Extraction et interprétation sémantique de tables anciennes : défis et perspectives

Solenn Tual [1], Nathalie Abadie [1], Joseph Chazalon [2], Bertrand Duménieu [3], Julien Perret [1]

[1] LASTIG, Université Gustave Eiffel, IGN-ENSG
[2] LRE, EPITA
[3] CRH, EHESS-CNRS

Conférence Ingénierie des connaissances (PFIA 2025) / Journée Humanités et IA
4 juillet 2025

(1) (2) (3)

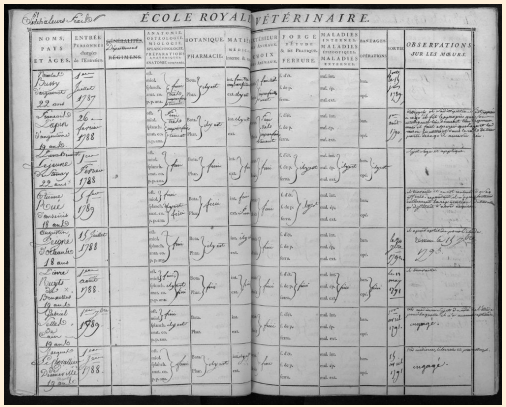
Contexte

- **Quantités significatives de documents tabulaires** dans les services d'archives, les musées et bibliothèques
- Informations intéressantes **pour différentes disciplines** (histoire, démographie, économie historique, géographie, etc.)

Exemple : documents tabulaires numérisés et diffusés par les Archives Départementales du Val-de-Marne (94)

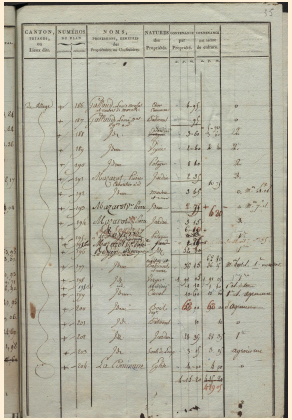
5 866 images

Registres détaillés des élèves inscrits à l'école vétérinaire de Maisons-Alfort (1760-1817)



24 012 images

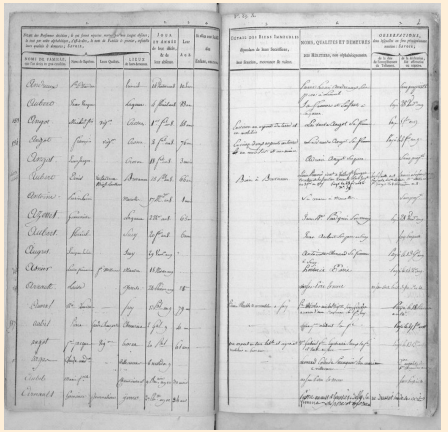
Registres d'états de sections du cadastre napoléonien (1809-)*



** Registres d'initialisation du cadastre (T0)*

163 371 images

Tables alphabétiques de décès, successions et d'absences (1759-1974)



Contexte

- **Quantités significatives de documents tabulaires** dans les services d'archives, les musées et bibliothèques

- Informations intéressantes **pour différentes disciplines** (histoire, démographie, économie, historique, géographie, etc.)

- **Numérisation** progressive de ces documents...
  - **Diffusion des images sur le Web**, pas des informations qu'elles contiennent

⇨ Documents qui **demeurent difficiles à exploiter dans leur intégralité** (volumes importants)

Enjeu

Faciliter l’accès aux connaissances structurées sous la forme de tables et contenues dans des documents historiques.

Enjeu

Faciliter l'accès aux connaissances structurées sous la forme de tables et contenues dans des documents historiques.

⇨ Travaux existants sur la représentation d'informations issues de documents historiques sous la forme de graphes de connaissances : [1] [2]

- Ne considèrent pas les tables (textes en prose majoritairement)
  - Les tables ont des caractéristiques différentes des textes en prose.

[1] V. Nundloll, et al. Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. Heliyon, 8(10) :e10710, 2022.
[2] N. Jain et al. Generating Domain-Specific Knowledge Graphs : Challenges with Open Information Extraction. In Proc. 1st Int. Workshop on Knowledge Graph GenerationFrom (Text2KG 2022)

Enjeu

Faciliter l'accès aux connaissances structurées sous la forme de tables et contenues dans des documents historiques.

⇨ Travaux existants sur la représentation d'informations issues de documents historiques sous la forme de graphes de connaissances : [1] [2]

- Ne considèrent pas les tables (textes en prose principalement)
  - Les tables ont des caractéristiques différentes des textes en prose.

⇨ **Comment produire des graphes de connaissances à partir de tables historiques issues de documents numérisés ?**

[1] V. Nundloll, et al. Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. Heliyon, 8(10) :e10710, 2022.
[2] N. Jain et al. Generating Domain-Specific Knowledge Graphs : Challenges with Open Information Extraction. In Proc. 1st Int. Workshop on Knowledge Graph GenerationFrom (Text2KG 2022)

Plan

1. **Description** des principales caractéristiques des tables historiques

2. Passage en revue des travaux portant sur :
   - L'**extraction d'informations (IE) à partir de tables anciennes numérisées**
   - L'**interprétation sémantique de tables (STI)**

3. **Mise en perspective de ces deux disciplines dans le but de produire des graphes de connaissances** à partir de tables historiques

Plan

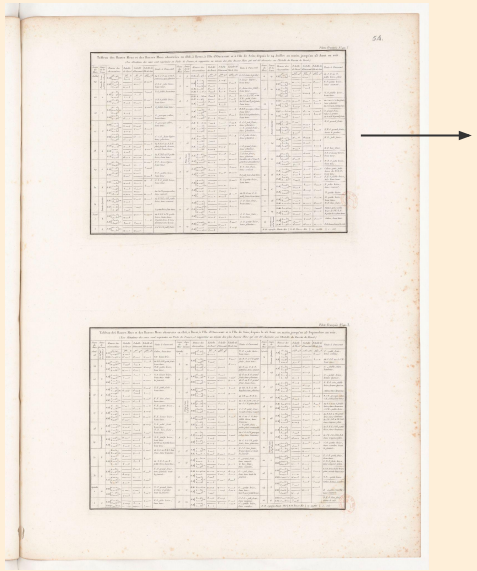
1. **Description** des principales caractéristiques des tables historiques

2. Passage en revue des travaux portant sur :
   - L'**extraction d'informations (IE) à partir de tables anciennes numérisées**
   - L'**interprétation sémantique de tables (STI)**

3. **Mise en perspective de ces deux disciplines dans le but de produire des graphes de connaissances** à partir de tables historiques

Caractéristiques des tables historiques

- **Manuscrites, imprimées, préimprimées** et remplies manuellement
- **Recours aux tables : isolées dans un ouvrage comme corps d'un document**



