu")

lireTousLesNoms(precedent)→
  @(detecterNom)
  nom = CATCH(lireNom(precedent),
                T_NOM)
  @(dessous)
  LOOP(lireTousLesNoms(nom))

lireNom(precedent) →
  reco = TRY(reconnaitNomINT, T_NOM)
  res = (si reco=="IDEM"
         alors precedent
         sinon reco)

reconnaitNomINT →
  reconnaitNom
  ou RAISE_QUESTION("Quel est le nom
     de cet ancien propriétaire ?",
     T_NOM)
```

Notre mise en œuvre d'une approche itérative (7/11)

Exemple d'échange entre le niveau local et le niveau global

Itération 1 : traitement

Niveau global

Module d'interprétation de page

Traiter **page 1**

{}

(retour)

```
zone1 → T_NOM("CURE_DE_SAVIGNY")
zone2 → Q("Quel est ce nom?", T_NOM)
zone3 → T_NOM("IDEM")
zone4 → T_NOM("IDEM")
```

Module d'interprétation de page

Traiter **page 2**

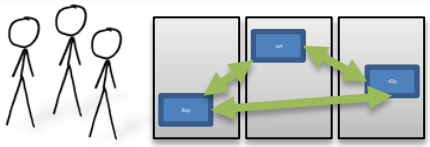
{}

(retour)

```
zone1 → T_NOM("PIGIS")
…
```

Itération 1 : interaction asynchrone

Niveau global



Répondre questions

```
page1,zone2 → Q("Quel est ce nom?", T_NOM)
page3,zone1 → Q("Quel est ce nom?", T_NOM)
page3,zone2 → Q("Quel est ce nom?", T_NOM)
page4,zone1 → Q("Quel est ce nom?", T_NOM)
```

(retour)

```
page1,zone2 → T_NOM("IDEM")
page3,zone1 → T_NOM("BRUNEL")
page3,zone2 → T_NOM("COCHARD")
page4,zone1 → T_NOM("IDEM")
```

Notre mise en œuvre d'une approche itérative (8/11)

Exemple d'échange entre le niveau local et le niveau global

Itération 2 : traitement — *etc.*

Niveau global

Module d'interprétation de page

Traiter **page 1**

```
zone1 → T_NOM("CURE_DE_SAVIGNY")
zone2 → T_NOM("IDEM")
zone3 → T_NOM("IDEM")
zone4 → T_NOM("IDEM")
```

(retour)

