

Extraction et interprétation sémantique de tables anciennes : défis et perspectives

Solenn Tual¹, Nathalie Abadie¹, Joseph Chazalon², Bertrand Duméniou³, Julien Perret¹

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.)
- **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

Pilote Français N° 101

Tableau des Hautes Mers et des Basses Mers observées en 1816, à Brest, à l'Île d'Ouessant et à l'Île de Sein; depuis le 24 Juillet au matin jusqu'au 25 Aout au soir.
(Une observation des vagues est exprimée en Pieds de France, et rapportée au niveau des plus basses Mers qui ont été observées sur l'Échelle du Bassin de Brest.)

Jours de la Mer	Heures de l'observation	Echelle de Brest	Echelle d'Ouessant	Echelle de l'Île de Sein	Vents à Ouessant	Jours de la Mer	Heures de l'observation	Echelle de Brest	Echelle d'Ouessant	Echelle de l'Île de Sein	Vents à Ouessant	Jours de la Mer	Heures de l'observation	Echelle de Brest	Echelle d'Ouessant	Echelle de l'Île de Sein	Vents à Ouessant
14	10	10	10	10	de S. O. au N. O. petite bruyante pluvieuse	15	10	10	10	10	de S. O. au N. O. petite bruyante pluvieuse	16	10	10	10	10	de S. O. au N. O. petite bruyante pluvieuse
20	10	10	10	10	N. O. faible bruyante	21	10	10	10	10	N. O. grand frais, bruyante	22	10	10	10	10	N. O. grand frais, bruyante
26	10	10	10	10	N. O. faible bruyante	27	10	10	10	10	N. O. grand frais, bruyante	28	10	10	10	10	N. O. grand frais, bruyante
34	10	10	10	10	N. O. grand frais, bruyante	35	10	10	10	10	N. O. grand frais, bruyante	36	10	10	10	10	N. O. grand frais, bruyante
37	10	10	10	10	N. O. grand frais, bruyante	38	10	10	10	10	N. O. grand frais, bruyante	39	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	40	10	10	10	10	N. O. grand frais, bruyante	41	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	42	10	10	10	10	N. O. grand frais, bruyante	43	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	44	10	10	10	10	N. O. grand frais, bruyante	45	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	46	10	10	10	10	N. O. grand frais, bruyante	47	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	48	10	10	10	10	N. O. grand frais, bruyante	49	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	50	10	10	10	10	N. O. grand frais, bruyante	51	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	52	10	10	10	10	N. O. grand frais, bruyante	53	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	54	10	10	10	10	N. O. grand frais, bruyante	55	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	56	10	10	10	10	N. O. grand frais, bruyante	57	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	58	10	10	10	10	N. O. grand frais, bruyante	59	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	60	10	10	10	10	N. O. grand frais, bruyante	61	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	62	10	10	10	10	N. O. grand frais, bruyante	63	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	64	10	10	10	10	N. O. grand frais, bruyante	65	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	66	10	10	10	10	N. O. grand frais, bruyante	67	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	68	10	10	10	10	N. O. grand frais, bruyante	69	10	10	10	10	N. O. grand frais, bruyante
39	10	10	10	10	N. O. grand frais, bruyante	70	10	10	10	10	N. O. grand frais, bruyante	71	10	10	10	10	N. O. grand frais, bruyante

H. B. - réimpression Haute-Mer | N. M. - Basse-Mer | m. - mètres | s. - secondes

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

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

Table pré-imprimée complétée manuellement

Groupe de lignes formant un foyer

Texte orienté

Idem, matérialisés par différents mots ou symboles

Abréviations

NUMERO	NOM	PRENOM	ANNEE DE NAISSANCE	SEX	ETAT	PROFESSION	REMARQUES
175	Guthmann	Bélisa	1878	fé	ép.	PH. Bercy	
176	Morocq	Jean	1878	hom.	ép.	PH. Bercy	
177	Guthmann	Blaise	1878	hom.	ép.	PH. Bercy	
178	Guthmann	Blaise	1878	hom.	ép.	PH. Bercy	
179	Morocq	Jean	1878	hom.	ép.	PH. Bercy	
180	Morocq	Jean	1878	hom.	ép.	PH. Bercy	
181	Morocq	Jean	1878	hom.	ép.	PH. Bercy	
182	Morocq	Jean	1878	hom.	ép.	PH. Bercy	
183	Morocq	Jean	1878	hom.	ép.	PH. Bercy	