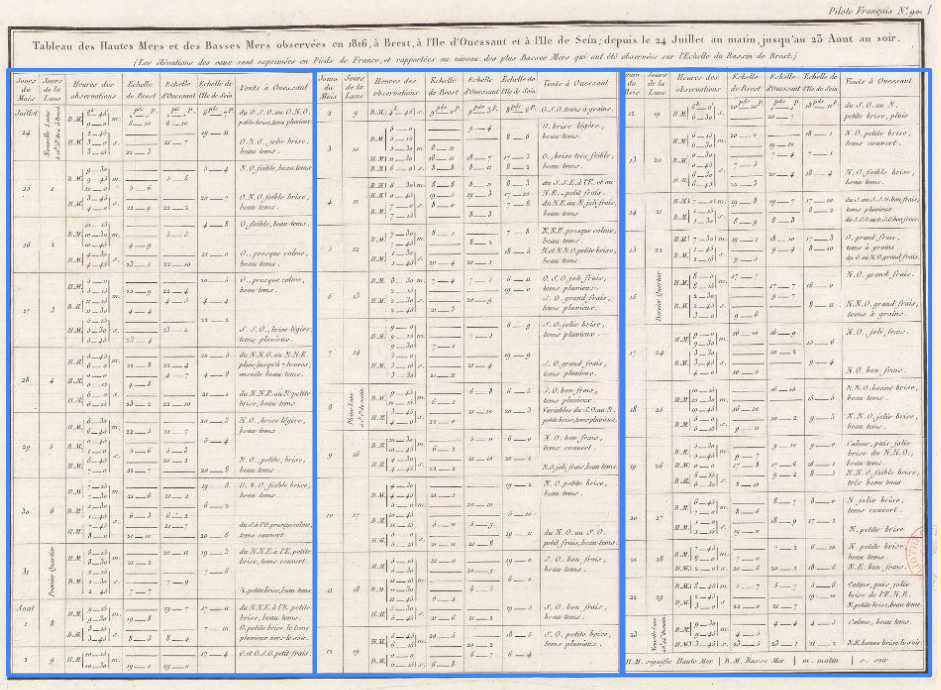
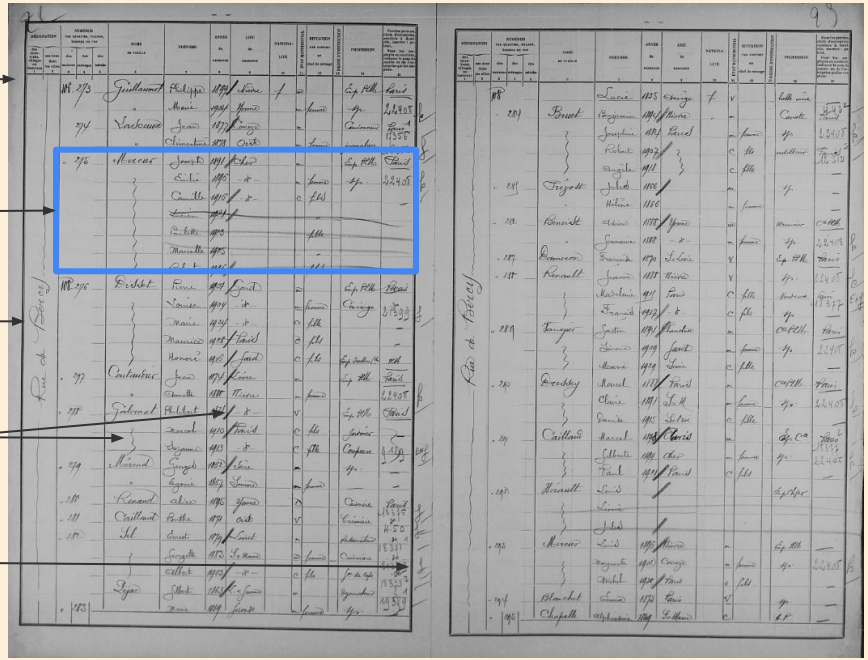


Tableau des Hautes Mers et des Basses Mers observées en 1816, à Brest, à l'Ile d'Ouessant et à l'Ile de Sein; depuis le 24 Juillet au matin, jusqu'au 25 Août au soir.

Tableau des hautes et basses mers observées en 1816 dans le Finistère [...]. Pilote Français, Tome 1. BNF

Caractéristiques des tables historiques

- Grande **variabilité des structures et des relations** entre les éléments de la table (cellules, lignes, colonnes).
- Respect relatif de la structure lors du remplissage manuel des tables



Registre de recensement de la population à Bercy. 1931. Vue 12/128. AD75

Caractéristiques des tables historiques

- **Variations de structures dans un même corpus**

Contribuable.s

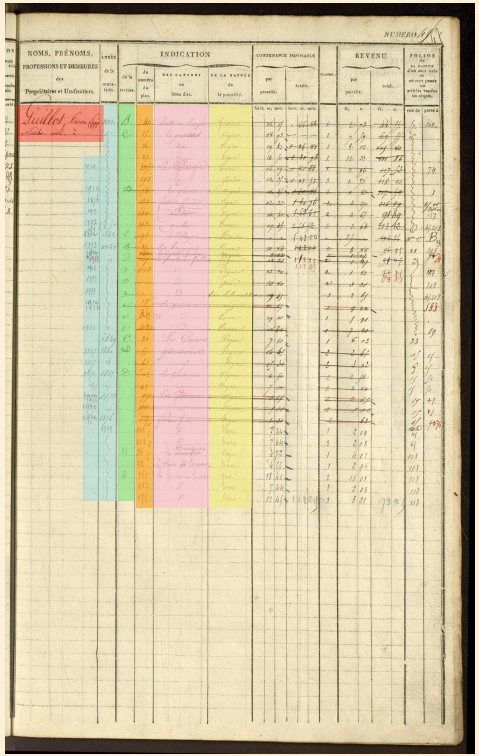
Section

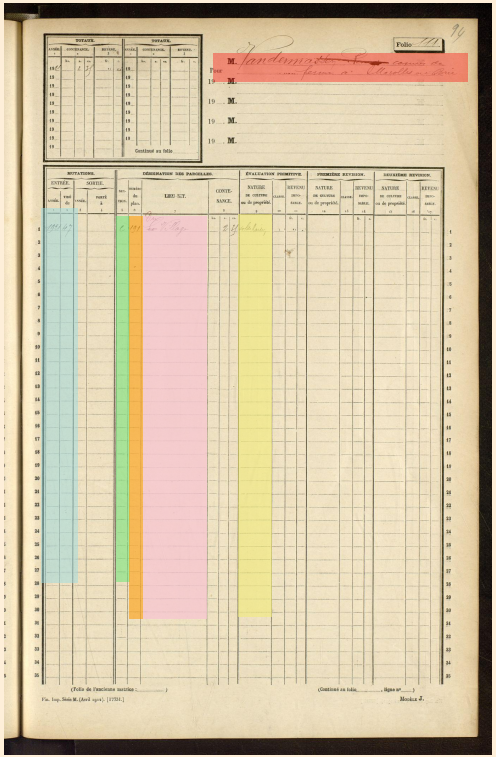
Parcelle

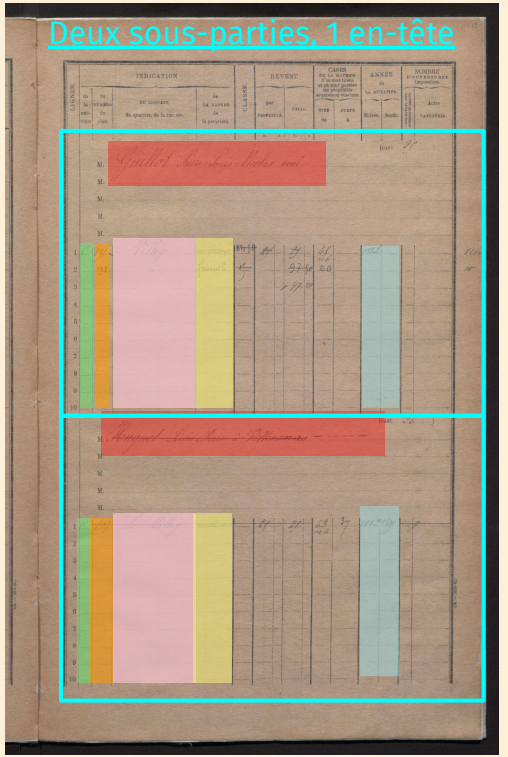
Adresse

Nature

Dates de validité







Registres cadastraux (matrices des propriétaires), Marolles-en-Brie, 1810-1932. AD94

Plan

1. **Description** des principales caractéristiques des tables historiques

2. Passage en revue des travaux portant sur :
   - L'**extraction d'informations (IE) à partir de tables anciennes numérisées**
   - L'**interprétation sémantique de tables (STI)**

3. **Mise en perspective de ces deux disciplines dans le but de produire des graphes de connaissances** à partir de tables historiques

Extraction d’informations (IE) dans des documents tabulaires historiques

- Localiser, transcrire et organiser le texte contenu dans des images de tables vers une représentation informatique structurée exploitable automatiquement

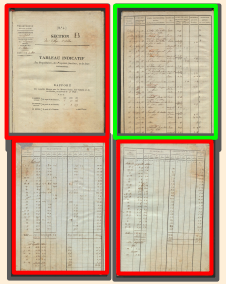
- Tâches (selon les approches, certaines étapes pourront être fusionnées) :

Collecte

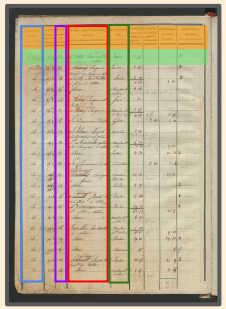
GLAM

Images, métadonnées, fiches d’inventaires

Détection de tables



Reconnaissance de la structure



Reconnaissance du texte

Village
22
Costy Jean Marie tailleur de pierres à Ablon
Jardin

Structuration

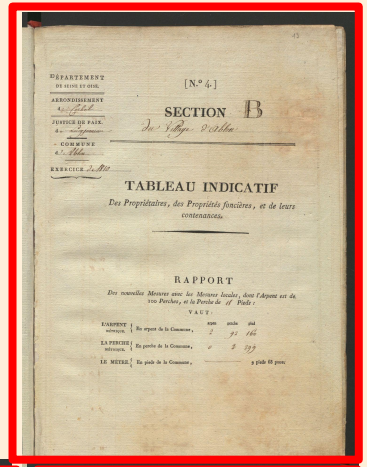
Adresse	N°	Contribuable	Nature
Village	22	Costy Jean Marie tailleur de pierres à Ablon	Jardin

