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, Ecole des Hautes Etudes en Sciences Sociales

{solenn.tual, nathalie-f.abadie, julien.perret}@ign.fr, joseph.chazalon@epita.fr, bertrand.dumenieu@ehess.fr

Résumé

Les documents historiques contenant des tables représentent une source d'informations précieuse dans divers domaines. Si les institutions patrimoniales numérisent massivement ces documents pour en faciliter l'accès, les connaissances structurées qu'ils contiennent demeurent difficilement accessibles faute de pouvoir être requêtées. Cet article propose une revue des méthodes d'extraction d'informations dans des tables historiques numérisées et d'interprétation sémantique de tables tout en identifiant leurs limites. Les défis et perspectives associés à chaque tâche sont identifiés afin de proposer une chaîne de traitement visant à extraire et à structurer les informations contenues dans des tables historiques sous la forme de graphes de connaissances.

Mots-clés

Interprétation sémantique de tables, Graphes de connaissances, Extraction d'informations, HTR, NER, Documents historiques.

Abstract

Historical documents containing tables represent a valuable source of information in various fields. Although heritage institutions are digitising these documents on a massive scale to facilitate access, the structured knowledge they contain remains difficult to access because it cannot be queried. This article reviews methods for extracting information from digitised tables and for semantically interpreting tables, while identifying their limitations. The challenges and perspectives associated with each task are identified in order to propose a processing chain aimed at extracting and structuring the information contained in historical tables in the form of knowledge graphs.

Keywords

Semantic table interpretation, Knowledge graphs, Information extraction, HTR, NER, Historical documents.

1 Introduction

Les fonds d'archives regorgent de documents contenant des tables, formes privilégiées d'organisation et de présentation des données prisées par les administrations, grandes productrices de registres. On en retrouve également communément dans les ouvrages imprimés, notamment à visée scientifique, comme dans des documents manuscrits divers tels que livres de comptes, contrats, index géographiques.

Longtemps peu valorisés par les institutions patrimoniales, cette masse de documents historiques tabulaires constitués d'une feuille à des dizaines de milliers de pages est aujourd'hui graduellement diffusée sous forme numérisée. Cette diffusion se limite toutefois généralement à un partage sur le Web des images des pages numérisées, sans possibilité d'interroger les tables qu'elles contiennent.

La sémantique d'une table est essentiellement portée par l'organisation tabulaire elle-même : les lignes guident le regroupement de l'information en unités cohérentes, tandis que les colonnes apportent une information sémantique fine à l'échelle des champs. Les différents éléments d'un tableau et l'information qu'ils contient sont caractérisées par une grande densité de mentions d'entités et des relations riches. Reconnaître et préserver cette organisation particulière tout en extrayant sa sémantique constitue donc un enjeu d'accès et de valorisation pour ces fonds d'archives. C'en est aussi un, majeur, pour les sciences sociales quantitatives, les grands corpus de documents historiques tabulaires restant largement sous-exploités en raison des coups prohibitifs de traitements qu'ils impliquent.

Faciliter l'accès aux connaissances structurées en tables et contenues dans des documents historiques tabulaires est l'enjeu de cet article.

Il existe des travaux d'extraction et représentation sémantique des informations contenues dans des documents historiques sous la forme de graphes de connaissances [27, 21, 36, 14]. Cependant, les tables anciennes ne font pas partie des types de documents considérés, ces derniers étant principalement des textes en prose, confrontés à des problématiques d'extraction d'informations qui leur sont propres.

L'extraction en masse des informations contenues dans des

documents historiques tabulaires est confrontée à de nombreux défis : forte variabilité de la structure des tables, mélange de textes manuscrits et imprimés, présence d'abréviations ou encore usage de vocabulaires spécifiques. En outre, une fois le texte brut de ces tables transcrit, il ne donne qu'une vue limitée des connaissances disponibles dans l'intégralité de la collection. Annoter sémantiquement les tables d'un corpus offre la possibilité de lier les informations à travers les pages et les documents pour constituer des bases de connaissances sérielles et cohérentes ouvrant la perspective d'analyses longitudinales à grande échelle.
L'association des méthodes d'extraction d'informations (IE) dans des documents et d'interprétation sémantique de tables (STI) semble particulièrement pertinente pour parvenir à cet objectif. Aussi, après un passage en revue des caractéristiques des tables dans les documents historiques (section 2), nous présentons les tâches et des méthodes d'extraction d'informations dans de tels documents (section 3). Nous décrivons ensuite les principales tâches et approches d'interprétation sémantique de tables (section 4). Après avoir identifié les perspectives et les verrous scientifiques restants pour exploiter conjointement ces deux disciplines, nous proposons une chaîne de traitement de documents historiques tabulaires réconciliant les techniques d'annotation à des fins d'extraction d'informations dans des images et celles d'interprétation sémantique de tables à l'aide d'un graphe de connaissances de domaine (section 5).

2 Tables dans les documents historiques

Les données considérées dans cet article sont des tables issues de documents historiques. Une table est une structure bidimensionnelle composée de n lignes et m colonnes. La cellule est le plus petit élément à l'intersection d'une ligne et d'une colonne. Un document historique est numérisé à partir d'une source physique papier, imprimée ou manuscrite, conservée dans un service d'archives, un musée ou encore une bibliothèque. Ces documents sont généralement décrits par des métadonnées qui fournissent des éléments de contexte souvent indispensables à leur compréhension. Les types de tables présentes dans des documents, historiques ou non, sont très variés. Les tables peuvent être classées dans deux catégories [24] : *les tables de mise en forme*, utilisées pour structurer du contenu sans cohérence sémantique, et les *tables classiques*, caractérisées par une forte cohérence entre les lignes et les colonnes et qui contiennent des connaissances interprétables. Nous ne nous intéressons ici qu'aux tables classiques pour lesquelles il est possible d'utiliser des systèmes d'interprétation sémantique de tables. Ces tables classiques peuvent être décrites selon trois dimensions [24], qui, quand elles sont combinées, permettent de définir finement différents types de tableaux :

- la **structure** : tables imbriquées ou divisées, cellules fusionnées, cellule contenant des énumérations de valeurs ;
- les **relations internes** entre les cellules, les lignes et les colonnes : tables relationnelles, tables d'entités, matrices ;
- l'**orientation** : tables horizontales, tables verticales, matrices.