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

A historical cadastral register page from Marolles-en-Brie, 1810-1932. The page is titled 'NOMENCLATURE' and contains a table with columns for 'NOMS, PRÉNOMS, PROFESSIONS ET DÉSIGNATIONS', 'INDICATION', 'CONTRIBUTION PROPRETIÉRE', 'REVENU', and 'FOLIO'. The table is filled with handwritten entries. A red box highlights the first row, which is the name 'GUILLET'. A blue box highlights the first column, which contains the names of the contributors. A green box highlights the second column, which contains the section names. A yellow box highlights the third column, which contains the parcel numbers. A pink box highlights the fourth column, which contains the nature of the parcels. A light blue box highlights the fifth column, which contains the dates of validity.

A historical cadastral register page from Marolles-en-Brie, 1810-1932. The page is titled 'NOMENCLATURE' and contains a table with columns for 'NOMS, PRÉNOMS, PROFESSIONS ET DÉSIGNATIONS', 'INDICATION', 'CONTRIBUTION PROPRETIÉRE', 'REVENU', and 'FOLIO'. The table is filled with handwritten entries. A red box highlights the first row, which is the name 'GUILLET'. A blue box highlights the first column, which contains the names of the contributors. A green box highlights the second column, which contains the section names. A yellow box highlights the third column, which contains the parcel numbers. A pink box highlights the fourth column, which contains the nature of the parcels. A light blue box highlights the fifth column, which contains the dates of validity.

A historical cadastral register page from Marolles-en-Brie, 1810-1932. The page is titled 'NOMENCLATURE' and contains a table with columns for 'NOMS, PRÉNOMS, PROFESSIONS ET DÉSIGNATIONS', 'INDICATION', 'CONTRIBUTION PROPRETIÉRE', 'REVENU', and 'FOLIO'. The table is filled with handwritten entries. A red box highlights the first row, which is the name 'GUILLET'. A blue box highlights the first column, which contains the names of the contributors. A green box highlights the second column, which contains the section names. A yellow box highlights the third column, which contains the parcel numbers. A pink box highlights the fourth column, which contains the nature of the parcels. A light blue box highlights the fifth column, which contains the dates of validity.

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) :