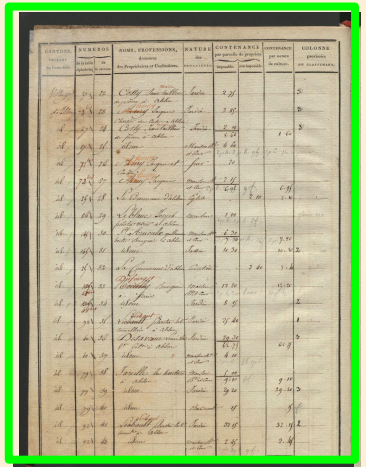
Détection des tables

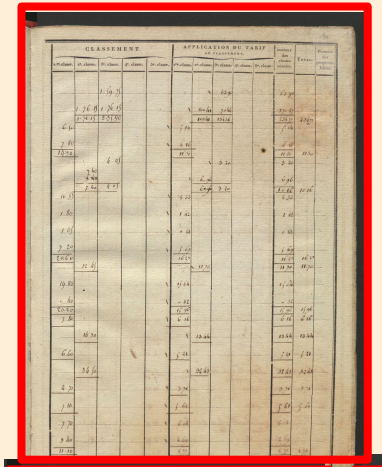
Exemple : classification de pages

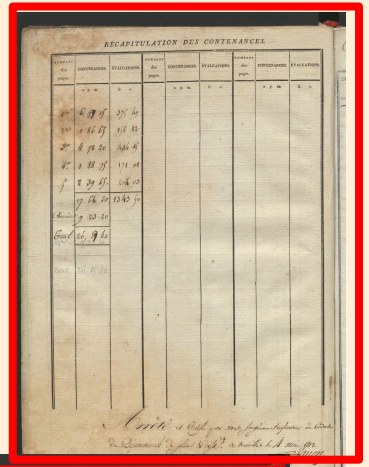
Différents types de pages : couverture, tableau principal, récapitulatif intermédiaire, résumé, ...

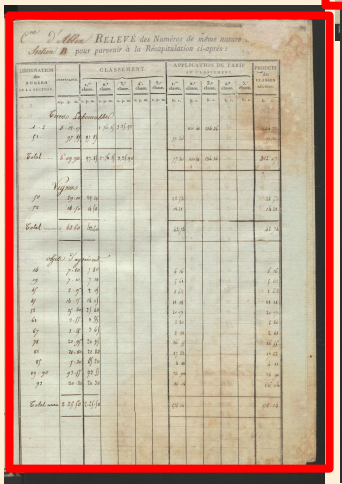
Registres d'état de sections (cadastre), Marolles-en-Brie, 1810. AD94

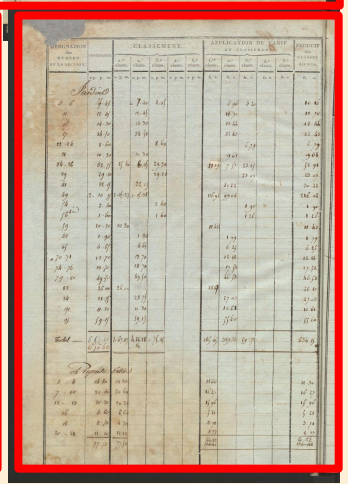


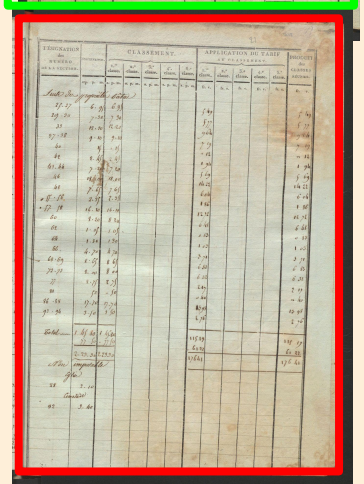


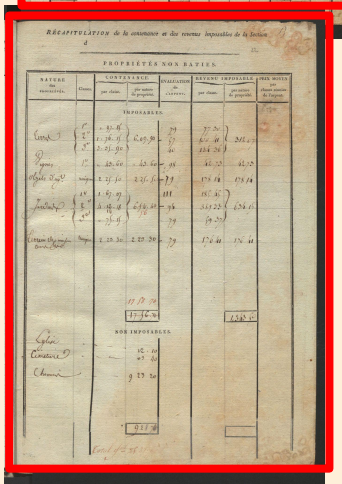


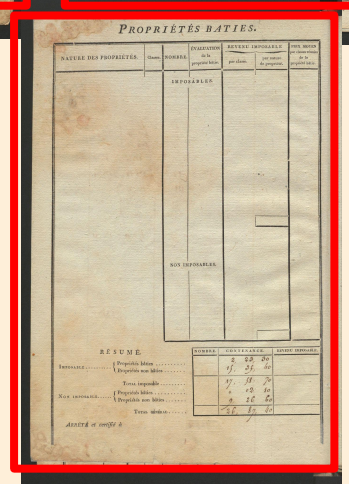






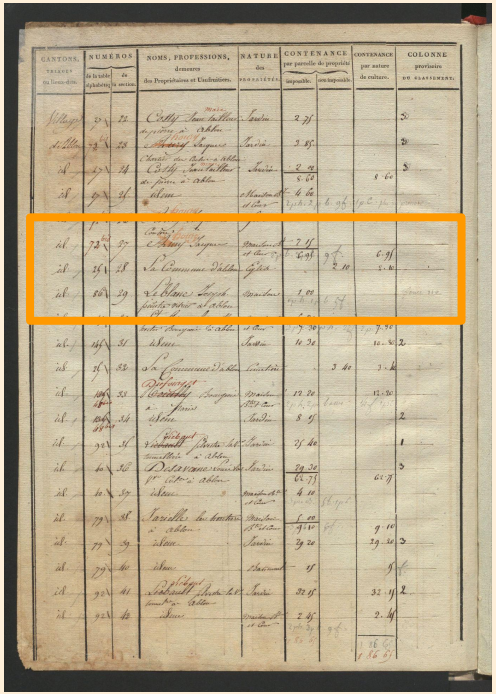






Reconnaissance de la structure et du texte

Extraction du contenu captif de l'image dans un format numérique textuel structuré



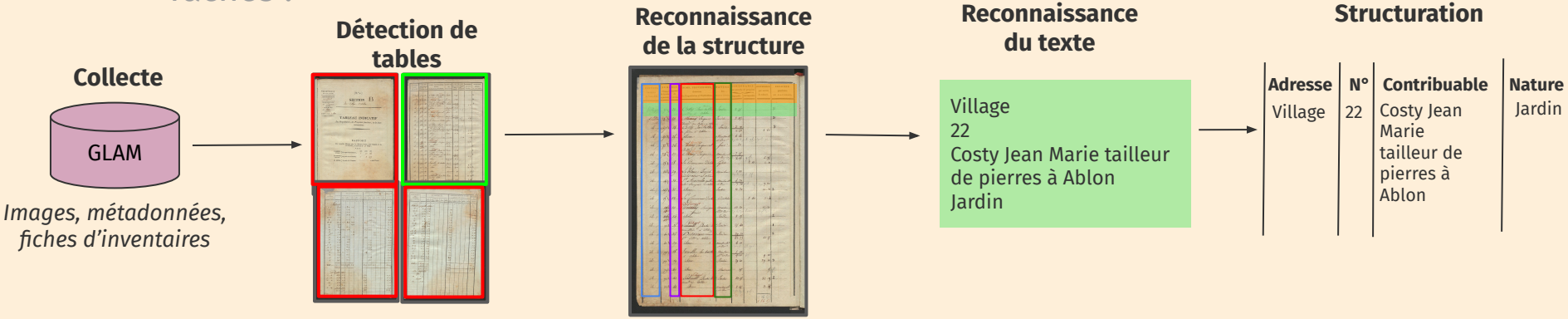
Registres d'état de sections (cadastre), Marolles-en-Brie, 1810. AD94

Transcription structurée

Numéro de parcelle	Adresse	Contribuable	Numéro du contribuable	Nature
..	..	..	..	..
27	id	*houry* Amy Jacques	72bis	Maison Bt et Cour
28	id	La Commune d'ablon	25	Eglise
29	id	Leblanc Joseph Peintre vitrier à ablon	86	Maison
..	..	..	..	..