Pour décrire des tables issues de documents historiques, cette taxonomie doit cependant être complétée par des éléments propres aux sources anciennes. Les tables considérées dans cet article sont originellement des documents "papier" — manuscrits, imprimés ou préimprimés et complétés manuellement — qui ont été conservés puis numérisés. Elles n'existent pas nativement dans un format numérique, comme c'est le cas des tables présentes sur le Web et traditionnellement considérées pour l'interprétation sémantique de tables. Ainsi, la **nature du document historique**, **sa structure et la place de la table dans celui-ci** ainsi que son **état de conservation** sont autant de paramètres qui modifient la compréhension de la table et l'accès aux informations qu'elle contient. Les types de documents historiques qui contiennent des tables sont très variés (cf. Figure 1). Par exemple, les registres utilisés pour le recensement ou le cadastre, les journaux de bord, les ouvrages scientifiques et les manuels contiennent des tables. La source peut être composée d'un seul document ou faire partie d'une collection plus vaste. Les registres issus d'une même collection ayant existé sur une longue période, comme le cadastre, contiennent des informations similaires dans des tables dont la structure varie d'une administration à l'autre et a évolué au cours du temps (cf. Figure 1 (a) et (b)). La table peut constituer l'élément principal du document ou servir de complément à d'autres contenus, tels que du texte ou des illustrations. Elle peut s'étendre sur plusieurs pages, constituer le corps principal du document ou alors ne couvrir qu'une partie d'une page. Les colonnes de la table sont généralement réparties sur une page simple ou sur une double-page. Un registre peut contenir une série de tables élémentaires (présentées sur une page simple ou double), susceptibles d'être agrégées en une seule en raison de leur structure commune. La fragmentation d'une table en plusieurs pages (et donc images) risque néanmoins de compliquer la reconstitution de sa structure. Par exemple, la reliure du registre est susceptible de masquer les colonnes voisines (et donc le texte). Elle peut également entraîner un décalage entre les pages d'une double-page, désalignant ainsi les cellules d'une même ligne. La structure initiale d'une table est souvent préimprimée. La complétion manuscrite des documents permettait aux scripteurs d'ajouter des lignes, colonnes ou cellules supplémentaires qui complexifient la lecture du document. L'état de conservation du document physique impacte l'accès à l'intégralité ou non des informations contenues dans la table. Par exemple, un registre qui contient une table distribuée dans plusieurs pages ne sera pas complet si une page est manquante ou que son contenu est illisible. Lors de la numérisation, il est essentiel de conserver l'ordre des pages afin de préserver la logique de lecture des informations.
Ainsi, bien que fortement structurées, les tables issues de documents historiques restent difficiles à exploiter dans leur totalité. Le texte qu'elles renferment ne devient acces-