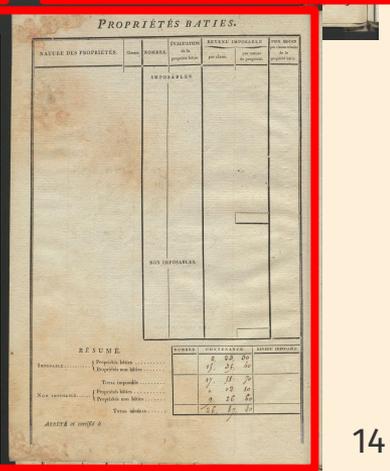
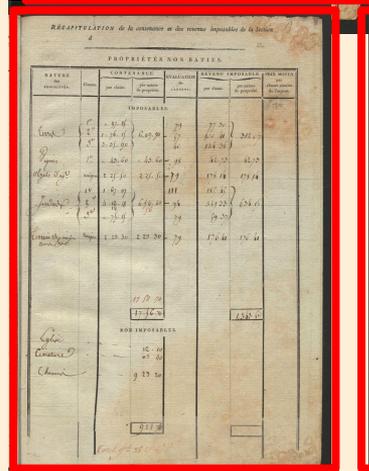
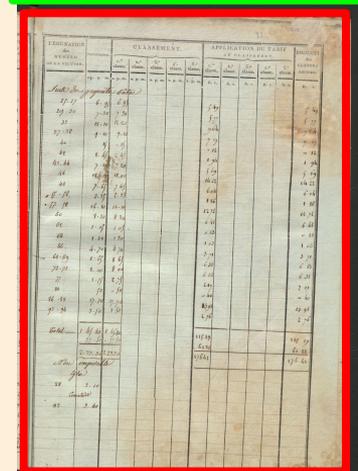
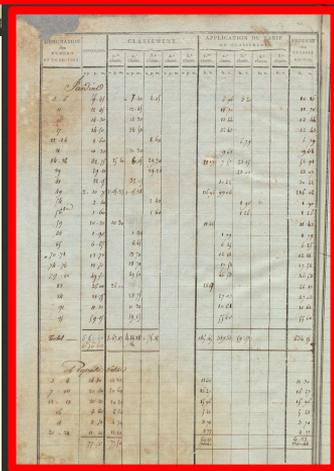
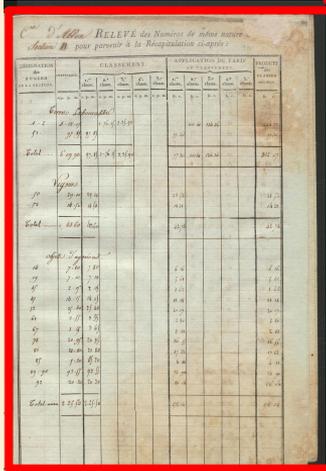
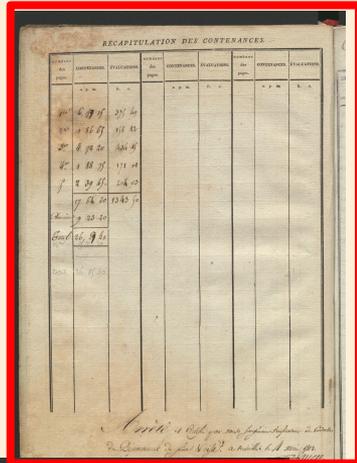
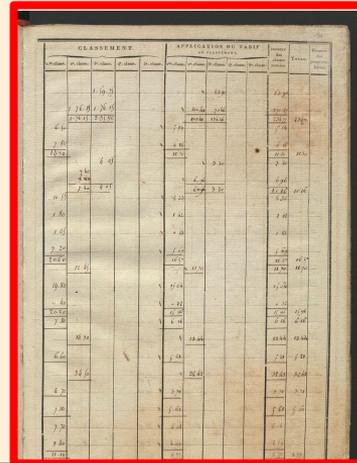
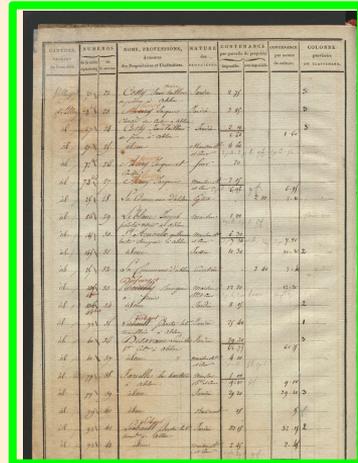
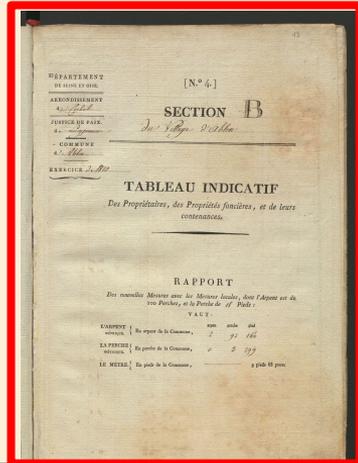


Détection des tables

Exemple :
classification de pages

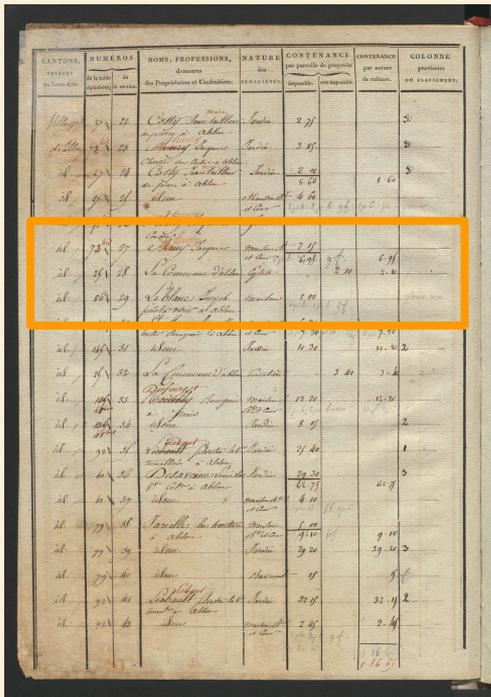
Différents types de pages :
couverture, tableau principal, récapitulatif
intermédiaire, résumé, ...

