Un pipeline pour connecter cartes anciennes et bases de connaissances historiques

Projet Atlas Historique Nouvelle-Aquitaine

Jean Pylouster – Ingénieur d'études CNRS (UMR CeRCA Poitiers)

En collaboration avec :
Mathieu Chartier – Doctorant (XLIM Poitiers)
Guillaume Bourgeois – Maître de conférence (CRIHAM-Poitiers)
Christine Plumejeaud – Ingénieure de recherche (UMR MIGRINTER – Poitiers)

Atlas historique de la Nouvelle-Aquitaine

- **L'Atlas historique de la Nouvelle-Aquitaine** assemble et met à disposition l'ensemble des connaissances historiques et cartographiques concernant cette région, de la protohistoire jusqu'à nos jours.

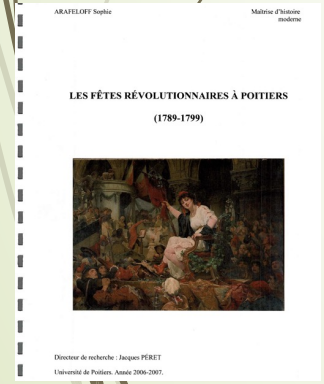
- Ce projet en humanités numériques s'appuie sur un **consortium de laboratoires de recherche ressortant de quatre universités régionales**, en association avec les grands acteurs de la conservation documentaire et patrimoniale, archéologique et historique en Nouvelle-Aquitaine.

- https://atlas-historique-nouvelle-aquitaine.huma-num.fr/

Traitement des contenus

Plusieurs étapes clés dans la chaîne de traitement :

1. Numérisation des contenus (scans, traitements de texte…)
2. Traitement avec **ABBYY** (OCR) et **Transkribus** (HTR)
3. Conversion en PDF si nécessaire (permet l'indexation par page)
4. Extraction des contenus et métadonnées vers des fichiers JSON (Python)
5. Nettoyage des textes « bruités » (Python + manuel)


ARAFELOFF Sophie

Maîtrise d'histoire moderne

LES FÊTES RÉVOLUTIONNAIRES À POITIERS

(1789-1799)

Directeur de recherche : Jacques PÉRET

Université de Poitiers. Année 2006-2007.


```
{
    "content": {
        "pages": {
            "1": "ARAFELOFF Sophie \n Maîtrise d'histoire moderne \n LES FETES
                REVOLUTIONNAIRES A POITIERS \n (1789-1799) \n Directeur de
                recherche : Jacques PÉRET \n Université de Poitiers. Année
                2006-2007.",
```