```
zone1 → T_NOM("CURE_DE_SAVIGNY")
zone2 → T_NOM("IDEM")
zone3 → T_NOM("IDEM")
zone4 → T_NOM("IDEM")
```

Module d'interprétation de page

Traiter **page 3**

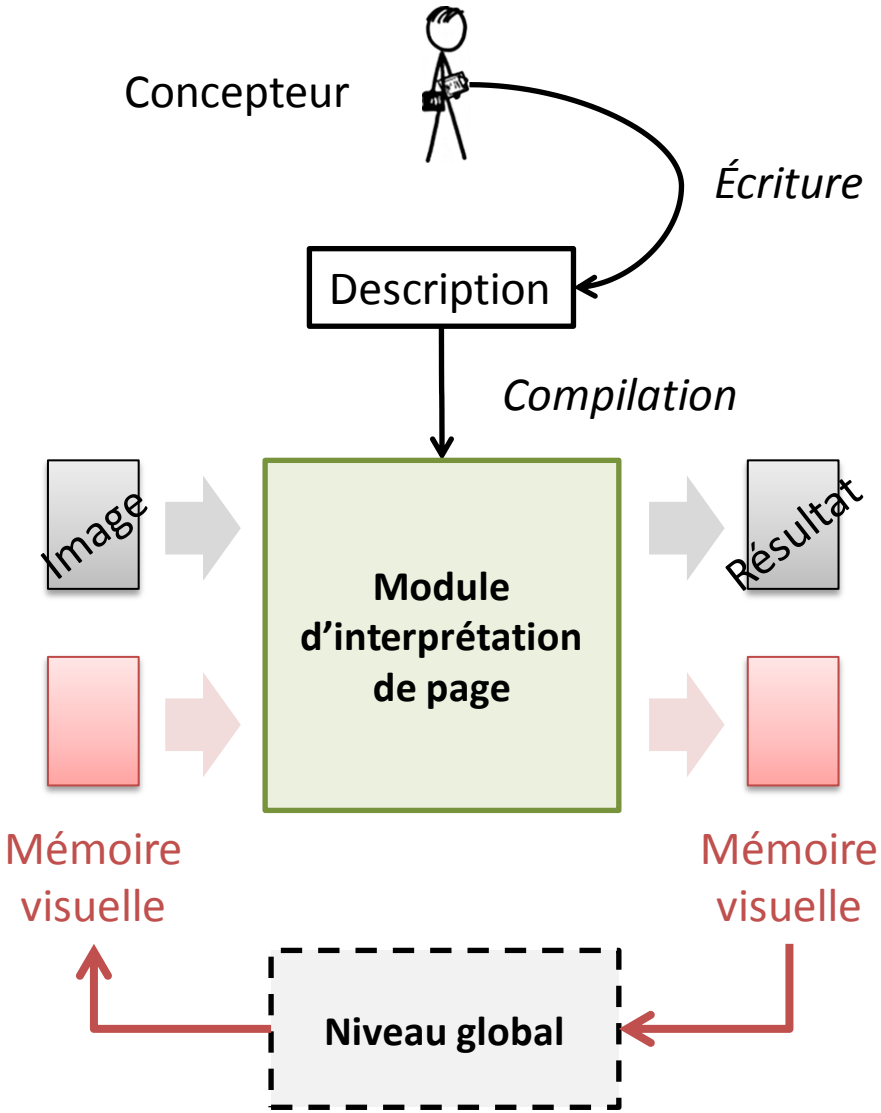
```
page3,zone1 → T_NOM("BRUNEL")
page3,zone2 → T_NOM("COCHARD")
```

Résultat associé

Ancien prop.
Cure de Savigny
Cure de Savigny
Cure de Savigny
Cure de Savigny

Notre mise en œuvre d'une approche itérative (9/11)

Adapter une approche déclarative



Extension en 3 étapes

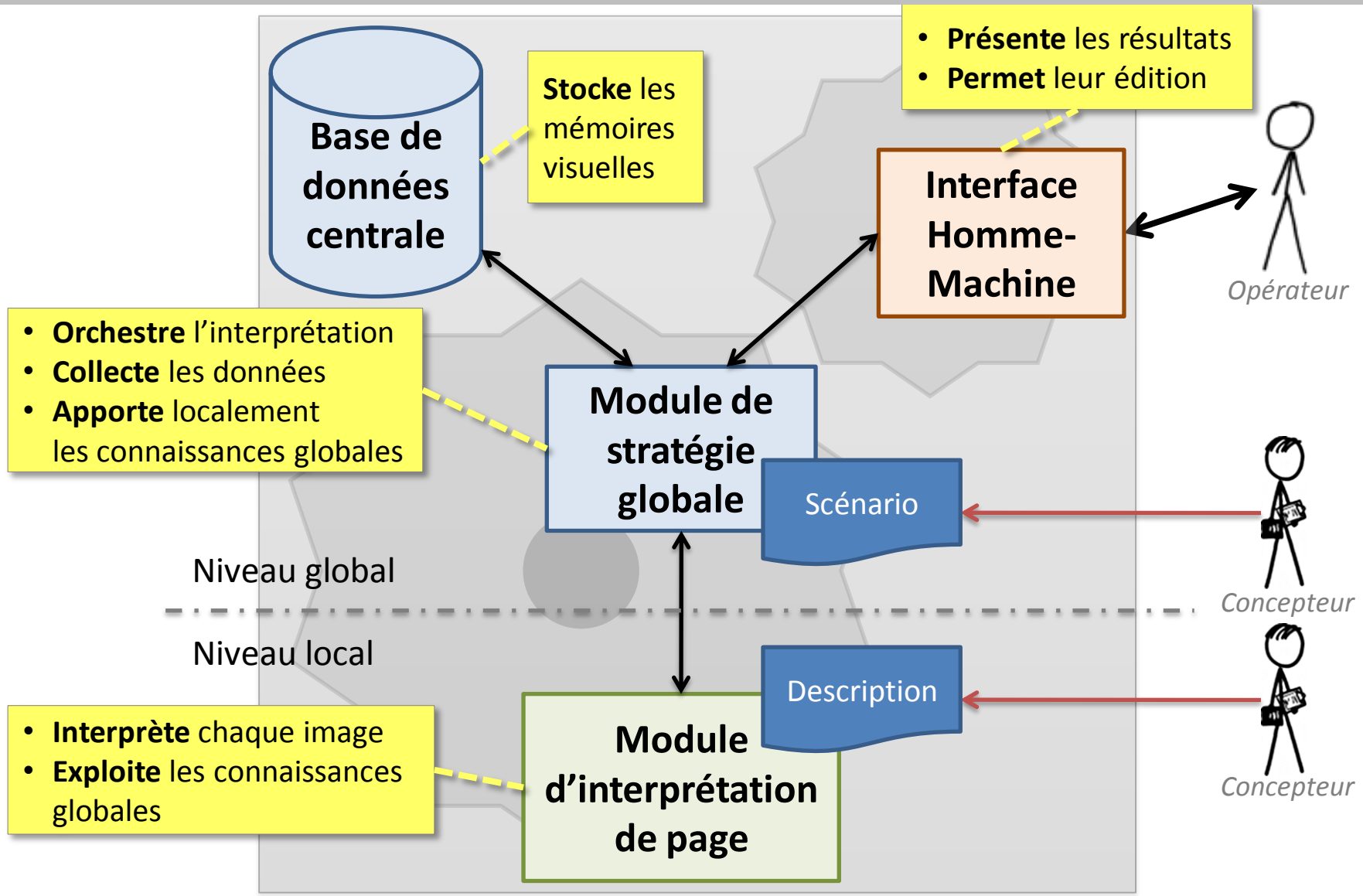
1. **Mémoire visuelle**
   ➔ Support de la communication avec l'environnement

2. **Langage de description**
   ➔ Introduire de nouveaux opérateurs d'interaction

3. **Architecture à deux niveaux**
   ➔ Niveau global : collecte et distribue les données
   ➔ Niveau local : produit et exploite les données

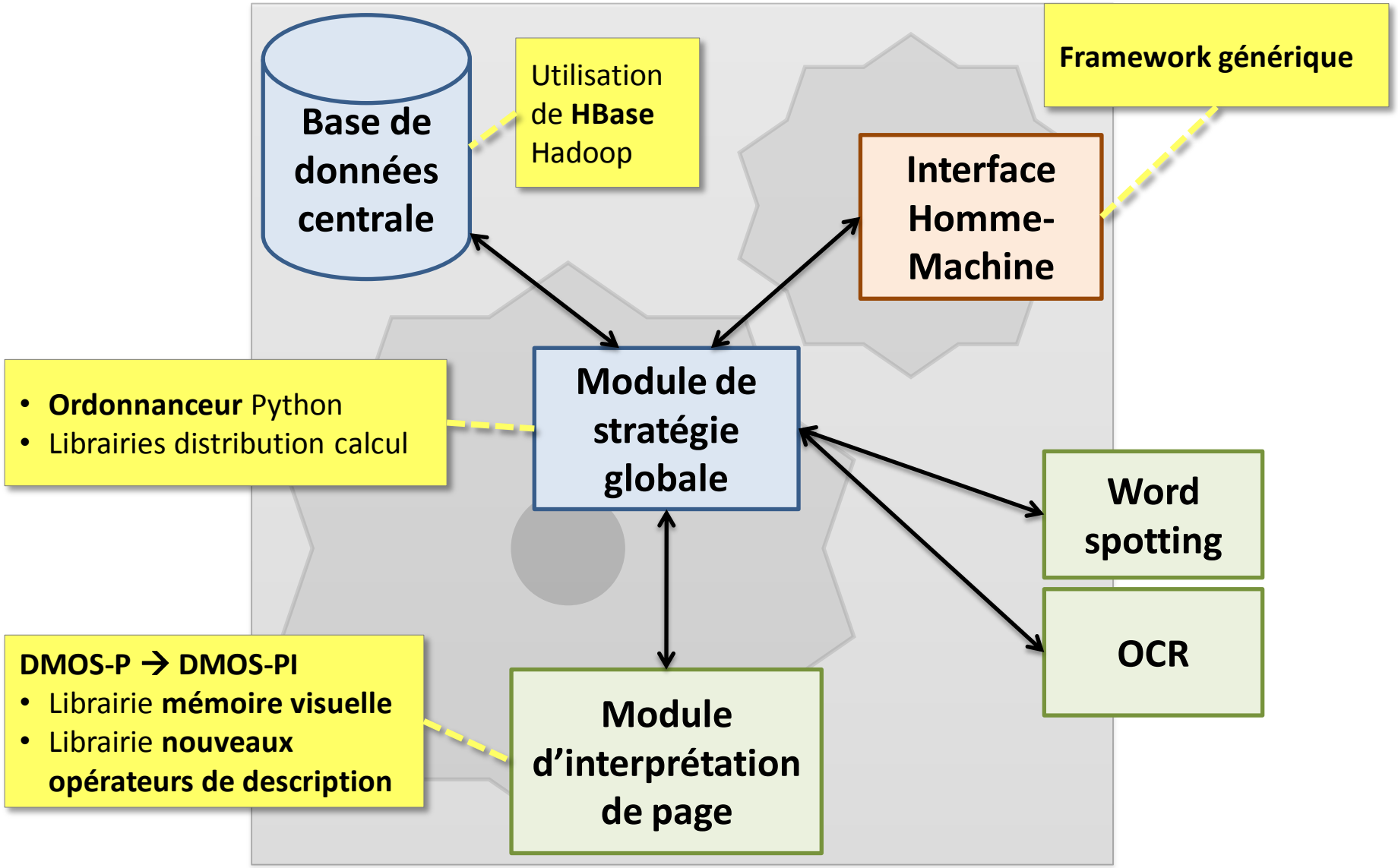
Notre mise en œuvre d'une approche itérative (10/11)

Architecture à deux niveaux



Notre mise en œuvre d'une approche itérative (11/11)

Implémentation de notre architecture



Bilan des contributions

Méthode pour rendre une approche déclarative itérative