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



Transcription structurée

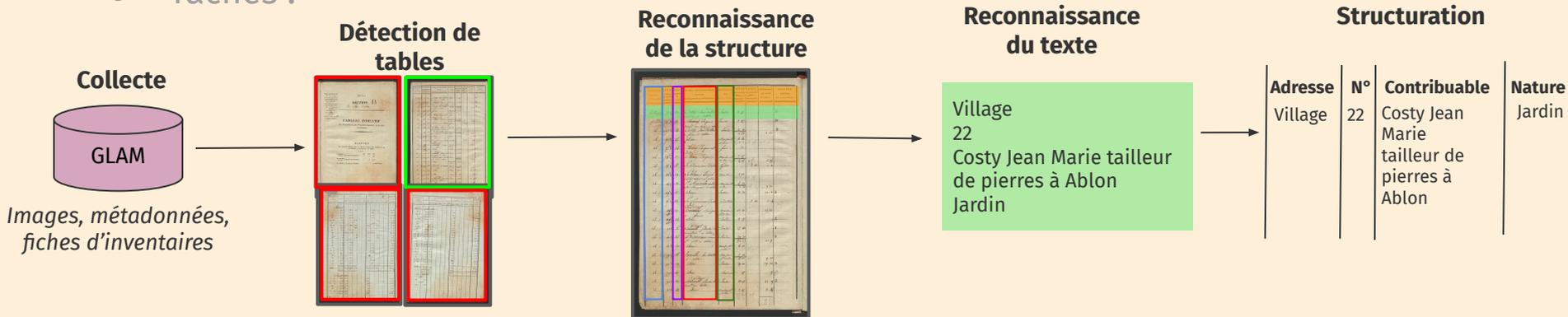


Numéro de parcelle	Adresse	Contribuable	Numéro du contribuable	Nature
				
..
27	id	houry Amy Jacques	72 ^{bis}	Maison B ^t et Cour
28	id	La Commune d'ablon	25	Eglise
29	id	Leblanc Joseph Peintre vitrier à ablon	86	Maison
..

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

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

- Tâches :



- Approches :

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

Approches séquentielles

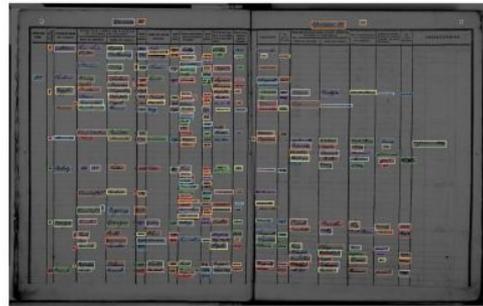
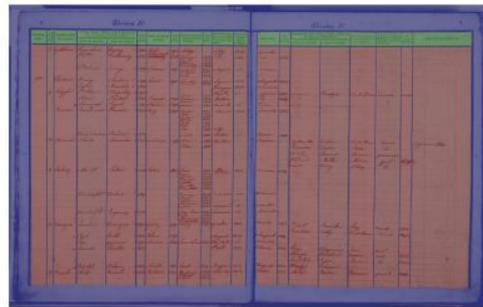
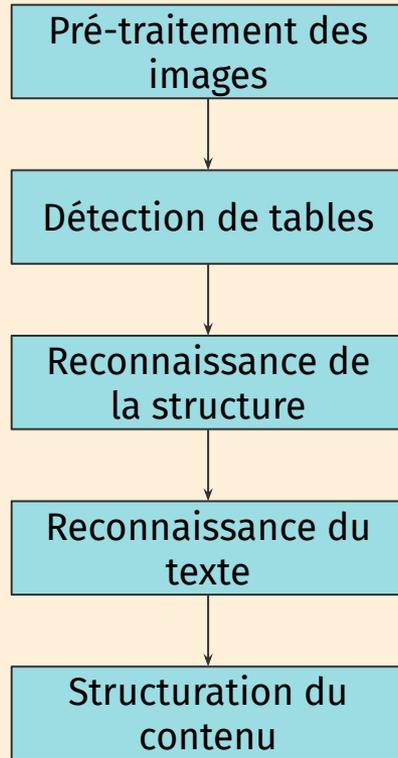