Analyse de l'OCR

Après traitement du fureteur, on obtient un fichier JSON avec un élément par page (résultat) de document.


```
[
    {
        "page_num": 20,
        "page_content": "Ce sont les communes de Chail (1259 ha), ! 'Enclave de la Martinière (1141 ha), Maisonnais (516 ha), Mazières-sur-Béronne (94
        "page_content_no_punct": "Ce sont les communes de Chail 1259 ha 'Enclave de la Martinière 1141 ha Maisonnais 516 ha Mazières-sur-Béronne 949 h
        "filepath": "JSON-cleaned/PyMuPDF\\Histoire-contemporaine\\Deux-Sevres",
        "file": "Allain-Nathalie-Lactivite-du-juge-de-paix-a-Melle.json"
    },
    {
        "page_num": 41,
        "page_content": "42 tolérance se multiplient. Beaucoup voient dans cette récente effervescence un moyen d'augmenter les gains. Pourtant, la mu
        "page_content_no_punct": "42 tolérance se multiplient Beaucoup voient dans cette récente effervescence un moyen d'augmenter les gains Pourtant
        "filepath": "JSON-cleaned/PyMuPDF\\Histoire-contemporaine\\Charente-Maritime",
        "file": "Alles-Christelle-Les-filles-publiques-a-La-Rochelle.json"
    },
    {
        "page_num": 72,
        "page_content": "Nombre de testateurs par paroisses \n (1766, 1807)' \n (liasses 3 13 474 a 3 13 1016 I \n Paroisses \n Saint - Barthélémy \n
        "page_content_no_punct": "Nombre de testateurs par paroisses 1766 1807 ' liasses 3 13 474 a 3 13 1016 I Paroisses Saint - Barthélémy Notre - I
        "filepath": "JSON-cleaned/PyMuPDF\\Histoire-moderne\\Charente-Maritime",
        "file": "Alves-de-Souza-Stephanie-Dechristianisation-lacisation-etude-des-comportements-devant-la-mort-a-La-Rochelle.json"
    },
    {
        "page_num": 83,
        "page_content": "1 \n \n V \n N Préc \n Saint - Barthélémy \n 17 \n 17 \n 5 \n 18 \n 6 \n Notre - Dame \n 14 \n 4 \n 3 \n 5 \n 8 \n Sans préci
        "page_content_no_punct": "1 V N Préc Saint - Barthélémy 17 17 5 18 6 Notre - Dame 14 4 3 5 8 Sans précisions 8 5 0 3 9 Saint - Jean 9 3 0 4 2
        "filepath": "JSON-cleaned/PyMuPDF\\Histoire-moderne\\Charente-Maritime",
        "file": "Alves-de-Souza-Stephanie-Dechristianisation-lacisation-etude-des-comportements-devant-la-mort-a-La-Rochelle.json"
    },
    {
        "page_num": 106,
        "page_content": "103 \n Les entreprises à capitaux britanniques en Poitou Charentes \n Sociétés \n Activités \n Localisation \n Effectif \n Or
        "page_content_no_punct": "103 Les entreprises à capitaux britanniques en Poitou Charentes Sociétés Activités Localisation Effectif Origine Sal
        "filepath": "JSON-cleaned/PyMuPDF\\Geographie",
        "file": "ANDRIEUX-France-1994-Les-Britanniques-en-France-une-immigration-fine.json"
    },
```

Analyse NLP

- C'est une étape primordiale pour la suite du workflow.
- On utilise un analyseur NLP pour traiter ce fichier : **Spacy** avec le corpus, **CamemBERT** pour le traitement du français.
- Actuellement, on ne s'intéresse qu'aux noms de lieux (commune, rue…)
- Nécessite des règles fines d'analyse NLP pour extraire les entités nommés des textes
- Une fois traité, nous obtenons un fichier au format CSV

```
lieu;type;def;page;ref
Charente;;nmod;2;Tricaud-Cedric-Henri-Thebault-homme-politique-charentais-atypique.json
Angoulême;;nmod;2;Tricaud-Cedric-Henri-Thebault-homme-politique-charentais-atypique.json
Saint-Cybard;;nmod;2;Tricaud-Cedric-Henri-Thebault-homme-politique-charentais-atypique.json
```

