Rapport IA et Application 2

Romain Guenneau,
Abdellah Hassani,
Simon,
Ritchy Agnesa

20 avril 2025

Contents

1 Introduction

Les actes de mariage anciens offrent des informations clés sur la société française du début du XXe siècle, notamment en matière d'endogamie sociale et de mobilité intergénérationnelle. Bien que disponibles en ligne sous forme d'images, ces documents restent difficiles à exploiter automatiquement. Ce projet vise à tester la faisabilité de l'extraction d'informations à partir de ces actes grâce à la reconnaissance optique de caractères (OCR), à structurer les données extraites dans une base, et à en évaluer la qualité. L'étude portera principalement sur les actes de 1913 et 1920 issus des Archives de Paris.

2 Extraction des données

Nous avons commencé par télécharger manuellement les actes de mariage disponibles sur le site des Archives de Paris, en ciblant les années 1913 et 1920. Pour chaque année, un total de 31 images a été collecté, constituant ainsi un échantillon de test représentatif pour l'évaluation des méthodes de reconnaissance optique de caractères (OCR).

3 Méthodes de reconnaissance OCR

Dans un précédent projet, nous avions expérimenté des solutions open source telles que `TesseractOCR` et `EasyOCR`. Bien que simples à implémenter, ces outils se sont révélés peu fiables sur des documents manuscrits : les lettres et chiffres étaient fréquemment mal interprétés, nécessitant une correction manuelle fastidieuse, ce qui annulait l'intérêt de l'automatisation. Nous avions tentés différents pré-traitements pour essayer de les améliorer, sans réel succès.

Face à ces limites, nous nous sommes tournés vers des modèles de langage (LLM), en particulier GPT, qui avaient donné d'excellents résultats en OCR dans nos précédents essais. Les dernières versions de ces modèles permettent non seulement une retranscription fidèle du texte manuscrit, mais aussi une extraction structurée des informations via un prompt bien conçu.

Dans ce projet, nous avons utilisé l'API de `Gemini`, le LLM développé par Google. Bien que sa mise à disposition gratuite soit appréciable, elle impose une limite stricte de 51 requêtes par clé API, ce qui a restreint notre capacité à traiter un grand volume de données.

L'un des principaux avantages de cette approche est sa polyvalence : un prompt bien formulé permet d'obtenir à la fois la transcription complète de l'acte et les données extraites dans un format structuré. Nous avons ensuite automatisé la conversion des résultats en fichiers `JSON`, puis en `CSV`, afin de constituer progressivement une base de données exploitable.

L'exécution du processus sur les 31 actes de 1913 requiert environ 10 minutes, un délai non négligeable à l'échelle de grands corpus. Toutefois, la précision des résultats obtenus est remarquable, avec des transcriptions quasi parfaites, comme l'illustre l' exemple ci-après.