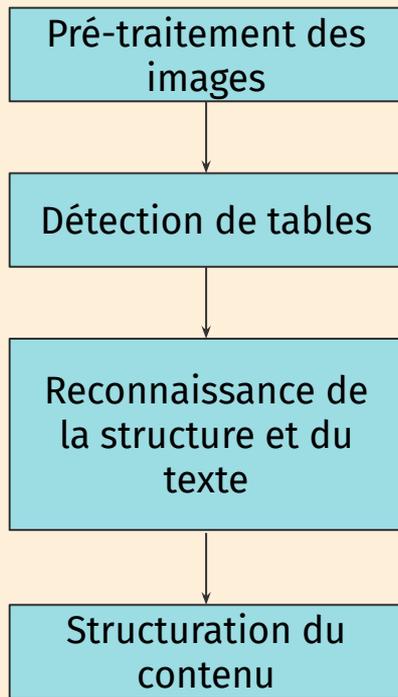
Fig. 3 Results from the main steps of the extraction pipeline on the second page of the 1832 census

[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



DAN [2],
DANIEL [3]

DÉSIGNATION		NUMÉROS PAR QUARTIER, VILLAGE BOULEVARD OU RUE			NOMS DE FAMILLE	PRENOMS	AGE	NATIONA- LITÉ	SITUATION PAR RAPPORT au chef de ménage	PROFESSION	12
des quartiers villages ou boulevards	DES NÉES dans les villes	des maisons	des maisons	des individus	6	7	8	9	10	11	12
Neully-le Réal	411	58	1861	Gendre	Pierre	75	fr.	chef	cult.	pat.	
		59	1862	Paraud	Marie	66	2	épouse	néant	mère	
		59	1863	Martin	Pierre	69	7	chef	métayer	pat.	
		59	1864	Joyoz	Suzanne	72	7	mère	néant	mère	
		59	1865	Martin	André	38	3	chef	métayer	pat.	

Fig. 5: Table header and first rows of a table from the census of the commune of Neully-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. Coquenat, 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 <i>end-to-end</i>
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] :

7 juin	Bâle	Suisse	0	1	Tchéquie
7 juin	Genève	Portugal	2	0	Turquie
11 juin	Genève	Tchéquie	1	3	Portugal
11 juin	Bâle	Suisse	1	2	Turquie
15 juin	Bâle	Suisse	2	0	Portugal
15 juin	Genève	Turquie	3	2	Tchéquie

Swiss national football team (Q165141) - CEA

city of Switzerland (Q54935504) - CTA

number of points /goals/set scored (P1351) - CPA

UEFA Euro 2008 (Q241864) - Topic Annotation

Switzerland - Turkey, 11 Jun 2008 (Q12012827) - Row-to-instance

Illustration des tâches de STI. Source: [1]