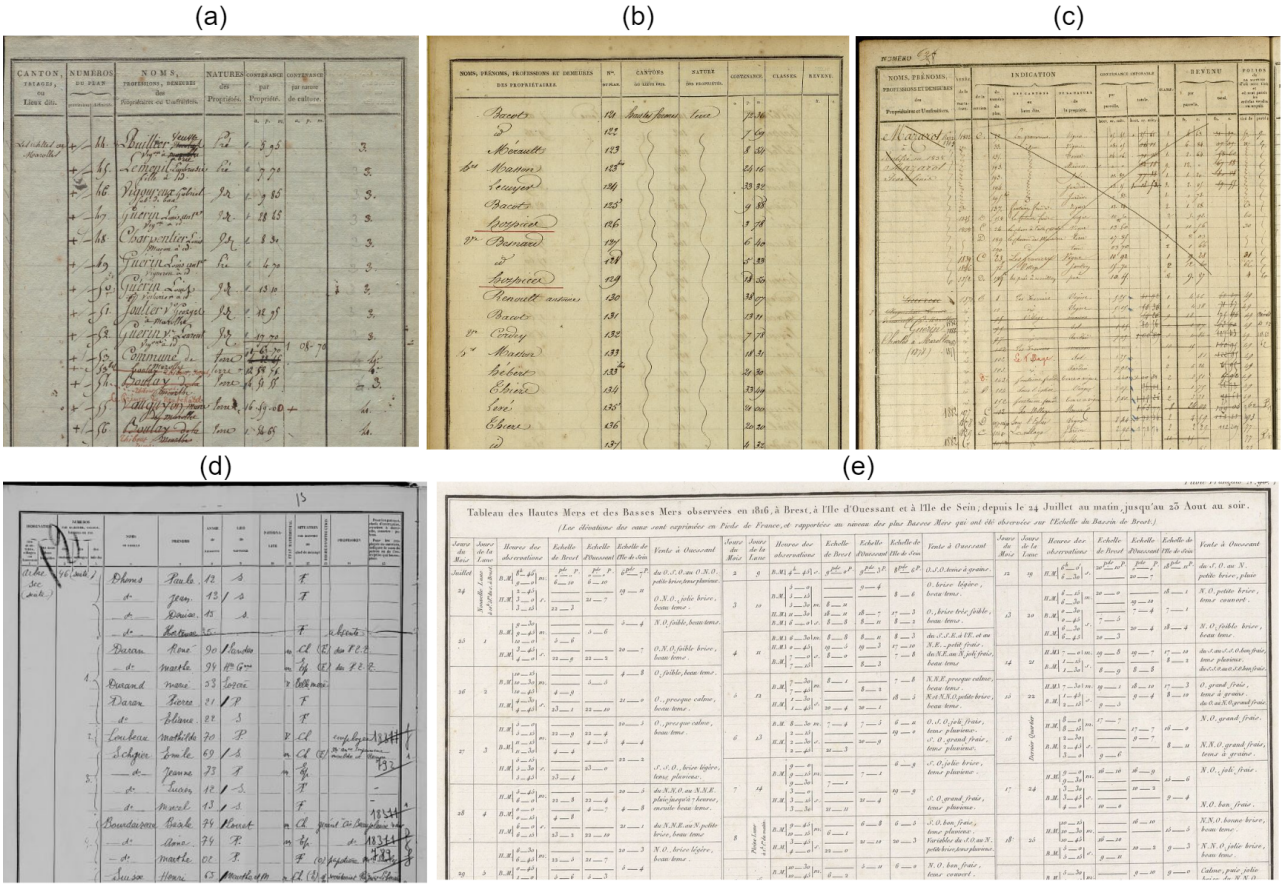

FIGURE 1 – **(a)** Page de registre d'états de sections de Marolles-en-Brie (1810) : table relationnelle horizontale. L'identité du sujet (parcelle) est partiellement masquée : l'identifiant de la section dans laquelle se trouve la parcelle est absent. Il faut se référer à la page de couverture du chapitre pour connaître l'identifiant de la section. **(b)** Page de registre d'états de sections d'Ivry (date inconnue, ultérieure à 1822) : même type de table que (a) mais l'ordre des colonnes est différent. **(c)** Matrice des contribuables de Marolles-en-Brie (1822-1914) : matrice, la colonne relative aux contribuables (première colonne ici) contient des cellules avec énumérations. Une page peut contenir plusieurs sous-tableaux relatifs aux propriétés de chaque contribuable. **(d)** Page de registre du recensement de Paris (1926). Des groupes de lignes doivent être formé pour reconstituer les ménages. **(e)** Tableau des hautes et basses mers observées en 1816 dans la région de Brest, extrait du Pilote français (tome 1, 1822)
Sources : Archives Départementales du Val-de-Marne, (a) 3P 387, (b) 3P 1631 et (c) 3P 389, (d) Archives de Paris D2M8 221, (e) Bibliothèque nationale de France, département Cartes et plans, GE CC-1194

sible aux machines qu'après une phase de transcription. Il doit être structuré pour permettre une recherche par attribut. Une fois cette étape franchie, les données obtenues ne fournissent que des informations partielles, souvent complexes à interroger et à croiser au sein d'une même collection. L'annotation sémantique du texte brut extrait constitue donc une approche pertinente pour valoriser pleinement le contenu de ces documents historiques.

3 Extraction d'informations dans des documents tabulaires historiques

L'extraction d'informations dans des documents numérisés a pour objectif de localiser, transcrire et organiser le texte qu'ils contiennent dans le but de traduire un contenu initialement exclusivement sous forme graphique vers une représentation informatique structurée exploitable automatiquement. Selon les approches, les étapes qui constituent ce processus peuvent être distinctes ou au contraire intégrées, comme c'est le cas dans les approches les plus modernes. En effet, ce domaine connaît des avancées significatives grâce au progrès de l'apprentissage profond.

3.1 Tâches et chaîne de traitement

Nous décrivons ci-dessous les principales tâches d'un processus d'extraction d'informations dans des documents historiques numérisés.

La classification des images permet d'identifier, au sein d'un répertoire de documents numérisés, celles qui nécessitent un traitement approfondi. En effet, les informations à extraire se trouvent souvent sur un type spécifique de pages. Lorsqu'elles sont disponibles, les métadonnées associées aux images décrivent généralement l'ensemble du répertoire sans distinction. Il est donc nécessaire de classer

les pages en fonction de leur nature (par exemple : couverture, tableau principal, récapitulatif intermédiaire, résumé) afin de ne conserver que celles pertinentes pour les étapes suivantes de la chaîne de traitement.

La reconnaissance de la mise en page du document consiste à comprendre la structure du document dans l'image. Par exemple, si l'image en entrée est une double-page où chaque page contient une partie des lignes d'une table, il faut détecter chaque page simple puis les différents éléments qui la composent (ligne ou colonne de la table, segment de texte).

La reconnaissance du texte (OCR/HTR) consiste à transcrire le texte. Pour des documents imprimés, les modèles utilisés sont des modèles d'OCR (*Optical Character Recognition*) alors que pour du texte majoritairement manuscrit ce sont des modèles d'HTR (*Handwritten Text Recognition*) qui sont utilisés. Le texte produit par ces systèmes contient généralement des erreurs de transcription, désignées sous le terme de "bruit".

La structuration des informations produites est le formatage final des données en sortie de la chaîne de traitement. Elle peut comprendre une phase de post-traitement du texte extrait et plus rarement d'enrichissement sémantique (reconnaissance des entités nommées, géocodage). Elle comprend également la production d'une base de données ou d'un fichier de données structurées (CSV, XML).

3.2 Approches

Il existe deux grands types de systèmes d'extraction d'informations : les approches classiques composées d'un enchaînement de modèles indépendants et les approches dites *end-to-end* où un même modèle réalise plusieurs étapes en une fois. Pour ces deux types d'approches, les régions contenant les tables sont généralement isolées au préalable grâce à un système de classification des pages ou de détection des régions.

Approches classiques Les approches classiques d'extraction d'informations dans des tables anciennes numérisées sont fondées sur un enchaînement de modèles où les résultats produits à une étape sont utilisés en entrée de l'étape suivante.

Après une phase de prétraitement des images (segmentation des double-pages en pages, redressage), *Constum et al. (2022)* [7] détectent la table dans la page, puis utilisent un modèle de segmentation pour extraire les lignes (sous la forme d'images) et enfin de les transcrire à l'aide d'un modèle d'HTR [10]. Un caractère spécial est utilisé pour matérialiser implicitement la séparation de la ligne en cellules. La structuration du contenu extrait est réalisée en post-traitement. La nature de chaque cellule est déduite de sa position dans la ligne. Ceci est possible car l'ordre des colonnes est connu *a priori* des auteurs et car la structure de l'ensemble des registres considérés ne varie pas.

Petit-Pierre et al. (2023) [28] proposent, à l'inverse, de détecter les colonnes et les segments de texte dans la page puis de regrouper ces segments en lignes *a posteriori*. Le texte est transcrit avec le modèle d'HTR proposé par *Puigcerver (2017)* [29].

Granell et al. (2023) [18] traitent des relevés météorologiques consignés dans des registres de navigation. La tâche de segmentation des lignes est réalisé par un modèle de classification des pixels [30]. La transcription du texte est réalisée à l'échelle de la ligne avec un modèle de type *Convolutionnal Recurrent Neural Networks* (CRNN).