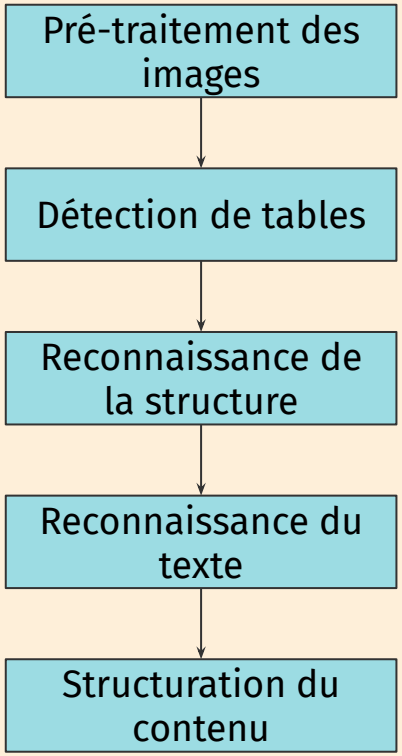
Extraction d’informations (IE) dans des documents tabulaires historiques

- Tâches :



- Approches :
  - Approches séquentielles
  - Approches *end-to-end*

Approches séquentielles



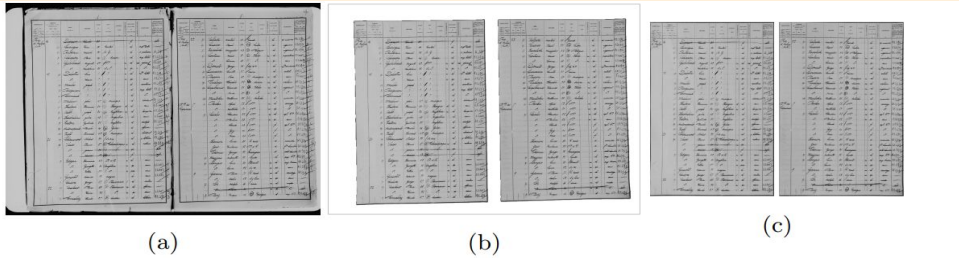


Fig. 3: (a) Example of a double page from the 1926 census. (b) Segmentation results (c) Dewarped tables.

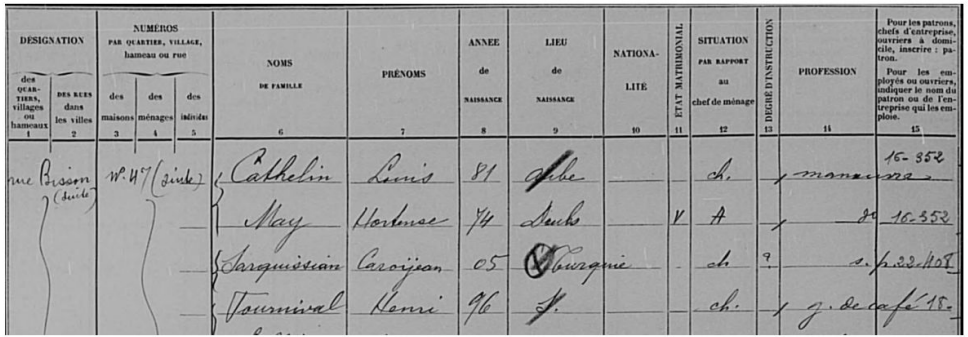
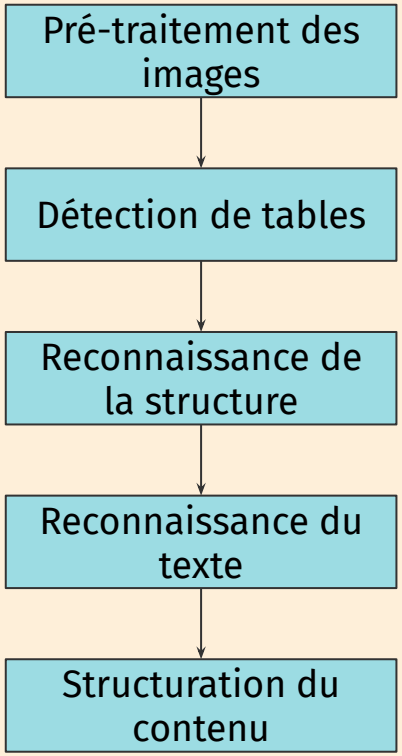
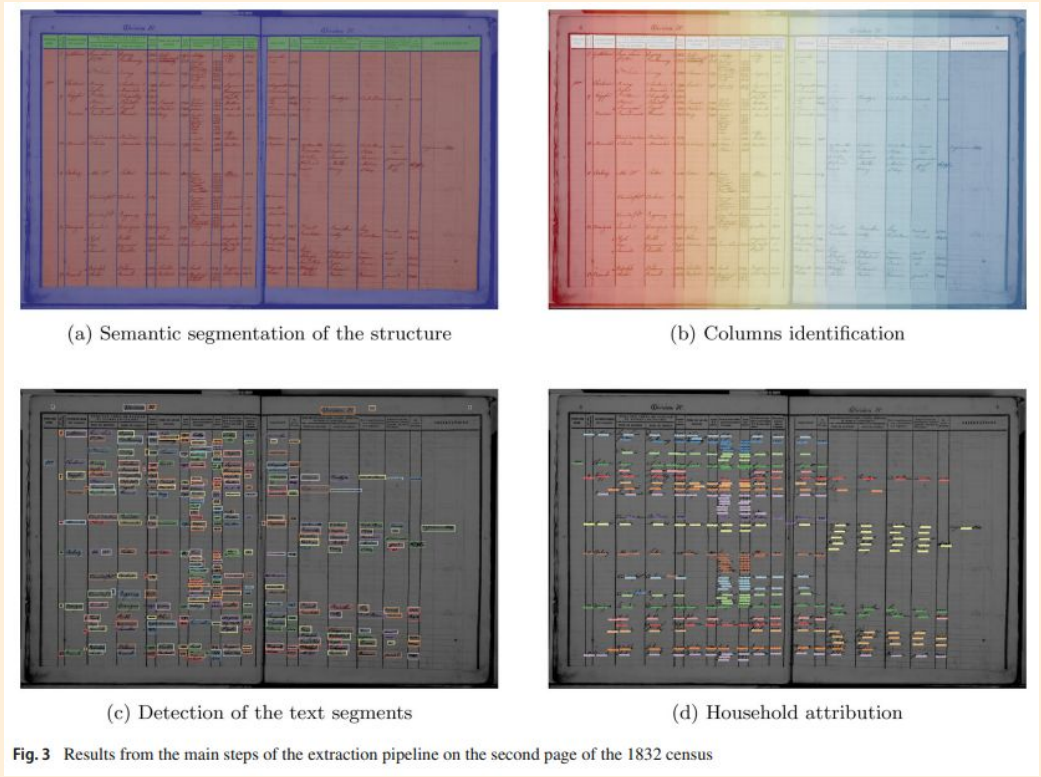


Fig. 1: Headers and first lines of a table from 1926. The corresponding annotation of the first line is "Cathelain/Louis/81/Aube/¤/¤/ch./¤/manoeuvre/?16-352". [1]

[1] T. Constum et al. Recognition and Information Extraction in Historical Handwritten Tables : Toward Understanding Early 20th Century Paris Census. In Document Analysis Systems, pages 143–157, 2022.
[2] R. Petitpierre et al. An end-to-end pipeline for historical censuses processing. IJDAR, 26(4) :419–432, 2023.

Approches séquentielles





(a) Semantic segmentation of the structure

(b) Columns identification

(c) Detection of the text segments

(d) Household attribution

Fig. 3 Results from the main steps of the extraction pipeline on the second page of the 1832 census

[2]

[1] T. Constum et al. Recognition and Information Extraction in Historical Handwritten Tables : Toward Understanding Early 20th Century Paris Census. In Document Analysis Systems, pages 143–157, 2022.