1. Mémoire visuelle
2. Trois nouveaux opérateurs d'interaction
3. Architecture à deux niveaux

Avantages

Automatisation

- Au niveau local
  - Fusion des données produites par l'environnement
  - Échange asynchrone d'informations
- Au niveau global
  - Collecte et circulation de l'information
  - Interprétation par itération

Conception simple

- Description de la page
  - Comme si l'information externe était déjà disponible
  - Prévoir la gestion de l'incertain
- Définition d'un scénario global
  - Enchaîne les traitements
  - Ne se préoccupe pas de la production locale des données

Plan

Que valide-t-on ?

Intérêt des outils proposés

- Conception simple
- Possibilité de comparer différents scénarios

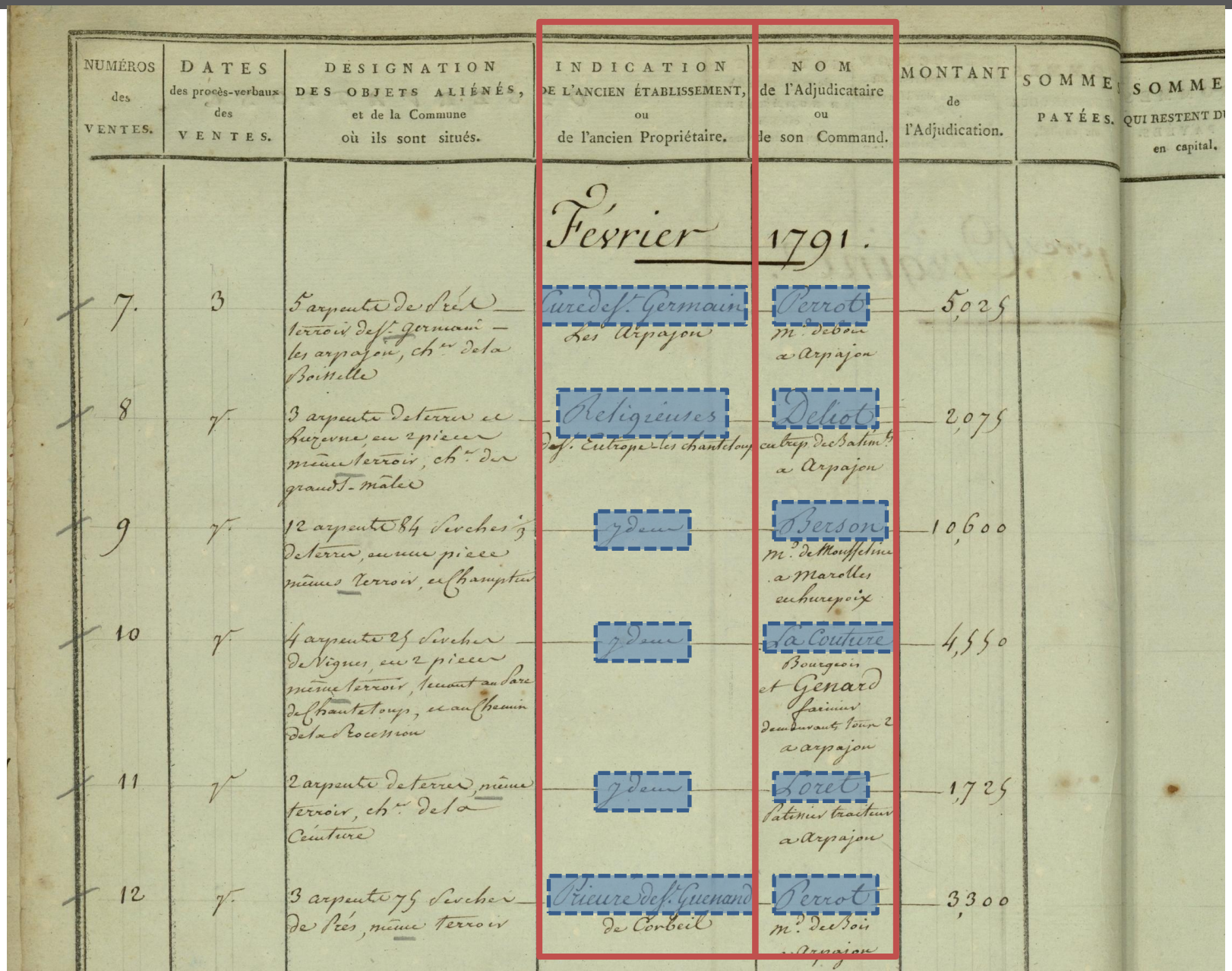
Intérêt d'une interprétation contextuelle et assistée

- Gain en qualité ou en coût de correction
- Évaluation du coût associé à un scénario délicat
  - Ne pas être dépendant de l'ergonomie des outils de correction
  - Éviter de comparer des actions de types différent

  ➔ Isoler des fragments de scénarios réels

Plan

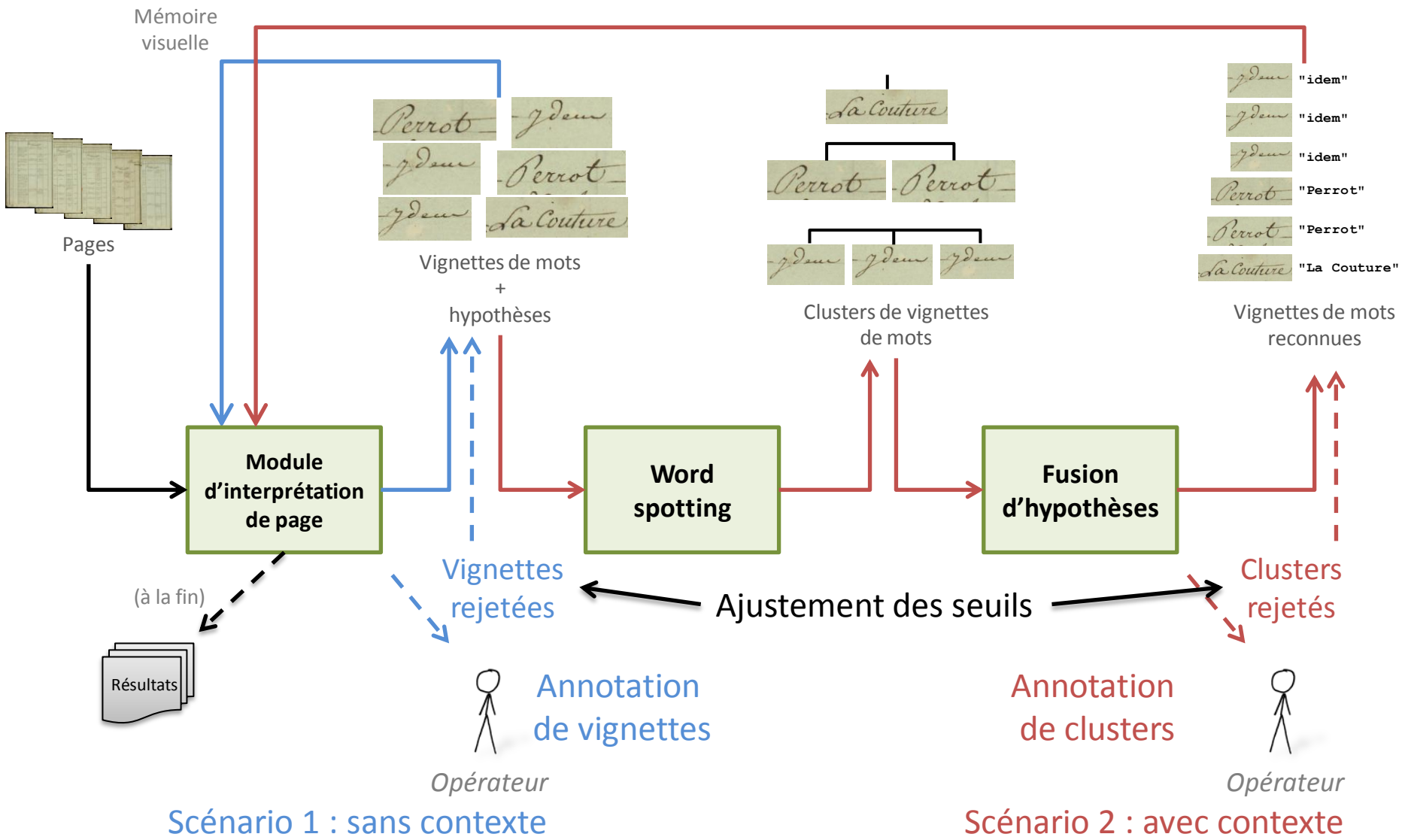