Pour traiter des tables astronomiques datant du XIV^{e} au XVI^{e} siècle, *Eberle et al. (2024)* [15] détectent les chiffres dans l'image et les transcrivent avec un modèle d'HTR entraîné pour reconnaître uniquement les chiffres allant de 0 à 9, puis de recomposer les nombres correspondants à la valeur de chaque cellule en post-traitement.

Le principal défaut des approches avec segmentation préalable des documents est que les erreurs de découpage des pages en lignes ou colonnes se propagent aux étapes suivantes. Par ailleurs, une fois la segmentation effectuée, les modèles qui réalisent les tâches suivantes ne peuvent s'appuyer que sur un contexte visuel et textuel réduit [3]. Enfin, des jeux de données d'évaluation doivent être produits pour chaque étape.

Approches end-to-end *Boilet et al. (2024)* [3] traitent les registres du recensement de différents départements français. La chaîne de traitement comprend une étape de classification des pages numérisées avec YOLOv8, suivie de la transcription structurée du contenu des pages entières de tables avec le modèle Document Attention Network (DAN) [11], un encodeur convolutif suivi d'un décodeur Transformer [37]. Cette approche permet de retrouver les informations de même type sans segmenter l'image ni matérialiser la position exacte des lignes et des colonnes. Les annotations sont réalisées selon un modèle "Clé-Valeur" [35] : pour chaque zone de l'image (ligne), la liste des informations à extraire (correspondant à des colonnes ou à des entités nommées) est définie.

Constum et al. [9] introduisent DANIEL, une architecture dérivée de DAN [11] destinée à capturer l'organisation hiérarchique de l'information. Il a été testé pour extraire le contenu d'actes d'états civils manuscrits mais présente des perspectives intéressantes pour le traitement de tables.

Parallèlement aux approches spécialisées sur un type de document particulier, certaines approches telles que DONUT [22] ont visé à proposer des modèles entraînés sur un grand nombre de tâches afin de pouvoir s'adapter rapidement à de nouveaux problèmes ou types de documents. Ces approches ont été améliorées avec la réutilisation de grands modèles de langage (LLMs) existants dans la construction de modèles vision-langage (VLMs), disposant alors des capacités de raisonnement supérieures, mais montrant jusqu'à présent une performance limitée dans l'analyse de tables modernes, comme noté par *Scius-Bertrand et al.* [33] mi-2024. Tout récemment, le modèle GOT-OCR [39] a réalisé de grands progrès dans l'interprétation d'images de documents à la densité de texte élevée, ainsi que dans l'analyse de tables modernes, permettant leur transcription et plus simplement la réponse à des questions simples, laissant présager de nouvelles avancées dans l'analyse de do-

cuments historiques à court terme. Cependant, aucune approche n'arrive encore à s'affranchir d'un entraînement spécifique sur les documents historiques à analyser.

3.3 Synthèse et limites

Dans le domaine de l'extraction d'informations à partir de documents historiques, les modèles Transformer [37] ont favorisé l'émergence d'approches dites *end-to-end*. Celles-ci offrent de nouvelles perspectives : l'extraction d'informations peut être réalisée pour des pages entières sans segmentation préalable ni perte d'information contextuelle. Le texte transcrit est enrichi par des caractères spéciaux qui traduisent la typographie ou même la structure du document. Les jeux de données d'entraînement ne nécessitent pas de localisation de chaque élément du tableau dans l'image. Contrairement aux approches classiques, il n'est plus indispensable de constituer un jeu de données spécifique pour chaque tâche (détection de zones, transcription) pour entraîner les modèles. Un seul jeu de données complet, dont les annotations sont des fichiers de textes structurés traduisant le contenu et la structure de l'image, est suffisant. Le mécanisme d'attention [37] permet de localiser les informations dans l'image *a posteriori*. Ce type d'approches génère du contenu structuré lisible par une machine, traduction numérique du tableau représenté dans l'image. Cependant, ces approches ciblent principalement le traitement des tables relationnelles classiques. Les tables à la structure plus complexe, telles que les matrices, demeurent peu étudiées dans la littérature, de même que l'extraction de texte multi-orienté, le fait d'avoir un sens de lecture uniforme restant une dimension importante lors de la transcription. Par ailleurs, les architectures Transformers sont parfois sujettes à des hallucinations. Elles entraînent la génération de résultats erronés (mais souvent crédibles), tels que l'ajout de lignes supplémentaires dans une table. Les chaînes de traitement d'IE appliquées aux tables de documents historiques vont rarement au-delà de la structuration des données sous la forme de bases de données relationnelles [28, 7, 18, 3]. Ceci ne permet pas d'établir les liens entre des mentions d'entités identiques. Par ailleurs, les requêtes se font directement sur les chaînes de caractères qui sont souvent impactées par des erreurs du modèles de transcription utilisé.

4 Interprétation sémantique de tables

L'interprétation sémantique de tables (STI) est un processus qui consiste à annoter les différents éléments qui composent un tableau (ligne, colonne, cellule, tableau entier) et leurs relations à l'aide des ressources d'un graphe de connaissances (KG) pour améliorer l'interprétation de leur sémantique. L'objectif est de faciliter le développement d'applications mettant en œuvre les données qu'elles contiennent, comme la recherche d'informations, la réponse aux questions ou bien la création de bases de connaissances. Ces dernières années, les approches visant à résoudre les différentes tâches de STI se sont multipliées, stimulées par des compétitions comme SemTab [1]. Ces compétitions ont mené à une définition précise des tâches et des méthodes d'évaluation. Les deux principales applications de la STI sont d'enrichir les données représentées dans une table à l'aide de connaissances supplémentaires et de créer ou de compléter des KGs avec des informations issues de ces tables. Les approches proposées dans la littérature traitent généralement des fichiers tabulaires (CSV) ou des tables HTML.
L'étude des tâches, des étapes et approches d'interprétation sémantique de tables présentées ici est une synthèse des articles d'état de l'art de *Liu et al. (2023)* [24] et *Cremashi et al. (2024)* [13]. *Liu et al. (2023)* établissent une taxonomie détaillée des types de tables et de métadonnées, une présentation des tâches de STI et des graphes de connaissances généralistes. Les jeux de données d'évaluation, les métriques et une synthèse des grandes étapes de la chaîne de traitement sont passés en revue ainsi que les différents types d'approches développées pour réaliser les tâches recensées (jusqu'en 2021). Une comparaison des performances des principales approches pour effectuer chaque type de tâche est réalisée. *Cremashi et al. (2024)* [13] proposent une nouvelle taxonomie très détaillée des tâches et approches de STI (de 2007 à octobre 2024) reposant sur 31 critères ainsi qu'une comparaison des outils existants. Les travaux les plus récents utilisant les modèles de langages pré-entraînés (comme BERT) et les grands modèles de langages (LLMs) sont pris en compte. La chaîne de traitement est finement détaillée, incluant davantage d'étapes que dans l'article de *Liu et al. (2023)* [24].
Nous renvoyons à ces publications le lecteur qui souhaite entrer dans le détail des approches de STI.