[2] R. Petitpierre et al. An end-to-end pipeline for historical censuses processing. IJDAR, 26(4) :419–432, 2023.

Approches *end-to-end*

Pré-traitement des images

↓

Détection de tables

↓

Reconnaissance de la structure et du texte

DAN [2], DANIEL [3]

↓

Structuration du contenu

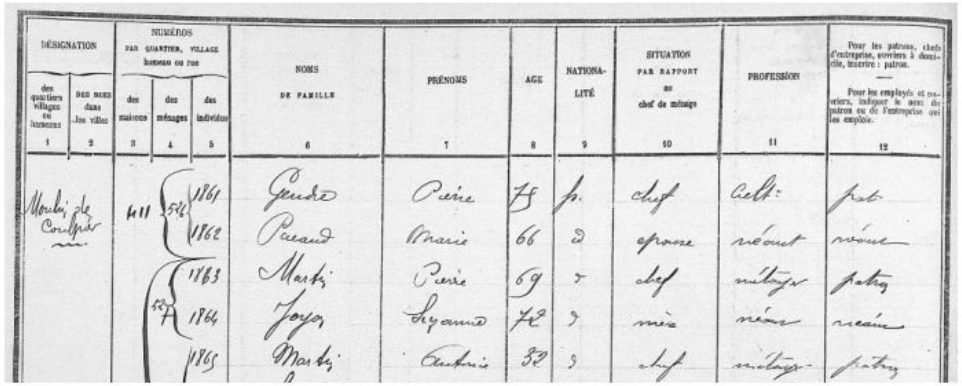


Fig. 5: Table header and first rows of a table from the census of the commune of Neuilly-le-Réal (department of Allier) in 1901. The label used to train the model for this part of the table is:

```
<s-h>Gendre <f>Pierre <o>cultivateur <l>chef <e>patron <a>75 <n>française
<s>Paraud <f>Marie <o>néant <l>épouse <e>néant <a>66 <n>idem
<s-h>Martin <f>Pierre <o>métayer <l>chef <e>patron <a>69 <n>idem
<s>Joyoz <f>Suzanne <o>néant <l>mère <e>néant <a>72 <n>idem
```

[1]

[1] M. Boillet et al. The Socface Project : Large-Scale Collection, Processing, and Analysis of a Century of French Censuses. ICDAR, pages 57–73, 2024.

[2] D. Coquenet, C. Chatelain, T. Paquet. DAN : a Segmentation-free Document Attention Network for Handwritten Document Recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence

[3] T. Constum. DANIEL : a fast document attention network for information extraction and labelling of handwritten documents. IJDAR, pages 1–23, 2025.

IE à partir de tables historiques : synthèse

- Progrès rapides avec les modèles à réseaux de neurones profonds, notamment les architectures Transformer
- Deux grands types d'approches :

	Approches séquentielles	**Approches *end-to-end***
Explicabilité	➕ (mais propagation d'erreurs)	➖
Jeux de données	1 pour chaque tâche qui nécessite l'entraînement d'un modèle	Réduction du nombre de jeux de données nécessaires
Contexte	➖	➕ (si extraction niveau page)

- ➡ Existe des approches qui permettent d'extraire du contenu de tables à grande échelle
- ➡ Nombreux défis et perspectives pour **parvenir à extraire des tables à la structure complexe** (cellules fusionnées/propagées, multiples orientations du texte)

Plan

1. **Description** des principales caractéristiques des tables historiques

2. Passage en revue des travaux portant sur :
   - L'**extraction d'informations (IE) à partir de tables anciennes numérisées**
   - L'**interprétation sémantique de tables (STI)**

3. **Mise en perspective de ces deux disciplines dans le but de produire des graphes de connaissances** à partir de tables historiques

Interprétation sémantique de tables (STI)

- Annoter les différents éléments qui composent un tableau et leurs relations à l'aide des ressources d'un graphe de connaissances (KG)
- Tâches [1][2] :

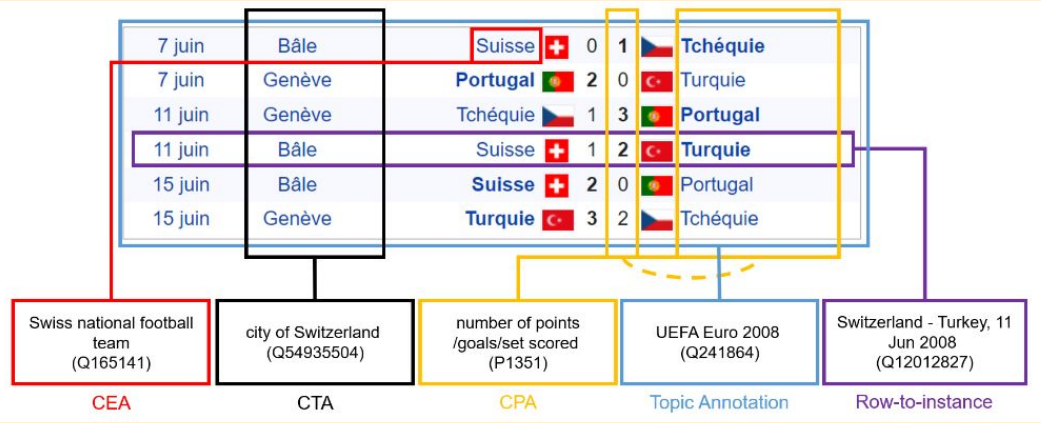


Illustration des tâches de STI. Source: [1]

- Annotation de cellules avec des entités (**CEA**)
  + Détection de nouvelles entités (CNEA) / Not In Lexicon
- Annotation de colonnes avec des types (**CTA**)
- Annotation de colonnes avec des propriétés (**CPA**)
- **Thématisation**
- **Correspondance ligne-instance**

Tableau Wikipédia décrivant des matchs de football de l'Euro 2000 annotés avec des entités de Wikidata

[1] J. Liu et al. From tabular data to knowledge graphs : A survey of semantic table interpretation tasks and methods. J. Web Semant., 76 :100761, 2023.
[2] M. Cremaschi et al. Survey on Semantic Interpretation of Tabular Data : Challenges and Directions, 2024. arXiv :2411.11891

Approches de STI

Approches heuristiques

Mesures de similarités, TF-IDF, vote majoritaire, méthodes probabilistes, etc.

ET

Reclassement des candidats

Ex : comparaison de labels d'entités et de mentions (CEA)

Ingénierie des caractéristiques

Extraction de caractéristiques

PUIS

Entraînement de modèles de machine learning (Random Forest, SVM, K-NN)

Ex : caractéristiques du contenu d'une colonne pour déterminer son type (CTA)

Apprentissage profond

Apprentissage de plongements
- De tables
- De graphes

PUIS

Comparaisons dans l'espace vectoriel

Jeux de données d'apprentissage nécessaire

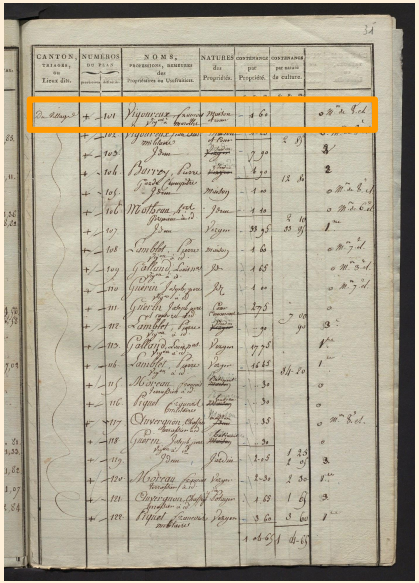
[1] J. Liu et al. From tabular data to knowledge graphs : A survey of semantic table interpretation tasks and methods. J. Web Semant., 76 :100761, 2023.
[2] M. Cremaschi et al. Survey on Semantic Interpretation of Tabular Data : Challenges and Directions, 2024. arXiv :2411.11891