Expérimentation 1 : transcription de patronymes (1/4)



NUMÉROS des VENTES.	DATES des procès-verbaux des VENTES.	DÉSIGNATION DES OBJETS ALIÉNÉS, et de la Commune où ils sont situés.	INDICATION DE L'ANCIEN ÉTABLISSEMENT, ou de l'ancien Propriétaire.	NOM de l'Adjudicataire ou de son Command.	MONTANT de l'Adjudication.	SOMME PAYÉES.	SOMME QUI RESTENT DU en capital.
			Février 1791.				
7.	3	5 arpents de Pré terroir de St. Germain les Arpajon, ch.n de la Boisselle	Cure de St. Germain Les Arpajon	Perrot M.d de bois a Arpajon	5,025		
8	7.	3 arpents de terre et bruyere, en 2 pieces même terroir, ch.n du grand-Mâle	Religieuses de St. Eutrope les Chantetoup	Deliot entrep.r des Batim.ts a Arpajon	2,075		
9	7.	12 arpents 84 perches ½ de terre, en une piece mêmes terroir, en Champtier	7 deux	Berson M.d de Mouffetine a Marolles en Hurepoix	10,600		
10	7.	4 arpents 25 perches de Vignes, en 2 pieces même terroir, tenant au parc de Chantetoup, et au chemin de la Roche-Mion	7 deux	La Couture Bourgeois et Genard fermier demeurant tous 2 a Arpajon	4,550		