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

- 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**

[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

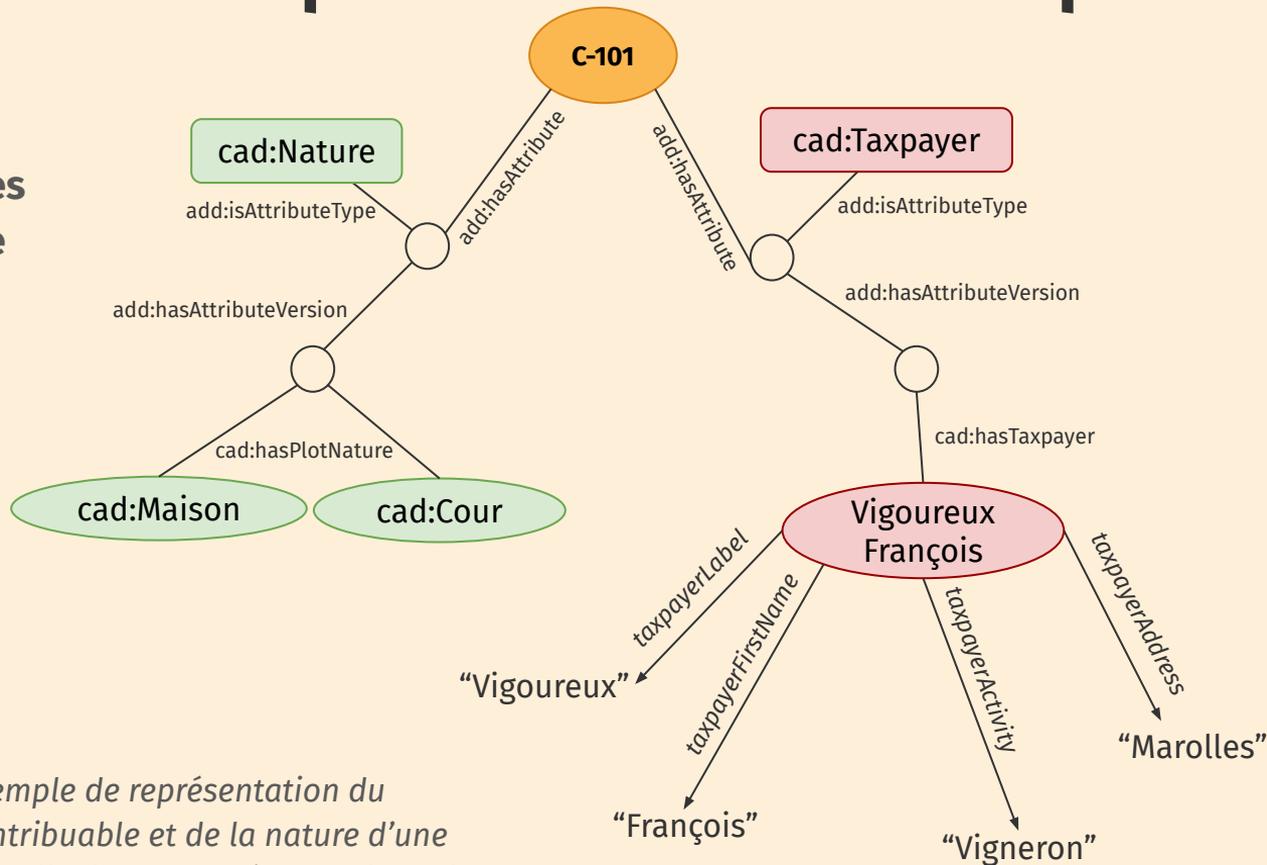
- Travaux nombreux pour annoter des **tables relationnelles ou d'entités**
 - Tables à la structure complexe, aux sujets cachés ou composés ?
- **Annotation souvent réalisée avec des graphes de connaissances encyclopédiques**
 - Graphes de connaissances de domaine ?
- Évaluations réalisées avec des **jeux 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

CANTON	NUMEROS	NOM S.	NATURE	contenance	contenance
TRAVERS	des P.L.S.	propriétaire, nature	de	par	personne
Tiers des	lignes	de l'occupation	Propriété	Propriété	de culture
	101	Vigoureux François vigneron	4 62	0 8 2 1/2	
	102	Vigoureux François vigneron	4 30	8 30	
	103	Vigoureux François vigneron	4 30	8 30	
	104	Vigoureux François vigneron	4 30	8 30	
	105	Vigoureux François vigneron	4 30	8 30	
	106	Vigoureux François vigneron	4 30	8 30	
	107	Vigoureux François vigneron	4 30	8 30	
	108	Vigoureux François vigneron	4 30	8 30	
	109	Vigoureux François vigneron	4 30	8 30	
	110	Vigoureux François vigneron	4 30	8 30	
	111	Vigoureux François vigneron	4 30	8 30	
	112	Vigoureux François vigneron	4 30	8 30	
	113	Vigoureux François vigneron	4 30	8 30	
	114	Vigoureux François vigneron	4 30	8 30	
	115	Vigoureux François vigneron	4 30	8 30	
	116	Vigoureux François vigneron	4 30	8 30	
	117	Vigoureux François vigneron	4 30	8 30	
	118	Vigoureux François vigneron	4 30	8 30	
	119	Vigoureux François vigneron	4 30	8 30	
	120	Vigoureux François vigneron	4 30	8 30	