Interprétation sémantique de tables : synthèse et limites

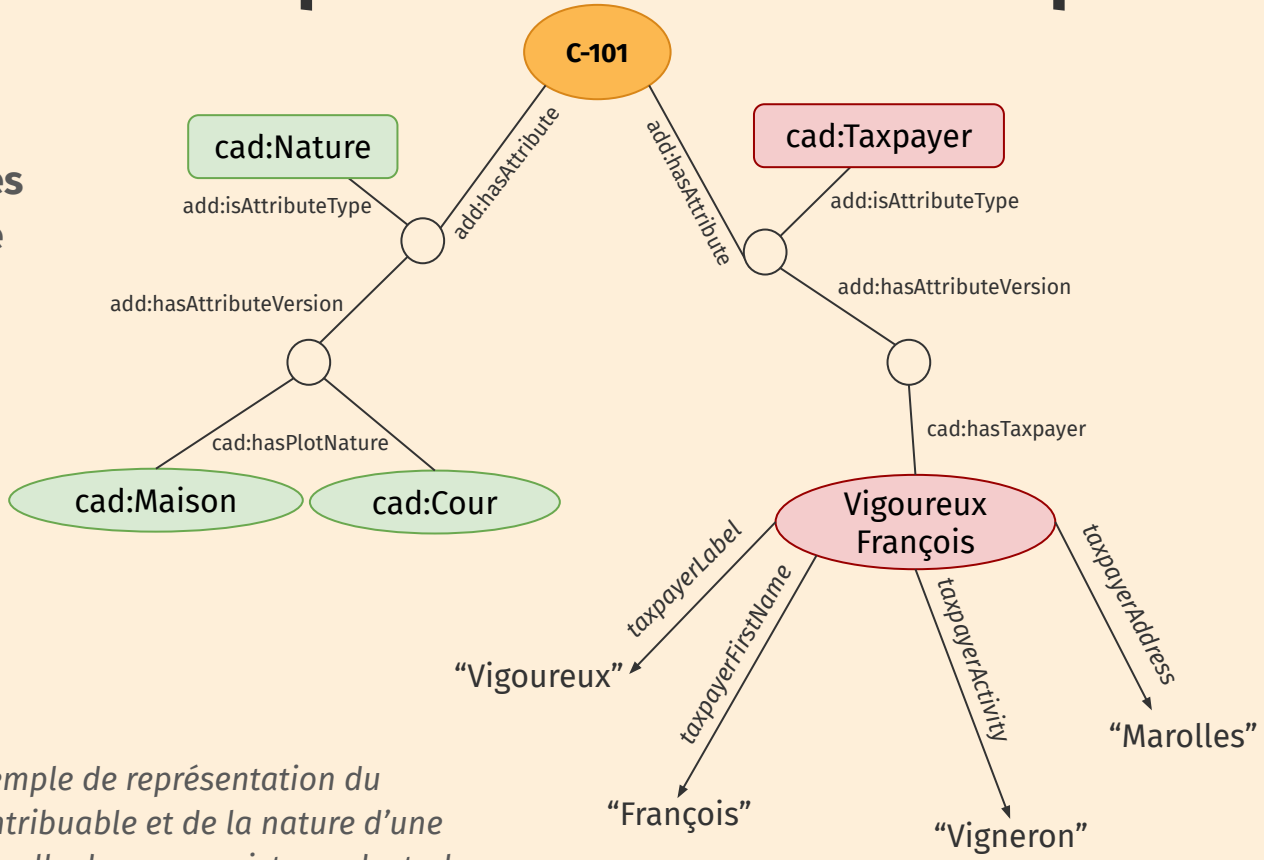
- Travaux nombreux pour annoter des **tables relationnelles ou d'entités**
  - Tables à la structure complexe, aux sujets cachés ou composés ?
- **Annotation souvent réalisée avec des graphes de connaissances encyclopédiques**
  - Graphes de connaissances de domaine ?
- Évaluations réalisées avec des **jeux de données d'état de l'art**
  - Cas d'usages sur des données "réelles"
- Toutes les approches **ne considèrent pas les entités *NIL (Not In Lexicon)***
  - Indispensable pour des applications historiques
- **Contexte sous-exploité**
  - Contexte très important pour comprendre des tables historiques

Interprétation sémantique de tables historiques

Combiner les chaînes de traitement d'IE et de STI ensemble pour structurer des informations et découvrir de nouvelles connaissances

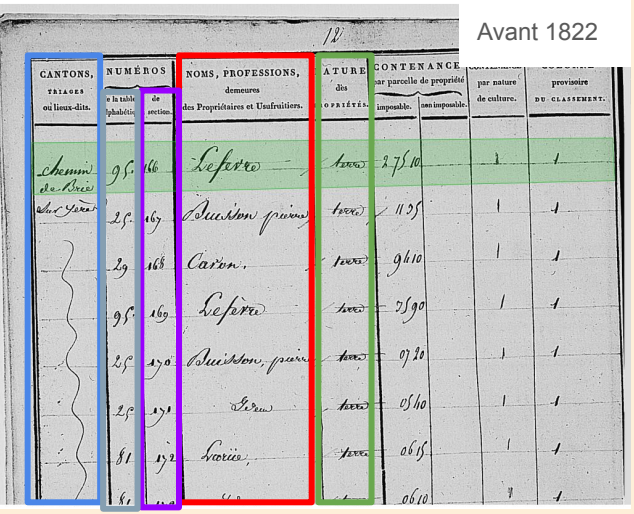


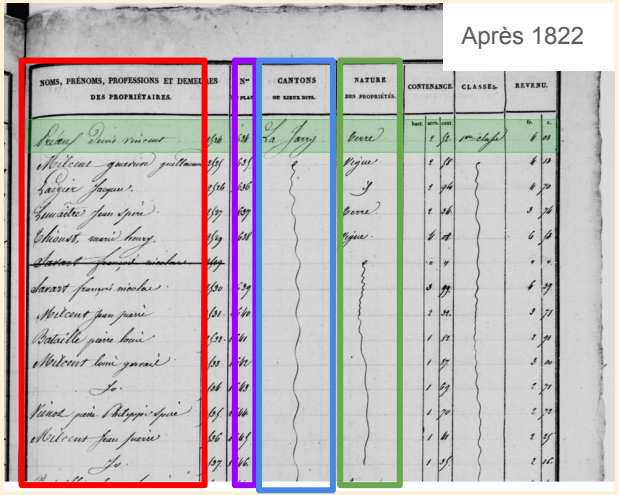
Exemple de représentation du contribuable et de la nature d'une parcelle dans un registre cadastral

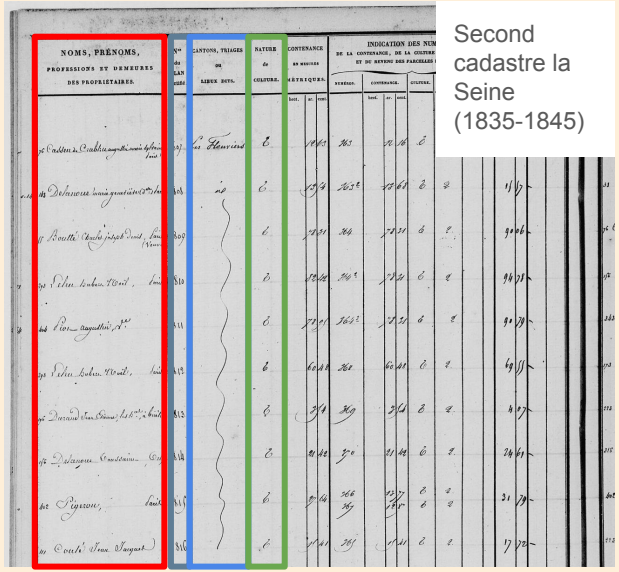


Défis et approches possibles

- Modèle d'annotation "image" guidé par une **ontologie de domaine** : résout certaines tâches de STI







CTA (exemple avec l'ontologie PeGazUs [1])

add:Landmark — cad:Taxpayer

xsd:string — cad:Nature — xsd:string

Type d'une ligne

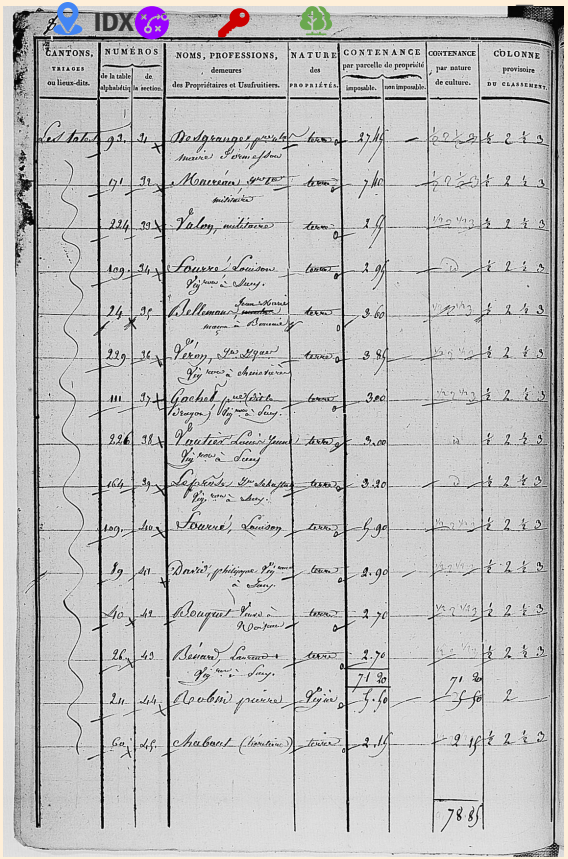
add:Landmark *add:isLandmarkType* cad:Plot