11	7.	2 arpents de terre, même terroir, ch.n de la Ceinture	7 deux	Loret Patissier traiteur a Arpajon	1,725		
12	7.	3 arpents 75 perches de Prés, même terroir	Prieure de St. Guenand de Corbeil	Perrot M.d de bois a Arpajon	3,300		

Expérimentation 1 : transcription de patronymes (2/4)

Comparaison de deux scénarios



Expérimentation 1 : transcription de patronymes (3/4)

Données traitées : environ 11 000 vignettes de patronymes

Évaluation

Approche	Taux d'annotation manuelle TM	Taux d'annotation automatique TA
Sans contexte	**1 action / vignette**	**Vignettes dont la valeur est** • **correcte** • **déterminée automatiquement**
Avec contexte (word spotting)	**1 action / cluster**	

Résultats

Approche	**Production de vérité terrain** Tolérance : 1% d'erreur	**Tâche d'indexation** Tolérance : 20% d'erreur
Sans contexte	TM = 78,8% **TA = 20,2%**	**TM = 20,2%** TA = 59,8%
Avec contexte (word spotting)	TM = 55,6% **TA = 43,4%**	**TM = 10,7%** TA = 69,3%

Expérimentation 1 : transcription de patronymes (4/4)

Description de la page

- Poser une question pour chaque vignette
- Poursuivre l'interprétation

➔ La question contient les hypothèses