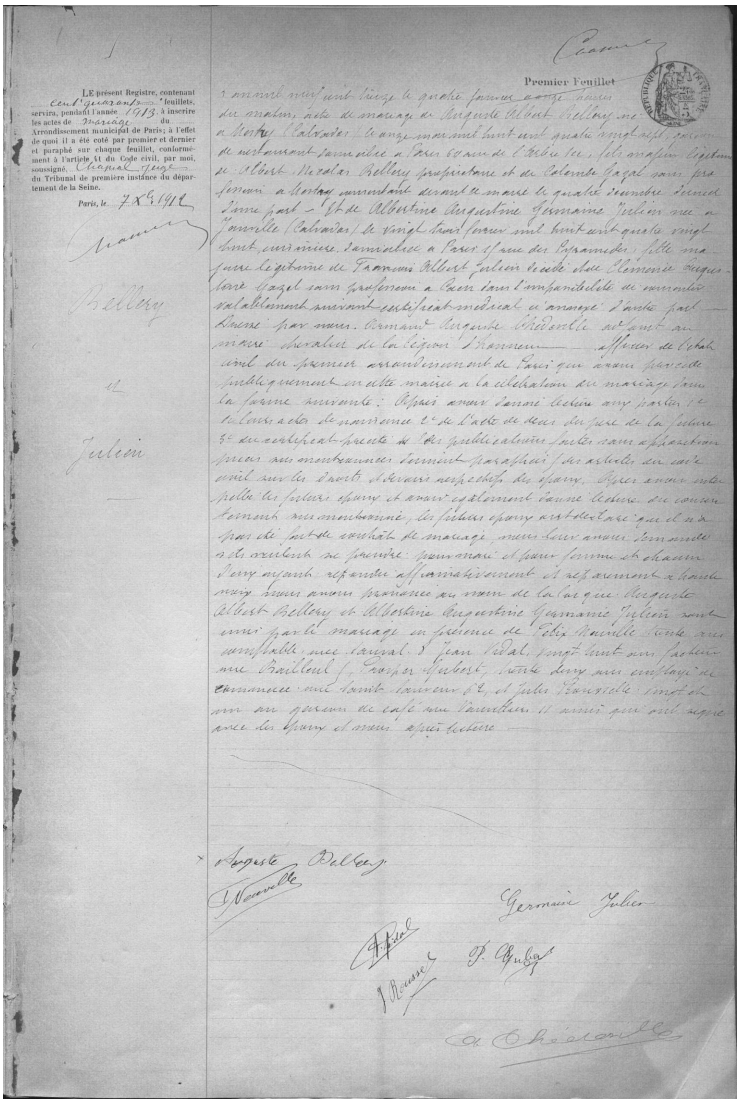

Figure 1: Archive de 1913

Retranscription

LE présent Registre, contenant cent quarante feuillets,
servira, pendant l'année 1913, à inscrire
les actes de mariage de l'Arrondissement municipal de Paris; à l'effet
de quoi il a été coté par premier et dernier
et paraphé sur chaque feuillet, conformément
à l'article 41 du Code civil, par moi,
soussignée, Chapal, Juge
du Tribunal de première instance du département de la Seine.
Paris, le 7/X/1912

Premier Feuillet

L'an mil neuf cent treize le quatre février à onze heures
du matin, acte de mariage de Auguste Albert Bellery né
à Mestres (Calvados) le onze mai mil huit cent quatre vingt sept, garçon
de restaurant demeurant à Paris 60 rue de l'arbre sec, fils majeur
de Albert Nicolas Bellery propriétaire et de Céleste Marguerite Rose
Gervais son épouse demeurant ensemble à Mestres (Calvados) d'une
et d'autre part de Albertine Augustine Germaine Julien née
à Janville (Calvados) le vingt trois février mil huit cent quatre vingt
quatre, cuisinière, domiciliée à Paris 11 rue des Pyramides, fille majeure
de Francis Albert Julien décédé et de Clémence Augustine
Louise Gazel sans profession à Caen dans l'impossibilité de se marier
valablement suivant certificat médical annexé d'acte fait
et signé par nous Armand Angèle Chicdelle, adjoint au
maire, chevalier de la légion d'honneur, officier de l'instruction
publique du premier arrondissement de Paris, qui avons procédé
publiquement en cette mairie à la célébration du mariage dans
la forme suivante. Après avoir donné lecture aux parties et
à leurs à chacun de nouveaux et de l'acte de deux des parents de la future
épouse, du chapitre pré-cité des publications faites sans opposition,
et des autres documents nécessaires, nous leur avons exposé en
leur présence les dispositions relatives aux droits et devoirs respectifs
des époux. Après avoir ensuite, reçu, de chacun des futurs époux et du
père de l'époux, leurs consentements respectifs. Nous avons,
en leur présence, fait également donner lecture du contrat
de mariage devant le maire et les témoins soussignés, puis nous les avons
requis s'ils veulent se prendre pour mari et femme, chacun d'eux
y ayant répondu affirmativement et séparément de sa main
droite, nous avons prononcé au nom de la loi que Auguste
Albert Bellery et Albertine Augustine Germaine Julien sont
unis par le mariage en présence de Félix Mérille (texte raturé :
comptable), rue Censival 8, et Jean Nidal, rue des

Milleuls, Grosser Gubert, et Jules Rousselle témoins domiciliés à Paris et majeurs, qui ont signé avec nous en présence du greffier du cabinet. Il est noté que l'acte est lu aux époux, témoins.