4.1 Tâches

L'annotation de colonnes avec des types (CTA) consiste à attribuer un type à chaque colonne de la table. La distinction entre les colonnes correspondant à des entités, qui seront associées à des classes de l'ontologie, et les colonnes correspondant à des valeurs alphanumériques (dates, nombres), est particulièrement utile pour la réalisation de cette tâche [13].

L'annotation de colonnes avec des propriétés (CPA) vise à associer une propriété du KG à une paire de colonnes pour caractériser leur relation.

L'annotation de cellules avec des entités (CEA) consiste à associer le contenu d'une cellule (aussi appelée mention) à une entité du KG.

La détection des cellules qui décrivent de nouvelles entités (CNEA) est une tâche définie plus récemment par *Cremaschi et al. (2024)* [13]. Elle consiste à détecter des mentions comme étant des entités absentes du graphe de connaissances [26] utilisé pour l'annotation. Dans la littérature, ces cellules sont souvent annotées avec le terme *NIL (Not In Lexicon)*.

La thématisation consiste à associer un concept du KG à l'ensemble de la table ;

1. https://www.cs.ox.ac.uk/isg/challenges/sem-tab/

La correspondance ligne-instance consiste à annoter une ligne de la table avec une entité du KG considérée comme le sujet de cette ligne. De ce fait, cette tâche est surtout pertinente dans le cas où la table considérée est une table relationnelle.
La tâche de CEA est particulièrement utile pour désambiguïser des entités mentionnées dans une table. Elle peut être assimilée à la tâche d'*Entity Linking* dans le domaine du traitement automatique du langage naturel [13] (TALN).
Les tâches de CTA et CPA permettent de transformer automatiquement ou semi-automatiquement des tables en reliant leur structure au schéma de données défini par une ontologie [13].

4.2 Chaîne de traitement

Cette section présente les étapes d'un processus complet d'interprétation sémantique de tables tel que décrit par *Cremashi et al. (2024)* [13].

La préparation des données comprend une phase de standardisation des formats de tables, de nettoyage et de normalisation du contenu des cellules et de compréhension du type de table considéré. Les types de traitements appliqués dépendent des valeurs alphanumériques contenues dans la table. Par exemple, il peut s'agir de correction d'erreurs de frappes [5], de suppression de contenus inutiles [1], d'uniformisation de la casse, de conversion des unités de mesures [16, 12], d'expansion des abréviations [1] ou de conversion de formats de dates ou de coordonnées géographiques [31]. Ces prétraitements facilitent et améliorent la qualité de l'interprétation des tables considérées ;

La classification des colonnes vise à distinguer les colonnes dont les cellules contiennent des entités du graphe et celles qui correspondent à des valeurs alphanumériques (chaîne de caractères, nombre, date) [41, 20]. Cette étape permet de sélectionner les colonnes pour lesquelles les tâches de CEA et de CTA doivent être réalisées.

L'annotation des colonnes contenant des valeurs alphanumériques a pour objectif de déterminer le type des données contenues dans les colonnes qui ne contiennent pas d'entités (dates, nombres et leurs unités, coordonnées géographiques) [5, 12]. Cette étape constitue une partie de la tâche de CTA.

La détection du sujet vise à déterminer la colonne de la table contenant l'entité qui est aussi le sujet de la ligne [4, 17]. Cette étape est particulièrement utile pour la tâche de CPA, car elle identifie la colonne racine de nombreuses propriétés [13].

La résolution d'entités correspond aux tâches de CEA et CNEA. La CEA peut être décomposée en trois sous-tâches : la détection des mentions dans la table, l'identification des entités du KG candidates au liage et des mentions sans entités correspondantes dans le KG ainsi que la désambiguïsation d'entités. La détection de mentions nécessite souvent une structuration plus fine du contenu d'une cellule, par exemple avec une étape de reconnaissance des entités nommées (NER) [13]. L'identification des entités candidates revient à sélectionner un sous-ensemble de ressources du KG et à les indexer suivant divers critères. Enfin, la désambiguïsation vise à départager des entités candidates très similaires. Le contexte s'avère particulièrement utile pour résoudre cette sous-tâche.

La prédiction des types des colonnes contenant des entités est le résultat final de la CTA. Cette étape peut comprendre une phase de présélection des types des entités de chaque colonne ou être le résultat de prédictions d'un algorithme entraîné avec des données similaires [31].

La prédiction des prédicats associés à des paires de colonnes est le résultat final de la CPA.

4.3 Approches

Il existe trois grandes familles d'approches d'interprétation automatique de tables [24]. Elles peuvent être utilisées pour traiter les différentes étapes et tâches de la chaîne de traitement décrites précédemment.

Les approches heuristiques Pour associer un élément de la table à la ressource du KG jugée la plus pertinente, les approches heuristiques s'appuient sur les métriques et les critères de décisions usuels en recherche d'information : similarités, TF-IDF, vote majoritaire ou encore méthodes probabilistes. Ces approches sont utilisées pour résoudre les tâches de CEA, CTA et de CPA. Elles peuvent introduire ou non une phase de reclassement des candidats qui améliore généralement les résultats.

L'ingénierie de caractéristiques repose sur l'extraction de caractéristiques lexicales et statistiques des éléments du tableau qui sont ensuite utilisées pour entraîner des modèles d'apprentissage classique (SVM, K-NN ou Random Forest). La tâche la plus traitée par ces méthodes est la CTA. Ces méthodes nécessitent de disposer de jeux de données pour entraîner les modèles utilisés.