\+ ***CPA :*** *se référer à l'ontologie de domaine qui décrit ces classes*

[1] C. Bernard et al. PeGazUs: A knowledge graph based approach to build urban perpetual gazetteers. EKAW 2024, 10.1007/978-3-031-77792-9_22

Défis et approches possibles

- Extraction d'informations : **approche *end-to-end* à l'échelle de la page**



+ un seul jeu de données (image → table numérique)
+ modèle qui traite les variations de structures de pages
+ fournit davantage de contexte au modèle pour chaque ligne à traiter
+ réduit les occasions de propager des erreurs
- hallucinations et effet "boite noire"

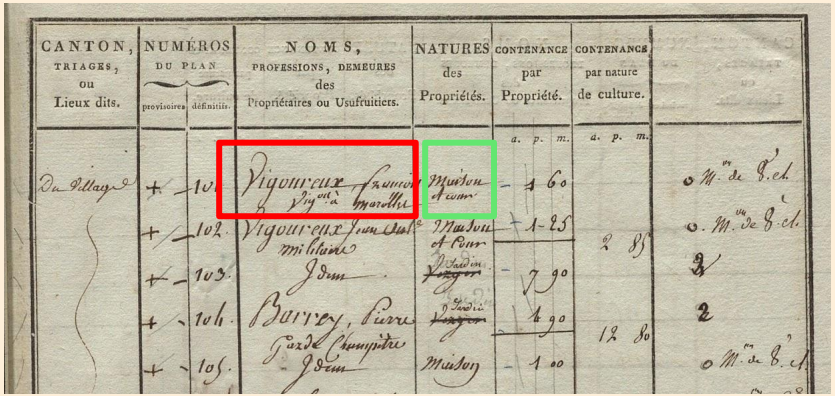
		IDX		
Les tates	31	93	Desgrange, p↑re↓ se la→marie	terre
§	320	171	Mucréau Et↑e↓ Et↑e↓	terre
§	33	234	Valon mitilaire	terre
§	34	109	Pourre S Louison	terre
§	35	24	Bellemaur,Jean Marie→maçon à Beneune	terre
§	36	229	Veron, V↑e↓ liques↓	terre
§	37	111	Gachet p↑re↓ louis→à Seury	terre
§		226	Vautier Lecis Jean S58	terre
§	39	164	Lepriase V↑e↓ setustiers à Vig en	terre
§	40	109	Seurré Louison	terre
§	411	89	David philippe Vig↑on↓	terre
§	42	40	Bouguet Veuve	terre
§	43	26	Benard Laureur→Vig VN	terre
§	414	211	Robin pierre	Vigne

Table d'un registre d'états de sections annotée automatiquement avec DAN. [1]

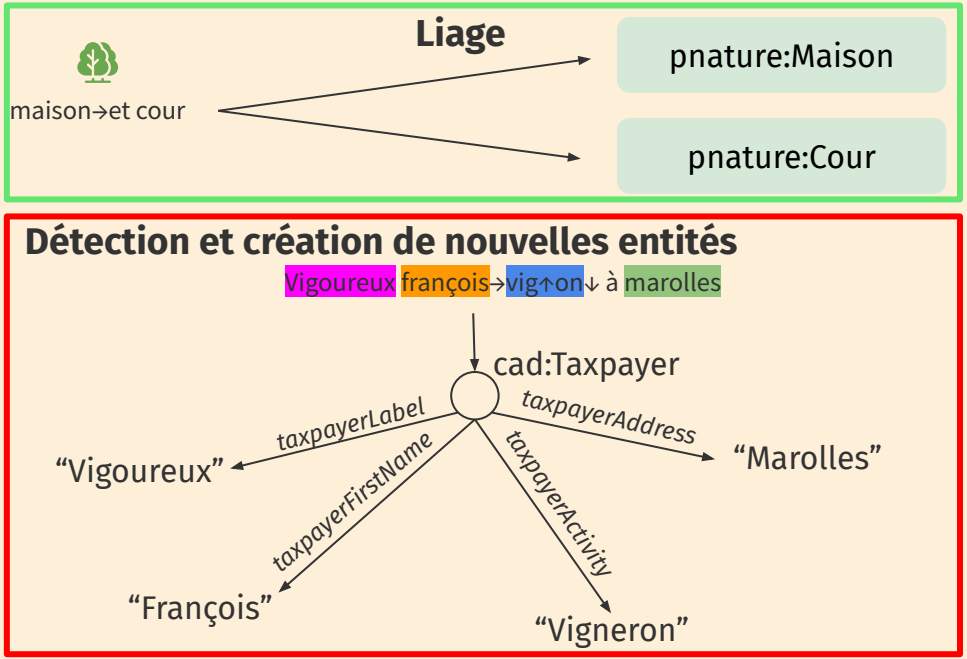
[1] Tual et al. An end-to-end pipeline for knowledge graph population from 19th-century land registry digitised tables. TPDL25, Sep 2025, Tampere, Finland. https://hal.science/hal-05118320

Défis et approches possibles

- Interprétation sémantique de tables
  - Graphes de connaissances généralistes (peu/pas utilisables avec ces données)
    - ⇒ **Annotation à l'aide de l'ontologie de domaine**
    - ⇒ Enrichissement progressif de cette ontologie avec la **découverte de nouvelles entités**



Exemple d'interprétation sémantique des tables du cadastre ancien (ontologie PeGazUs [1])



[1] C. Bernard et al. PeGazUs: A knowledge graph based approach to build urban perpetual gazetteers. EKAW 2024, 10.1007/978-3-031-77792-9_22

Défis et approches possibles

- **Collecte et utilisation des métadonnées** fournies avec les tables historiques
  - Utile pour la **thématisation des tables et l'enrichissement de leur contenu**
  - **Uniformisation des métadonnées et de leurs formats**

- **Production de métadonnées supplémentaires** à partir des documents

- **Maintenir le lien entre la source (image) et l'information produite**
  - Utile aux chercheurs en SHS et aux institutions patrimoniales
  - Confrontation possible entre le résultat obtenu automatiquement et la lecture de l'utilisateur
  - Très forte compatibilité entre les graphes de connaissances et le protocole IIIF [1] [2]

[1] Web Annotation Data Model, W3C, https://www.w3.org/TR/annotation-model/
[2] Simplest Annotation, IIIF Cookbook, https://iiif.io/api/cookbook/recipe/0266-full-canvas-annotation/

Conclusion

- Combiner les chaînes de traitement d'IE et de STI ensemble : **Nombreuses perspectives pour extraire, structurer et rendre accessible des connaissances**
    - **IE pour extraire la table dans un format tabulaire numérique**
    - **STI pour annoter les concepts et propriétés présents dans la table** et en découvrir des nouveaux
    - Ontologie utilisée par les deux tâches
- Permettre une **exploration à grande échelle** de documents tabulaires historiques

Perspectives

- **Évaluer la chaîne de traitement** proposée

⇨ Implémentation réalisée avec les registres d'états de sections du cadastre napoléonien

[Accepté] Solenn Tual, Nathalie Abadie, Joseph Chazalon, Bertrand Duménieu, Julien Perret. **An end-to-end pipeline for knowledge graph population from 19th-century land registry digitised tables.** 29th International Conference on Theory and Practice of Digital Libraries (TPDL25), Sep 2025, Tampere, Finland. https://hal.science/hal-05118320

Perspectives

- **Évaluer la chaîne de traitement** proposée

⇨ Implémentation réalisée avec les registres d'états de sections du cadastre napoléonien

[Accepté] Solenn Tual, Nathalie Abadie, Joseph Chazalon, Bertrand Duménieu, Julien Perret. **An end-to-end pipeline for knowledge graph population from 19th-century land registry digitised tables.** 29th International Conference on Theory and Practice of Digital Libraries (TPDL25), Sep 2025, Tampere, Finland. https://hal.science/hal-05118320

- **Modèles/méthodes/chaînes de traitement généralistes pour les tables historiques ?**

Merci pour votre attention !