```
page →
  @(colonneAncienProp)
  lireTousLesNoms("inconnu")

lireTousLesNoms(precedent)→
  @(detecterNom)
  nom = CATCH(lireNom(precedent),
              T_NOM)
  @(dessous)
  LOOP(lireTousLesNoms(nom))

lireNom(precedent) →
  reco = TRY(reconnaitNomINT, T_NOM)
  res = (si reco=="IDEM"
         alors precedent
         sinon reco)
```

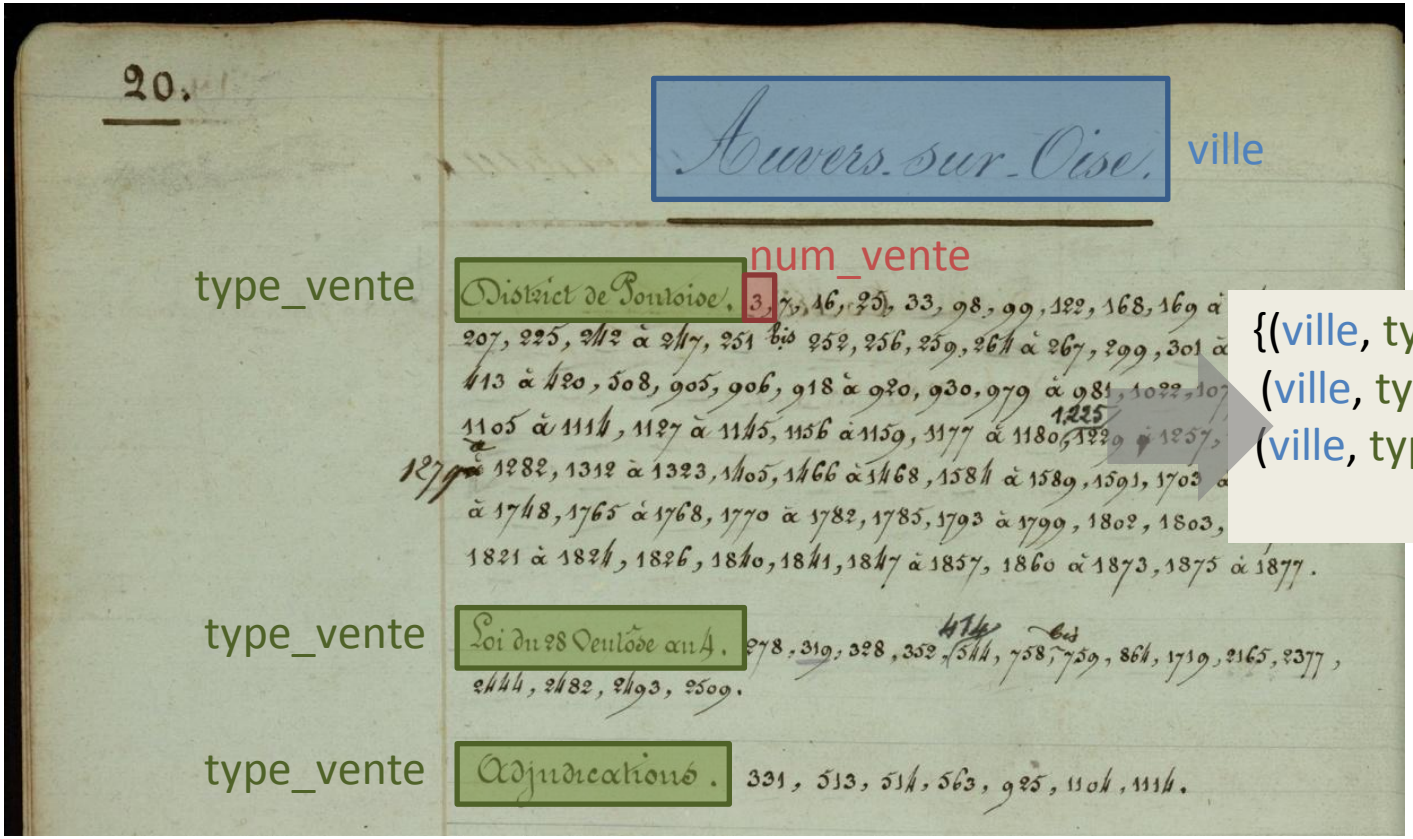
```
reconnaitNomINT →
  lsthypo = reconnaitNom
  RAISE_QUESTION("Confirmer nom.",
                 lsthypo,
                 T_NOM)
```

À retenir

- **Contexte + interprétation itérative = classifieur global**
- Quantité de travail
  ➔ ajusté au niveau global
- Améliorer le classifieur global
  ➔ pas d'impact sur la description locale

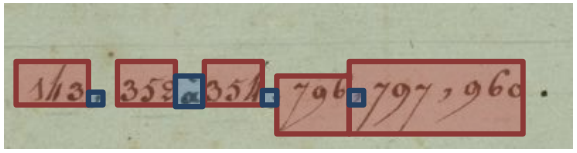
Plan

Expérimentation 2 : correction de sous-segmentation (1/4)



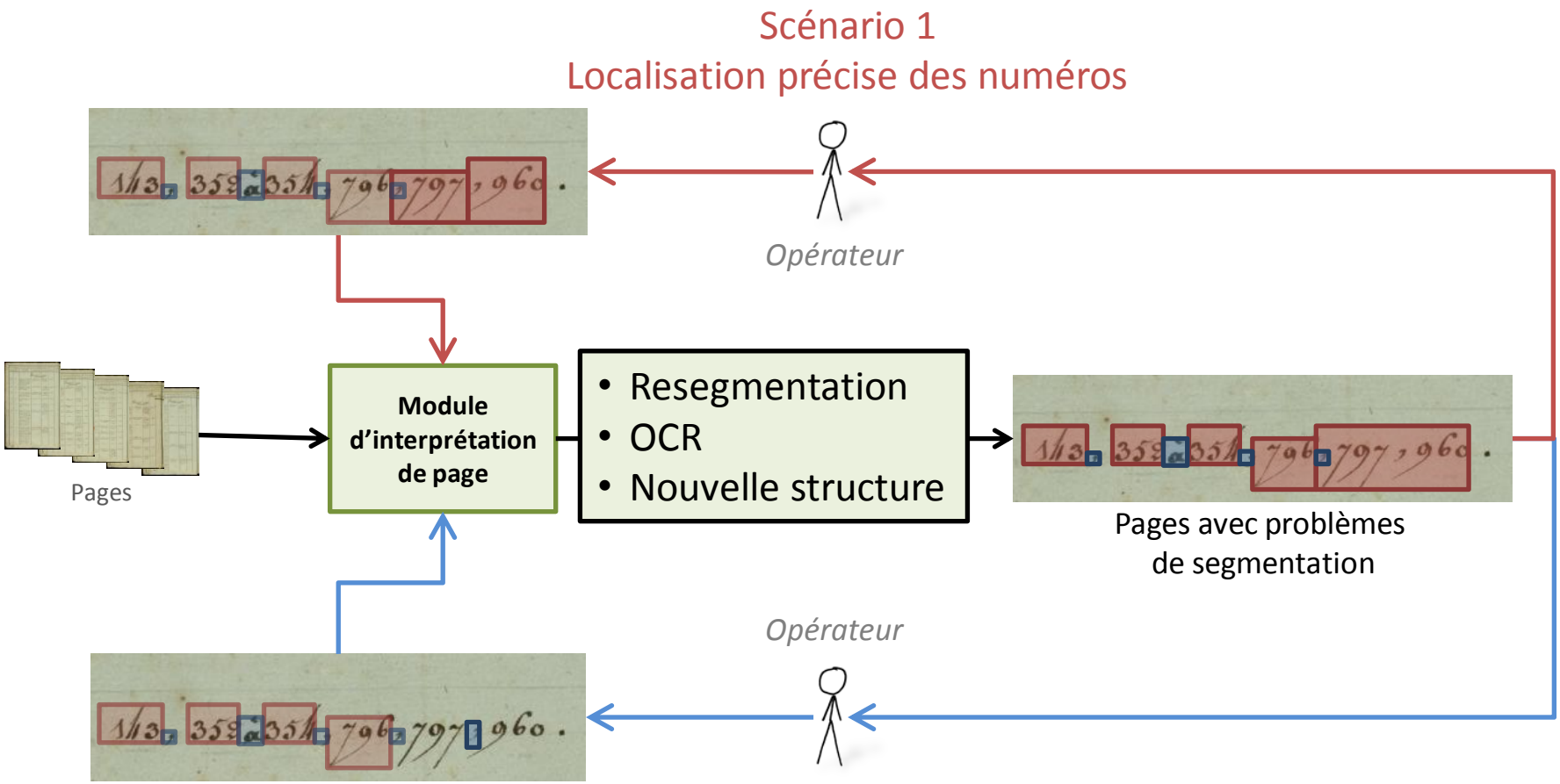
{(ville, type_vente, num_vente),
(ville, type_vente, num_vente),