L'apprentissage profond est basé sur l'entraînement de réseaux de neurones profonds. Nous pouvons distinguer deux groupes de méthodes [24] : l'apprentissage de plongements représentant les éléments de la table (cellules, lignes, colonnes) afin de les comparer dans l'espace vectoriel et l'apprentissage de plongement des entités du graphe. Comme pour l'ingénierie des caractéristiques, des jeux de données doivent être produit pour entraîner les modèles. Les méthodes basées sur l'utilisation des LLM connaissent actuellement de multiples développements [40].

4.4 Synthèse et limites

Malgré les avancées récentes, il demeure encore de nombreux défis à résoudre pour annoter sémantiquement des tables complexes comme celles issues de documents historiques. Comme pour l'IE, la plupart des approches se concentrent sur l'annotation de tables à la structure simple, principalement des tables d'entités ou tables relationnelles horizontales à sujet monocellulaire. Or, il existe une grande diversité de tables aux structures plus complexes, notamment parmi les tables de documents historiques. Nous identifions par exemple des structures imbriquées ou à sujets cachés ou composés (identifiant à reconstituer à partir de plusieurs colonnes ou à l'aide du contexte). La grande ma-

jorité des approches de STI existantes ont été évaluées avec des jeux de donnés d'état de l'art, comme T2Dv2 [32] et Limaye [23]. Ceci est utile pour comparer les approches entre elles, mais ne considère pas l'adaptation à des tables originales. Par ailleurs, les méthodes existantes utilisent principalement des graphes de connaissances encyclopédiques comme Wikidata [38], DBpedia [25], ou YAGO [34] pour l'annotation. Ces KG ne sont pas pertinents pour traiter des documents historiques qui contiennent des entités inconnues de ces bases. Les approches qui considèrent l'incomplétude des graphes de connaissances qu'elles utilisent sont encore rares [13]. Cette hypothèse est pourtant indispensable pour annoter des registres anciens qui décrivent des entités majoritairement non encyclopédiques. Enfin, l'utilisation du contexte de la table, de ses métadonnées est également insuffisamment développée alors qu'elle est identifiée comme bénéfique aux différentes tâches [24].

5 Interprétation sémantique de tables historiques : défis et perspectives

Nous pouvons définir l'interprétation sémantique de tables historiques comme un processus qui combine l'extraction d'informations dans des documents tabulaires historiques et l'annotation sémantique de tables dans une même chaîne de traitement. La revue des travaux existants montre qu'il n'existe pas de chaîne de traitement intégrant des approches de ces deux domaines pour extraire et gérer des connaissances contenues dans les tables anciennes. Leur articulation est pourtant particulièrement pertinente pour exploiter et interroger ce type de sources historiques en détail. Des connaissances isolées dans des documents tabulaires historiques peuvent être mises en relation dans un graphe afin de parcourir une collection de documents non plus d'images en images, mais d'entités en entités. Ceci facilite notamment la recherche d'informations dans des collections d'images de très grandes tailles.

Dans cette section, nous proposons une chaîne de traitement combinant extraction d'informations et interprétation sémantique de tables anciennes dans une perspective de peuplement de graphe de connaissances historiques puis nous recensons les défis qui demeurent pour y parvenir. Dans cette chaîne de traitement dont les étapes sont décrites ci-après, on suppose qu'une ontologie est déjà disponible pour représenter les connaissances contenues dans les tables.

La collecte des tables consiste à réunir les documents numérisés issus d'une ou plusieurs collections, accompagnés de leurs métadonnées. Ces métadonnées fournissent un contexte global sur les tables à traiter et peuvent être restructurées [3] afin d'en faciliter l'exploitation, notamment lorsque leurs structures varient. Ce sont des sources de connaissances privilégiées pour thématiser les tables à traiter et ainsi faciliter leur classification à l'étape suivante.

La classification des images est une étape récurrente dans les chaînes d'IE étudiées. Elle est nécessaire pour identifier les images qui contiennent les tables à traiter. Elle peut bénéficier d'une étape de thématisation préalable des répertoires d'images à l'aide de leurs métadonnées. La classification des images (pages de couvertures, résumés intermédiaires, synthèses) est utile pour fournir davantage de contexte aux tables et parfois pour compléter les informations qu'elles contiennent. Ces connaissances additionnelles peuvent venir compléter la thématisation initiale réalisée à l'aide des métadonnées.

La reconnaissance de la structure du tableau et du texte, réalisée avec des approches *end-to-end* utilisant des architectures Transformer comme DAN [11] ou DANIEL [8] permet d'extraire le contenu de chaque page sans segmentation des images et sans localisation du texte au préalable. Grâce à ce fonctionnement intégré, ces approches évitent la propagation d'erreurs d'une tâche d'IE à l'autre tout en permettant un gain de temps significatif pour produire des jeux de données d'entraînement [3]; la production de données intermédiaires n'étant plus nécessaire.

En outre, ces approches permettent de résoudre, manuellement, les tâches de CTA et CPA lors de la définition du modèle d'annotation des images. En effet, l'utilisateur souhaite généralement extraire un ou plusieurs attributs des objets décrits dans la table. Ces attributs correspondent très souvent aux colonnes et sont généralement décrits sous la forme de classes dans une ontologie de domaine produite à partir de connaissances expertes des documents et du domaine auquel ils appartiennent. Aussi, les clés du modèle d'annotation "Clé-Valeur" utilisé pour produire les jeux de données d'entraînement de DAN correspondent à des classes de l'ontologie ou à des valeurs alphanumériques attendues, tandis que les valeurs du modèle d'annotation "Clé-Valeur" correspondent quant à elles, pour une ligne donnée, au contenu des cellules situées dans les colonnes considérées.

Par ailleurs, le mécanisme d'attention d'architectures telles que DAN permet d'absorber les variations d'emplacement et d'intitulés des colonnes qui contiennent les mêmes informations dans des tables appartenant à une même collection.

Le cas des colonnes qui peuvent contenir plusieurs types d'entités ou de valeurs alphanumériques constitue, en revanche, un défi qui reste à traiter. L'identification automatique du type des colonnes et des propriétés associées, dans le cas de documents aussi spécifiques que les documents historiques, reste également non résolue ce qui limite les possibilités de traitement massif de document tabulaires anciens sans intervention humaine.