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

Avant 1822

CANTONS, PRÉLÈVES ou lieux-dits.	NUMÉROS de la table hypothécaire de section.	NOMS, PROFESSIONS, demeures des Propriétaires et Usufructuaires.	ATURE des propriétés, ou nées, imposés, non imposés.	CONTENANCE par parcelle de propriété par nature de culture.	provisionnaire DU CLASSEMENT.
Chemin de Buis	95. 66	Lefèvre	terre	275 10	1
Chemin de Buis	25. 67	Buisson jeune	terre	11 25	1
	29. 68	Caron	terre	96 10	1
	95. 69	Lefèvre	terre	2590	1
	25. 70	Buisson jeune	terre	07 20	1
	25. 71	Servais	terre	05 10	1
	81. 72	Servais	terre	06 15	1
	81. 73		terre	06 10	1

Après 1822

NOMS, PRÉNOMS, PROFESSIONS ET DEMURES DES PROPRIÉTAIRES.	N° de parcelle.	CANTONS ou lieux-dits.	NATURE DES PROPRIÉTÉS.	CONTENANCE.	CLASSE.	REVENU.
Paris Denis vicent	166	La Ferté	terre	2 1/2	1	10
Millets guérin guillaume	167		terre	1 1/2	1	10
Paris Jacques	168		terre	1 1/2	1	10
Millets Jean Pierre	169		terre	2 1/2	1	10
Millets Jean Louis	170		terre	2 1/2	1	10
Millets Jean Louis	171		terre	2 1/2	1	10
Millets Jean Louis	172		terre	2 1/2	1	10
Millets Jean Louis	173		terre	2 1/2	1	10
Millets Jean Louis	174		terre	2 1/2	1	10
Millets Jean Louis	175		terre	2 1/2	1	10
Millets Jean Louis	176		terre	2 1/2	1	10
Millets Jean Louis	177		terre	2 1/2	1	10
Millets Jean Louis	178		terre	2 1/2	1	10
Millets Jean Louis	179		terre	2 1/2	1	10
Millets Jean Louis	180		terre	2 1/2	1	10

Second cadastre la Seine (1835-1845)

NOMS, PRÉNOMS, PROFESSIONS ET DEMURES DES PROPRIÉTAIRES.	N° de parcelle.	CANTONS, PRÉLÈVES ou lieux-dits.	NATURE DES PROPRIÉTÉS.	INDICATION DES SURS DE LA CONTENANCE, DE LA CULTURE ET DU REVENU DES PARCELLES.		
				ÉTENDUE.	CONTENANCE.	REVENU.
Millets Jean Louis	181		terre	2 1/2	1	10
Millets Jean Louis	182		terre	2 1/2	1	10
Millets Jean Louis	183		terre	2 1/2	1	10
Millets Jean Louis	184		terre	2 1/2	1	10
Millets Jean Louis	185		terre	2 1/2	1	10
Millets Jean Louis	186		terre	2 1/2	1	10
Millets Jean Louis	187		terre	2 1/2	1	10
Millets Jean Louis	188		terre	2 1/2	1	10
Millets Jean Louis	189		terre	2 1/2	1	10
Millets Jean Louis	190		terre	2 1/2	1	10
Millets Jean Louis	191		terre	2 1/2	1	10
Millets Jean Louis	192		terre	2 1/2	1	10
Millets Jean Louis	193		terre	2 1/2	1	10
Millets Jean Louis	194		terre	2 1/2	1	10
Millets Jean Louis	195		terre	2 1/2	1	10
Millets Jean Louis	196		terre	2 1/2	1	10
Millets Jean Louis	197		terre	2 1/2	1	10
Millets Jean Louis	198		terre	2 1/2	1	10
Millets Jean Louis	199		terre	2 1/2	1	10
Millets Jean Louis	200		terre	2 1/2	1	10