Informations extraites


```
{
"date_mariage": "quatre février 1913",
"lieu_mariage": "Mairie du premier arrondissement de Paris",
"nom_prenom_marie": "Auguste Albert Bellery",
"profession_marie": "garçon de restaurant",
"adresse_marie": "Paris 60 rue de l'arbre sec",
"nom_prenom_mariee": "Albertine Augustine Germaine Julien",
"profession_mariee": "cuisinière",
"adresse_mariee": "Paris 11 rue des Pyramides",
"nom_pere_marie": "Albert Nicolas Bellery",
"profession_pere_marie": "propriétaire",
"nom_mere_marie": "Céleste Marguerite Rose Gervais",
"profession_mere_marie": "épouse",
"nom_pere_mariee": "Francis Albert Julien",
"profession_pere_mariee": "décédé",
"nom_mere_mariee": "Clémence Augustine Louise Gazel",
"profession_mere_mariee": "sans profession"
}
```


4 Évaluation

4.1 Évaluation de l'OCR

La première étape consiste à évaluer la qualité de la transcription du texte à partir des images manuscrites. Pour cela, le texte extrait automatiquement par notre modèle pourrait être comparé à une transcription manuelle réalisée sur un échantillon d'actes de mariage.

Les métriques les plus appropriées dans ce contexte seraient :

- **CER (Character Error Rate)** : mesure le pourcentage d'erreurs au niveau des caractères (insertions, suppressions, substitutions)
- **Distance de Levenshtein** : mesure le nombre minimal d'opérations nécessaires pour transformer la transcription en texte de référence.

Cependant, la constitution d'un corpus de référence manuel s'avère difficile. En effet, les actes anciens sont souvent peu lisibles, ce qui rend la transcription humaine fastidieuse et très longue. Malgré cela, les premières observations qualitatives montrent que le modèle OCR utilisé fournit des résultats globalement satisfaisants, même sur des écritures anciennes.

4.2 Évaluation de l'extraction d'information

La seconde étape concerne l'extraction des informations structurées à partir du texte transcrit. Cette tâche revient à identifier et à extraire automatiquement les entités pertinentes (noms, dates, professions, lieux, etc.) dans un format exploitable.

Pour évaluer cette étape, un échantillon de documents peut être annoté manuellement avec les bonnes valeurs attendues pour chaque champ, puis comparé aux résultats extraits automatiquement. Cette comparaison peut être assimilée à un problème de classification d'entités nommées, et s'évalue à l'aide de métriques classiques comme la précision et le rappel. Les mêmes difficultés d'annotations précédemment citées se posent toujours c'est pourquoi nous avons choisi de ne pas le faire pour ce projet.

5 Conclusion

Ce projet a permis de démontrer la faisabilité de l'extraction d'informations structurées à partir d'actes de mariage anciens, en combinant des techniques de reconnaissance optique de caractères et l'utilisation de modèles de langage avancés. L'intégration des LLM comme Gemini s'est révélée particulièrement efficace, tant pour la transcription fidèle des documents que pour l'extraction automatique des informations pertinentes.

Les résultats obtenus sur l'échantillon de 31 actes de l'année 1913 sont très encourageants, avec une qualité de restitution proche de celle d'une annotation humaine. Toutefois, des limitations subsistent, notamment en ce qui concerne le coût en temps et les restrictions imposées par les API utilisées, qui freinent le traitement de volumes plus importants.
À terme, une généralisation de cette méthode à un plus grand corpus d'actes numérisés permettrait de constituer une base de données inédite, ouvrant la voie à des recherches approfondies en sciences sociales sur les dynamiques familiales, sociales et professionnelles du début du XX^{e} siècle. Le travail engagé dans ce projet constitue ainsi une première étape vers une exploitation systématique des archives d'état civil au service de la recherche.