Notons cependant que la qualité de la transcription fournie par le modèle peut avoir des conséquences importantes sur les étapes d'interprétation sémantique réalisées en aval de la chaîne de traitements.

La structuration des informations produites est très probablement l'étape qui peut bénéficier le plus des apports des approches d'interprétation sémantiques de tables. Celles-ci permettent en effet d'enrichir la transcription brute produite à l'étape précédente à l'aide d'approches de détection de sujet, de CEA et de CNEA permettant de proposer une structuration fine des données en sortie qui s'abstrait de celle de

la table en entrée, en particulier de l'ordre des colonnes, et facilite leur exploitation future.

Le cas de tables relationnelles, où le sujet correspond à une ligne, est certainement le plus simple à traiter. Il est cependant possible que l'identification unique du sujet ne soit réalisable qu'en intégrant des connaissances de thématisation produites lors des étapes de collecte et de classification des images ou en combinant les valeurs de plusieurs colonnes.

Les tâches de CEA et de CNEA permettent à cette étape de lier le texte brut produit par le modèle d'IE à des ressources du KG (notamment des concepts de taxonomies) ou de créer des entités qui vont être ajoutées au KG. Ces entités pourront ultérieurement être utilisées pour annoter de nouvelles tables. La principale difficulté pour leur mise en œuvre réside dans l'automatisation du choix de la méthode la plus appropriée selon le type d'entités ou de valeurs alphanumériques présents dans les différentes colonnes de la table transcrite.

De la même façon, dans le cas de nouvelles entités à intégrer au KG et à structurer au préalable, une étape de reconnaissance des entités nommées peut s'avérer nécessaire pour structurer des descriptions de personnes par exemple. L'automatisation du paramétrage de cette étape reste un verrou pour son exécution automatique sur de nombreux documents, sans intervention humaine.

L'identification automatique des entités mentionnées à plusieurs reprises dans une même table peut bénéficier des approches de détection de coréférence dans des documents [2] qui semblent particulièrement pertinentes pour rassembler les mentions d'un même objet qui apparaîtraient à plusieurs reprises et créer la ressource RDF correspondante selon le schéma de données décrit dans l'ontologie. Une piste intéressante pour réduire les erreurs liées à l'homonymie serait de contraindre la détection de coréférence à un sous-ensemble cohérent de documents au sein d'une même collection, par exemple dans un même registre ou un même espace géographique.

Lien aux sources La réalisation automatique des tâches d'IE et de STI permet ainsi de traiter de grandes collections d'images et de structurer les informations qu'elles contiennent sous la forme d'un graphe de connaissances. Cependant, maintenir le lien entre cette représentation dérivée du contenu de l'image et l'image elle-même demeure indispensable. L'image transmet des éléments de contexte utiles aux chercheurs en sciences humaines et sociales. Elle permet par ailleurs de confronter les informations lues aux données produites lors du processus automatique. Il est donc indispensable de développer des outils qui facilitent l'exploration des données par tous types d'utilisateurs. Le standard IIIF (*International Image Interoperability Framework*) [6] offre un ensemble de techniques pour diffuser des annotations d'images structurées selon le *Web Annotation Data Model* [19]. Ce standard est très utilisé par les services d'archives, les bibliothèques et les musées. Aussi, il serait particulièrement pertinent de visualiser les fragments de pages sur les images et d'afficher leur contenu enrichi sémantiquement (texte brut et ressources RDF associées, identifiées par URIs dans le graphe) dans une même interface. Le format de fichier employé pour la diffusion des annotations IIIF est le JSON-LD. Faisant partie des standards des Linked Open Data, il s'avère particulièrement adapté pour établir un lien entre la localisation des zones dans les images et les annotations structurées sous forme de ressources RDF générées par cette chaîne de traitement.

6 Conclusion

Cet article propose un passage en revue des tâches et principales approches d'extraction d'informations dans des tables historiques et d'interprétation sémantique de tables. Ces deux disciplines alliées ensemble offrent de nombreuses perspectives pour extraire, structurer et rendre accessible les connaissances contenues dans les tables de documents historiques numérisés en masse par les services d'archives ces dernières années. Après avoir identifié les limites des méthodes existantes, nous décrivons les étapes d'une chaîne de traitement spécialisée pour extraire et lier les données extraites des documents tabulaires historiques. Ces données sont structurées suivant une ontologie produite par des experts du domaine et des documents considérés. Le peuplement de l'ontologie avec les informations extraites des images facilite leur interrogation et leur mise en relation. Nous proposons également de développer des outils de visualisation et de requêtes du graphe de connaissances qui maintienne le lien aux sources numérisées et permette de déceler des erreurs. Le standard IIIF nous semble particulièrement pertinent pour mettre en œuvre cet objectif.
Cette approche ouvre de nouvelles perspectives pour la valorisation et l'exploitation de documents tabulaires historiques. Les travaux futurs viseront à évaluer cette chaîne de traitement sur un cas concret : les registres du cadastre napoléonien.

Remerciements

Ce travail a été soutenu financièrement par le Ministère des Armées – Agence de l'innovation de défense.

Références

[1] N. Abdelmageed and S. Schindler. JenTab : A Toolkit for Semantic Table Annotations. In *Proc. 2nd Int. Workshop on Knowledge Graph Construction*, volume 2873 of *CEUR Workshop Proceedings*, 2021.

[2] A. Arora, E. Silcock, M. Dell, and L. Heldring. Contrastive Entity Coreference and Disambiguation for Historical Texts. In *Proc. 2024 Conf. on Empirical Methods in Natural Language Processing*, pages 6174–6186, 2024.

[3] 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.

[4] Y. Chabot, T. Labbe, J. Liu, and R. Troncy. DAGoBAH : An End-to-End Context-Free Tabular Data Se-

mantic Annotation System. In *SemTab@ISWC*, pages 41–48, 2019.

[5] S. Chen, A. Karaoglu, C. Negreanu, T. Ma, J.-G. Yao, J. Williams, A. Gordon, and C.-Y. Lin. LinkingPark : An Integrated Approach for Semantic Table Interpretation. In *SemTab@ ISWC*, pages 65–74, 2020.

[6] IIIF Consortium. IIIF : How It Works. https://iiif.io/get-started/how-iiif-works/. Accessed : 2025-02-27.