(ville, type_vente, num_vente),
...}

Problème principal : sous-segmentation



Expérimentation 2 : correction de sous-segmentation (2/4)

Comparaison de deux scénarios



Scénario 2
Localisation rapide des séparateurs manqués

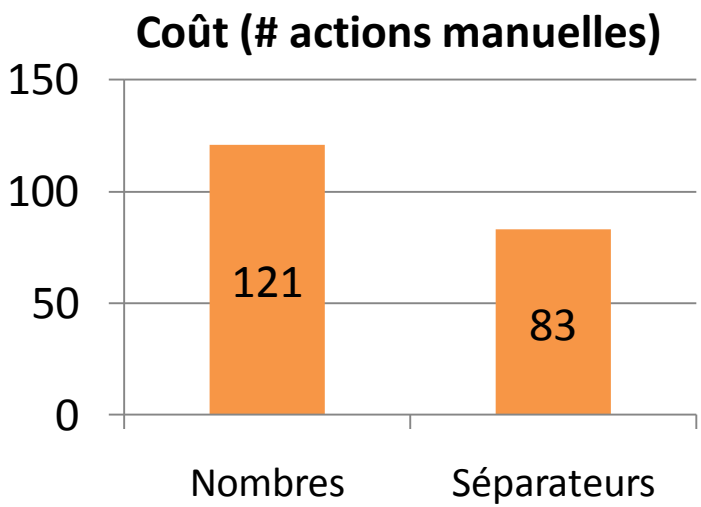
Expérimentation 2 : correction de sous-segmentation (3/4)

Données traitées : 50 pages (1637 vignettes de nombres)

Évaluation

- Quantité de travail manuel requis pour corriger les 121 nombres mal segmentés
- 1 action manuelle = …
  - 1 localisation précise de nombre
  - 1 localisation précise de séparateur

Résultats



Gain :
30% d'actions en moins
en corrigeant la cause des erreurs
(détection du séparateur)

Expérimentation 2 : correction de sous-segmentation (4/4)

Description de la page

- 1 description pour les 2 scénarios
  - ➔ Possibilité de les combiner
- Basée sur des questions optionnelles
  - ➔ Opérateur **OPT(Type, Règle)**