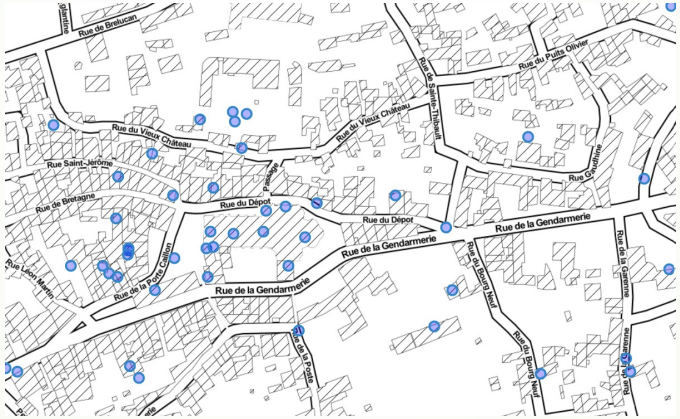
Géolocalisation

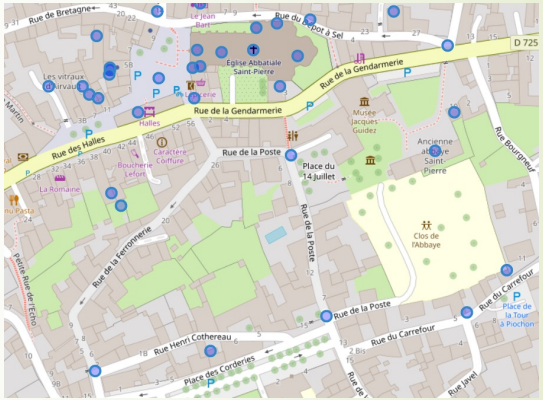
- Utilisation du géocodage d'OpenStreetMap
- Pour chaque lieu contenu dans le fichier CSV, on recherche les coordonnées GPS et on sauvegarde aussi la référence du mémoire d'origine
- On génère un fichier de sortie au format GeoJSON


```
{
    "type": "FeatureCollection",
    "properties": {"layer": "couche1"},
    "features": [
        {
        "type": "Feature",
        "geometry": {
            "type": "Point",
            "coordinates": ["46.82744", "-0.13786"],
            "properties": {
                "commune": "Château d'Airvault",
                "page": "5",
                "reference": "Chartier Mathieu - Topographie et développement morphologique d'Airvault et de Saint-Loup-sur-Thouet au Moyen Âge - Volume 1.pdf"
                }
            }
        },
```