[7] 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.

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

[9] Thomas Constum. *Extraction d'information dans des documents historiques à l'aide de grands modèles multimodaux*. Phd thesis, Normandie Université, France, 2024.

[10] D. Coquenet, C. Chatelain, and T. Paquet. End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1) :508–524, 2022.

[11] D. Coquenet, C. Chatelain, and T. Paquet. DAN : A Segmentation-Free Document Attention Network for Handwritten Document Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7) :8227–8243, 2023.

[12] M. Cremaschi, A. Rula, A. Siano, and F. De Paoli. Semantic Table Interpretation Using MantisTable. In *Proc. 14th Int. Workshop on Ontology Matching*, volume 2536 of *CEUR Workshop Proceedings*, pages 195–196, 2019.

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

[14] C. Díaz, J. Dunstan, L. Etcheverry, A. Fonck, A. Grez, D. Mery, J. L. Reutter, and H. R. Corral. Automatic Knowledge-Graph Creation from Historical Documents : The Chilean Dictatorship as a Case Study. In *Joint Proc. 2nd Workshop on KBC from PTLM (KBC-LM 2024) and 3rd Challenge on LM for KBC (LM-KBC 2024)*, volume 3853 of *CEUR Workshop Proceedings*, 2024.

[15] O. Eberle, J. Büttner, H. El-Hajj, G. Montavon, K.-R. Müller, and M. Valleriani. Historical insights at scale : A corpus-wide machine learning analysis of early modern astronomic tables. *Science Advances*, 10(43), 2024.

[16] B. Ell, S. Hakimov, F. Kaupmann, P. Braukmann, L. Cazzoli, A. Mancino, J. A. Memon, K. Rother, A. Saini, and P. Cimiano. Towards a Large Corpus of Richly Annotated Web Tables for Knowledge Base Population. https://pub.uni-bielefeld.de/record/2912802, 2017. dataset.

[17] S. Gottschalk and E. Demidova. Tab2KG : Semantic table interpretation with lightweight semantic profiles. *Semantic Web*, 13(3) :571–597, 2022.

[18] E. Granell, V. Romero, J. R. Prieto, J. Andrés, L. Quirós, J. A. Sánchez, and E. Vidal. Processing a large collection of historical tabular images. *Pattern Recognit. Lett.*, 170, 2023.

[19] W3C Web Annotation Working Group. Web Annotation Data Model. https://www.w3.org/TR/annotation-model/, February 2017. W3.org.

[20] T. Guo, D. Shen, T. Nie, and Y. Kou. Web Table Column Type Detection Using Deep Learning and Probability Graph Model. In *Web Inf. Syst. and Appl.*, pages 401–414, 2020.

[21] N. Jain, A. Sierra-Múnera, M. Lomaeva, J. Streit, S. Thormeyer, P. Schmidt, and R. Krestel. Generating Domain-Specific Knowledge Graphs : Challenges with Open Information Extraction. In *Proc. 1st Int. Workshop on Knowledge Graph Generation From (Text2KG 2022)*, volume 3184 of *CEUR Workshop Proceedings*, pages 52–69, 2022.

[22] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. Ocr-free document understanding transformer. In *ECCV*, pages 498–517.

[23] G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and searching web tables using entities, types and relationships. *Proceedings of the VLDB Endowment*, 3(1-2), 2010.

[24] 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.

[25] P. Mendes, M. Jakob, and C. Bizer. DBpedia : A Multilingual Cross-domain Knowledge Base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1813–1817, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[26] C. Möller. Knowledge Graph Population with out-of-KG Entities. In *The Semantic Web : ESWC 2022 Satellite Events.*, volume 13384, pages 199–214, 2022.

[27] V. Nundloll, R. Smail, C. Stevens, and G. Blair. 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.

[28] R. Petitpierre, M. Kramer, and L. Rappo. An end-to-end pipeline for historical censuses processing. *IJDAR*, 26(4) :419–432, 2023.

[29] J. Puigcerver. Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition ? In *ICDAR*, pages 67–72, 2017.

[30] L. Quirós. P2pala : Page to page layout analysis tookit. https://github.com/lquirosd/P2PaLA, 2017.

[31] S. K. Ramnandan, A. Mittal, C. A. Knoblock, and P. Szekely. Assigning Semantic Labels to Data Sources. In *ESWC*, pages 403–417, 2015.

[32] D. Ritze and C. Bizer. Matching Web Tables To DBpedia - A Feature Utility Study. In *Proc. 20th Int. Conf. on Extending Database Technology (EDBT)*, 2017.

[33] A. Scius-Bertrand, A. Fakhari, L. Vögtlin, D. Ribeiro Cabral, and A. Fischer. Are layout analysis and OCR still useful for document information extraction using foundation models ? In *ICDAR*, pages 175–191, 2024.

[34] F. M. Suchanek, M. Alam, T. Bonald, L. Chen, P-H. Paris, and J. Soria. YAGO 4.5 : A Large and Clean Knowledge Base with a Rich Taxonomy. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 131–140, Washington DC USA, July 2024. ACM.

[35] S. Tarride, M. Boillet, and C. Kermorvant. Key-Value Information Extraction from Full Handwritten Pages. In *ICDAR*, pages 185–204, 2023.

[36] S. Tual, N. Abadie, B. Duménieu, J. Chazalon, and E. Carlinet. Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du 19 ème siècle : application aux métiers de la photographie. In *IC*, 2023.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *arXiv :1706.03762*, 2017.

[38] D. Vrandečić and M. Krötzsch. Wikidata : a free collaborative knowledgebase. *Commun. ACM*, 57(10) :78–85, September 2014.

[39] H. Wei, C. Liu, J. Chen, J. Wang, L. Kong, Y. Xu, Z. Ge, L. Zhao, J. Sun, Y. Peng, C. Han, and X. Zhang. General OCR theory : Towards OCR-2.0 via a unified end-to-end model.

[40] T. Zhang, X. Yue, Y. Li, and H. Sun. TableLlama : Towards Open Large Generalist Models for Tables. In *NAACL*, pages 6024–6044, 2024.

[41] Z. Zhang. Effective and efficient Semantic Table Interpretation using TableMiner+. *Semantic Web*, 8(6) :921–957, 2017.