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	Mucrèau 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
§	39	164	Vautier Lecis Jean S58	terre
§	40	109	Lepriase V↑e↓ setustiers à Vig en	terre
§	411	89	Seurré Louison	terre
§	42	40	David philippe Vig↑on↓	terre
§	43	26	Bouguet Veuve	terre
§	44	26	Benard Lauteur→Vig VN	terre
§	414	211	Robin pierre	Vigne

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

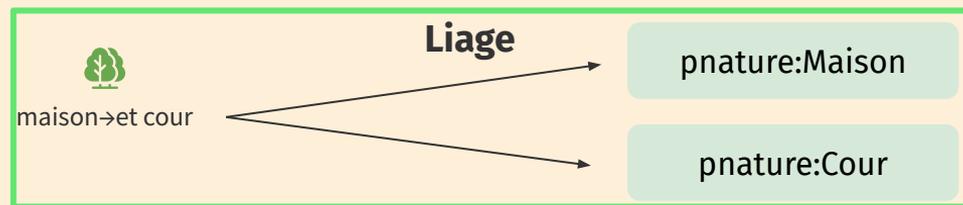
[1] Tual et al. An end-to-end pipeline for knowledge graph population from 19th-century land registry digitised tables. TPD125, Sep 2025, Tampere, Finland. <https://hal.science/hal-05118320>

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

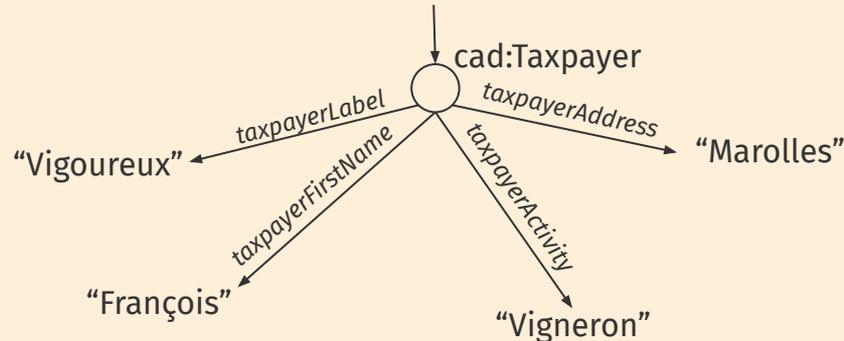
CANTON TRIAGES, ou Lieux dits.	NUMÉROS DU PLAN provisoire, définitif.	N O M S, PROFESSIONS, DEMEURES des Propriétaires ou Usufruitiers.	NATURES des Propriétés.	CONTENANCE par Propriété.	CONTENANCE par nature de culture.
				a. p. m.	a. p. m.
De Village	+ 101	Vigoureux François Marolles	Maison et cour	1 60	o. M. de B. ch.
	+ 102	Vigoureux Jean milhane	Maison et Cour	1 25	o. M. de B. ch.
	+ 103	Jean	Maison	7 90	2 85
	+ 104	Burroy Pierre Garde Champêtre	Maison	4 90	12 80
	+ 105	Jean	Maison	1 00	o. M. de B. ch.

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



Détection et création de nouvelles entités

Vigoureux François → vigon ↓ à marolles



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, EHES- CNRS

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



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

Manuscript page from 'ECOLE ROYALE VÉTÉRINAIRE' showing a botanical table. The table has columns for 'NOMS', 'PARTIES', 'BOTANIQUE', 'MATHÉMATIQUES', 'PHYSIQUES', 'CHIMIE', 'MÉTALLURIE', and 'OBSERVATIONS SUR LES MOEURS'. The text is handwritten and includes various botanical terms and observations.

Manuscript page showing a table with columns for 'LIGNON', 'VÉGÉTAL', 'NOM', 'SÉRIE', and 'OBSERVATIONS'. The table contains handwritten entries, likely related to botanical or agricultural studies.

Manuscript page showing a table with columns for 'NOM', 'SÉRIE', 'OBSERVATIONS', and 'REMARQUES'. The table contains handwritten entries, likely related to botanical or agricultural studies.

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. **Coquenot**, 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.