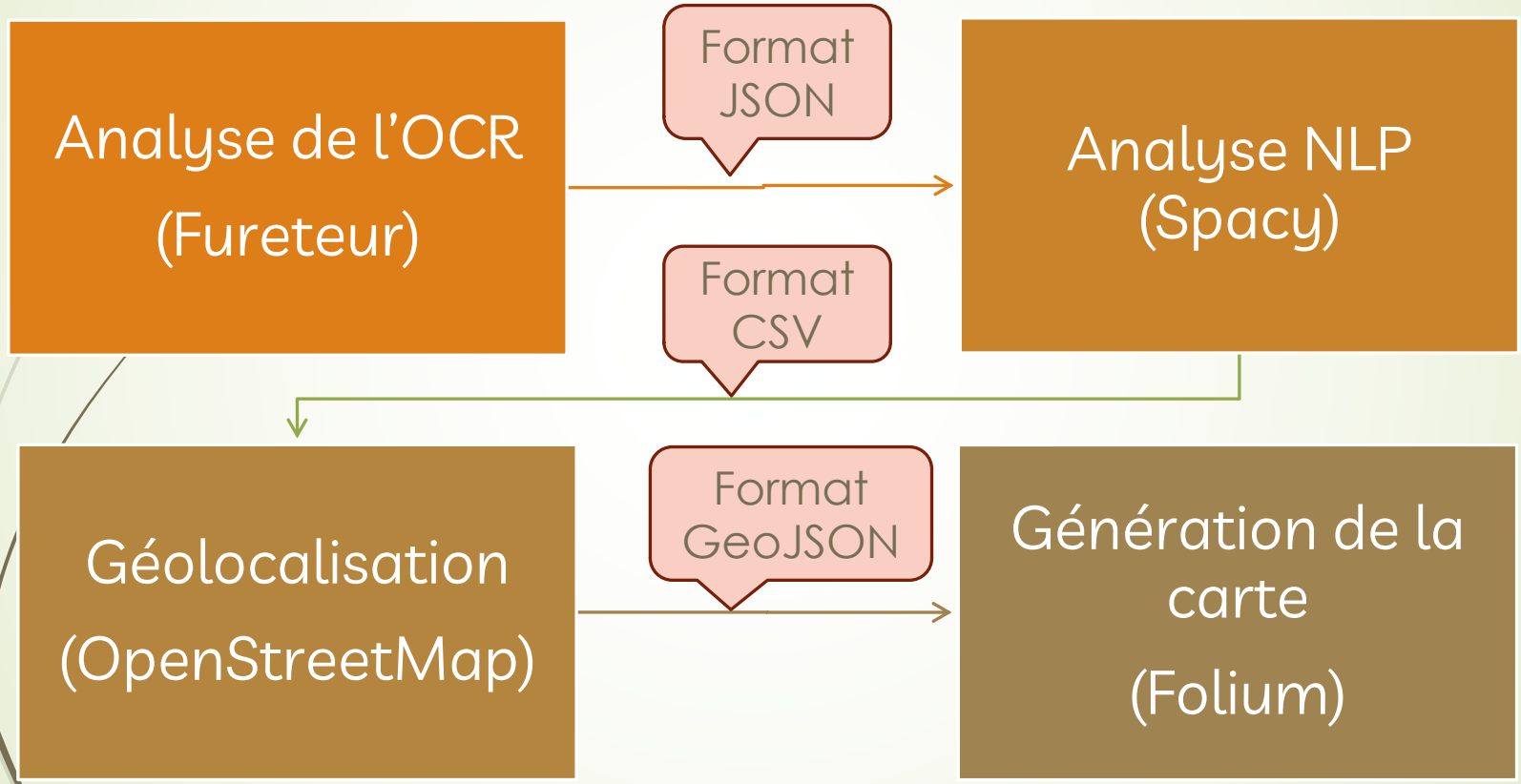
Génération de la carte

- Utilisation de l'API **Folium** pour la génération de la carte au format HTML avec le jeu de tuiles par défaut **OpenStreetMap**.
- Folium permet facilement d'utiliser plusieurs jeux de tuiles





Workflow (par NextFlow)



Pourquoi utilisé un pipeline ?

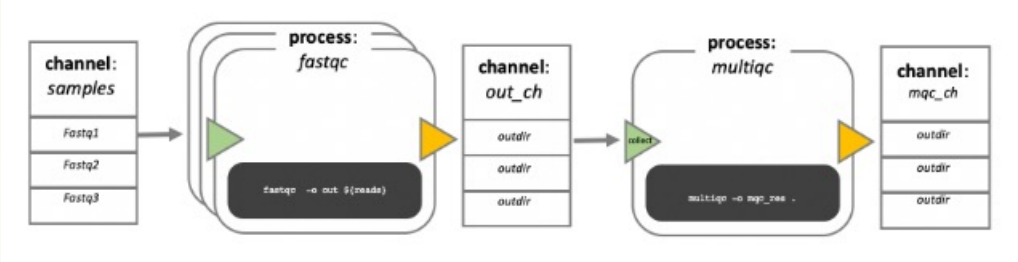
- Nécessité d'enchainer des tâches de collecte, de nettoyage et de traitements des données
- Gérer des données de provenance différente et des données volumineuses.
- Nécessiter de gérer des fonctions clefs :
  - **Gestion du temps d'exécution :** gestion de l'exécution du programme et fractionnement des tâches et des données pour qu'elles s'exécutent en même temps dans un processus appelé parallélisation
  - **Reproductibilité :** la spécification du pipeline signifie que le flux de travail produira les mêmes résultats lors de sa réexécution
  - **Réentrée :** reprise à partir des dernières étapes exécutées avec succès (point de contrôle)

Pourquoi Nextflow ?

- Pipeline utilisé en bioinformatique (et aussi en physique, en imagerie)
- Capable de gérer des données volumineuses
- Capable de fractionner des tâches pouvant être exécutées en même temps (parallélisation).
- Gestion de point de contrôle

Comment ça marche ?

Un workflow : processus, canaux et flux de travail.



Un peu de code … ATLAS Nextflow

Une démonstration … Reporting Nextflow

Suite du projet …

- Interface graphique pour la partie Workflow
- Améliorer le nettoyage post-OCR et post-HTR pour faciliter les extractions de données (entités nommées, triplets RDF…)

  → Exemple : automatiser des corrections en NLP
- Améliorer l'extraction de lieux pour le géoréférencement

  → Exemples : mieux relier un nom de rue ou un monument à une ville, gérer le surfacique et le parcellaire…
- Créer une base de connaissances autour du corpus (textes + cartes)

  → Exemple : faciliter les interactions entre les contenus et les cartes

Merci pour votre attention !