```
extraireNumVenteLigne →
   @(positionLigne)
   lstsep = OPT(T_SEP, detecterTousSep)
   OPT(T_NUM, extraireNbreEntreSep(lstsep))

detecterTousSep →
   sep = TRY(T_SEP, separateur)
   res = sep :: LOOP(detecterTousSep)

extraireNbreEntreSep (lstsep) →
   @(entreseparateur lstsep)
   TRY(T_NUM, extraireNombre)
   LOOP(extraireNbreEntreSep(lstsep))
```

À retenir

- **Correction de la cause des erreurs et non des conséquences**
- Possibilité de laisser la charge de la détection d'erreur à l'opérateur humain
- Possibilité de poser plusieurs questions de types différents

Plan

Utilisation du système en production

Données traitées (partenariat avec les Archives des Yvelines)

- Environ 1200 pages de documents d'archives
- 6700 lignes de ventes
- 11 000 vignettes de patronymes

Techniques mises en œuvre en conditions réelles

- Au niveau local
  - Descriptions avec plusieurs types de données échangées
  - Possibilité de détection d'erreur manuelle
  - Remise en cause en profondeur des résultats

  ➔ Avec des descriptions simples
- Au niveau global
  - Gestion centralisée du compromis automatique/manuel
  - Optimisation de séquences, regroupement de mots visuellement similaires

  ➔ Grâce à la séparation global/local facilitant la collaboration en conception

Plan

Conclusion

Notre contribution

Méthode pour mettre en œuvre une interprétation itérative

- Basée sur une approche déclarative
- Extension en 3 étapes
  1. **Mémoire visuelle**
  2. **Langage de description**
  3. **Architecture à 2 niveaux**

Avantages

- **Des outils proposés**
  - **Réintégration locale** de connaissances globales
  - Interaction **asynchrone**
  - Conception **centrée page**
- **De l'approche**
  - **Réutilisable** : basée sur des mécanismes standards
  - **Validée** : en théorie, en pratique, sur les expérimentations et en production

Perspectives

- Consolidation : vers l'**apprentissage de règles**
  - Difficulté d'amorçage du système réduite
  - Mécanisme itératif favorable à la remise en cause du modèle
- Extension : vers une **exploitation en consultation**
  - Utilisateurs actifs (crowdsourcing + traitements automatiques)