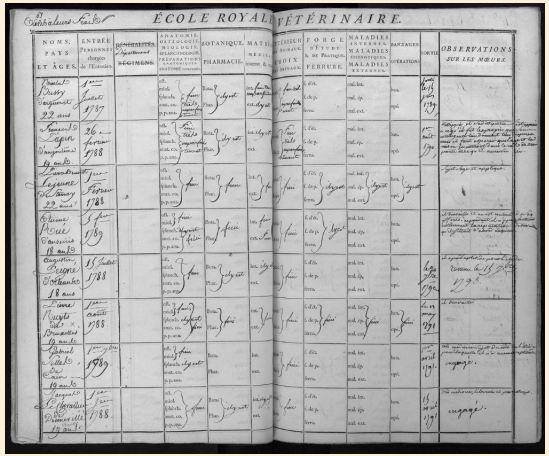
Extraction et interprétation sémantique de tables anciennes : défis et perspectives

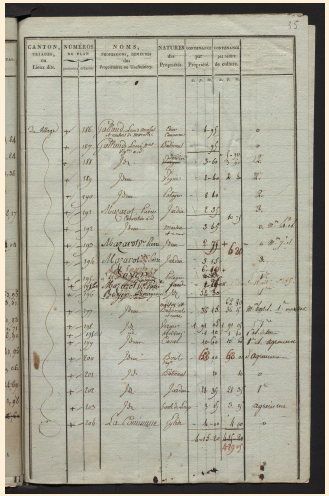
Solenn Tual [1], Nathalie Abadie [1], Joseph Chazalon [2], Bertrand Duménieu [3], Julien Perret [1]

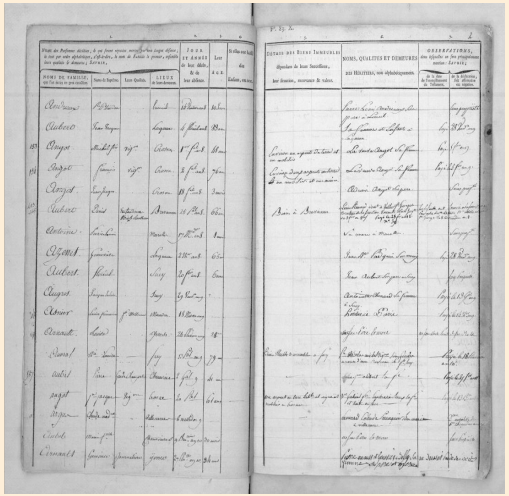
[1] LASTIG, Université Gustave Eiffel, IGN-ENSG
[2] LRE, EPITA
[3] CRH, EHESS-CNRS

{solenn.tual, nathalie-f.abadie, julien.perret}@ign.fr
www.umr-lastig.fr

joseph.chazalon@epita.fr
bertrand.dumenieu@ehess.fr







Références

M. **Boillet**, S. Tarride, Y. Schneider, B. Abadie, L. Kesztenbaum, and C. Kermorvant. The Socface Project: Large-Scale Collection, Processing, and Analysis of a Century of French Censuses. In ICDAR, pages 57–73, 2024.

M. **Cremaschi**, B. Spahiu, M. Palmonari, E. Jimenez-Ruiz. “Survey on Semantic Interpretation of Tabular Data: Challenges and Directions”, 2024. arXiv :2411.11891.

T. **Constum**, N. Kempf, T. Paquet, P. Tranouez, C. Chatelain, S. Brée, and F. Merveille. Recognition and Information Extraction in Historical Handwritten Tables: Toward Understanding Early 20th Century Paris Census. In Document Analysis Systems, pages 143–157, 2022.

T. **Constum**. DANIEL: a fast document attention network for information extraction and labelling of handwritten documents. IJDAR, pages 1–23, 2025.

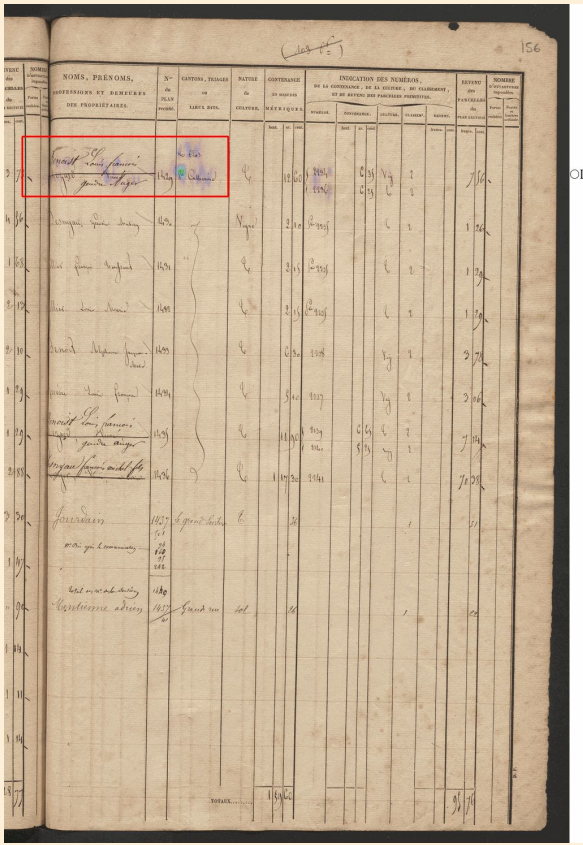
D. **Coquenet**, C. Chatelain, T. Paquet. DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition”. IEEE Transactions on Pattern Analysis and Machine Intelligence, inPress, ⟨10.1109/tpami.2023.3235826⟩.

N. **Jain** et al. Generating Domain-Specific Knowledge Graphs: Challenges with Open Information Extraction. In Proc. 1st Int. Workshop on Knowledge Graph Generation (Text2KG 2022)

J. **Liu**, Y. Chabot, R. Troncy, V.-P. Huynh, T. Labbé, and P. Monnin. “From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods.” J. Web Semant., 76: 100761, 2023.

V. **Nundloll**, et al. Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. Heliyon, 8(10): e10710, 2022.

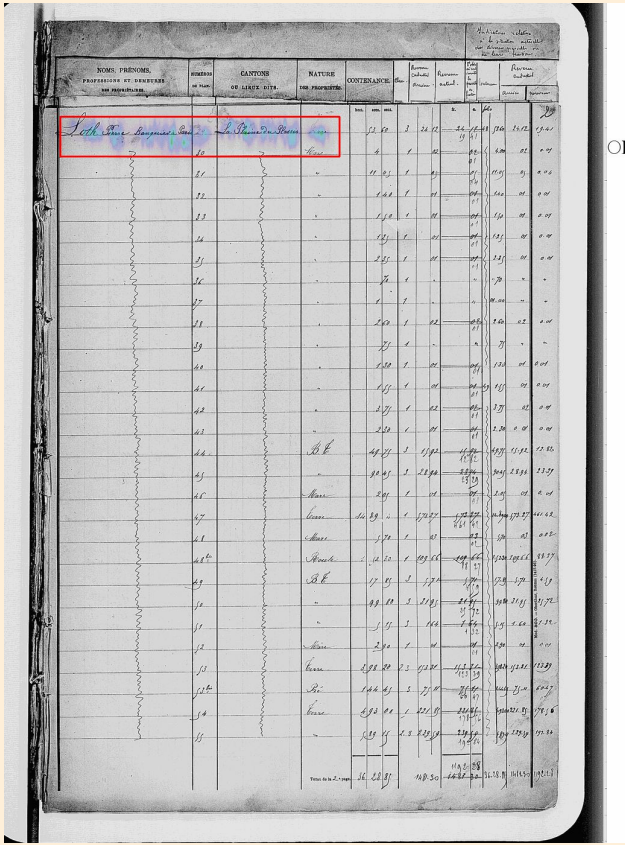
Hallucinations



○Le Clos→S↑te↓ Catherine ○Henoist Louis françois, F↑dre↓ Ruger ○1429 ○T ○2234→2236 ○Vig→T

Hallucinations une fois toutes les lignes prédites

Hallucinations



○La Plaine du Plessis ○Loth Pierre Banguier à Paris ○29 ○Terre

Désynchronisation de la lecture liée à la présence de nombreuses cases vides