# Le corpus des Ouvriers des deux mondes : des images et des URLs

Jean-Damien Généro

# Le corpus des *Ouvriers des deux mondes* : des images et des URLs

Jean-Damien Généro*

19 juin 2020

Si les documents d'archives ont une part prépondérante dans le projet Time us, ils ne représentent pas pour autant l'intégralité de sa documentation. Les imprimés sont également présents, sous la forme de trois importants dossiers : la collection de la presse ancienne lyonnaise, divers imprimés portant sur le textile en France au XIXe siècle, et le corpus des *Ouvriers des deux mondes*[1].

Les *Ouvriers des deux mondes* sont des enquêtes sociologiques réparties en 3 séries et 126 monographies[2]. Initiée par le sociologue Frédéric Le Play (1806-1882), la publication est assurée par la Société internationale des études pratiques d'économie sociale de 1857 à 1928 et représente un total de 13 volumes. Ceux-ci sont aujourd'hui intégralement consultables sur le site *Internet Archive*[3]. Nous allons nous intéresser dans ce billet aux fichiers de transcription de ces volumes et au lien entre ceux-ci et les images numérisées d'origine.

Le script LSE OD2M, écrit par Alix Chagué, avait automatiquement segmenté et transcrit les images, puis encodé et structuré en XML-TEI les textes bruts ainsi obtenus[4] ; la sortie avait résulté en 13 fichiers XML. Ces fichiers « sources » avaient ensuite été scindés en 222 fichiers XML correspondant à autant de divisions logiques des volumes : les monographies bien sûr, mais également les introductions, tables des matières et autres éléments de paratexte. Des opérations de vérification ont permis de réduire le nombre de fichiers à 192.

* *Ingénieur de recherche et développement stagiaire, Inria, équipe ALMAnaCH ; étudiant du Master "Technologies numériques appliquées à l'histoire" de l'École nationale des chartes. Lien vers le billet original.*

1. Aperçu des états sur le wiki Time Us.

2. Anthony, Lorry, « Les monographies des *Ouvriers européens* (1855, 1877-1879) et des *Ouvriers des deux mondes* (1857-1930). Inventaire et classification », dans *Les Études sociales. Les monographies de famille de l'École de Le Play*, nos 131-132, 2000, pp. 93-181, spec. p. 101.

3. Liens des différents volumes sur le wiki Time Us.

4. Voir son biller à propos de la Constitution d'un corpus textuel sur les monographies de Le Play.

# 1 Les images : stockage local ou distant ?

Le schéma d'encodage retenu conserve le lien entre l'image et sa transcription. Celui-ci s'exprime sous la forme d'un élément `<facsimile>` englobant un ensemble de balises `<graphic>`, dont l'attribut `@url` indique la localisation de l'image.

```
<facsimile xml:id="facs_451">
  <surface lrx="2721" lry="4415">
   <figure/>
   <zone rendition="printspace">
    <zone lrx="1357" lry="712" rendition="paragraph"
        ulx="352" uly="474" xml:id="facs_451_p_1"/>
    <zone lrx="1354" lry="864" rendition="paragraph"
        ulx="350" uly="750" xml:id="facs_451_p_2"/>
    <zone lrx="2430" lry="3952" rendition="paragraph"
        ulx="1419" uly="3772" xml:id="facs_451_p_24"/>
   </zone>
  </surface>
  <graphic url="../images/bin/
      lesouvriersdesde01sociuoft_0454.tif"/>
</facsimile>
```

Le script LSE OD2M a travaillé à partir d'images stockées localement après avoir été téléchargées depuis *Internet Archive* : chaque attribut `@url` contient donc un chemin local vers une image.

Ce stockage local répondait à un besoin spécifique lors de la phase de transcription. Il pose néanmoins problème pour la suite du traitement, dans la mesure où il ne garantit pas la portabilité du corpus.

Le site *Internet Archive* met à disposition du plus grand nombre des ressources digitalisées ou numériquement natives depuis son lancement en 1996. Son idée fondatrice est d'être un centre stable et durable d'archives digitales ; stocker en local les ressources qui en sont issues ne semble ainsi pas nécessaire. Il a donc très vite été question de substituer au chemin local l'url de l'image sur *Internet Archive*.

Pour ce faire, deux étapes étaient nécessaires :

— Rechercher les urls ;

— Écrire un script pour :

1. Itérer sur l'ensemble des fichiers XML ;
2. Comparer les chemins locaux aux urls ;
3. Éffectuer la substitution lorsque les deux correspondaient à la même image.

## 2 Recherche des urls

Éffectuer la substitution sur la base d'une expression régulière n'était pas envisageable, car la dénomination des images dans les urls sur *Internet Archive* et dans les des fichiers images n'était pas similaire.

La piste de la librairie python `internetarchive` a été explorée, mais là encore sans succès. C'est finalement l'analyse du code source des pages d'I*Internet Archive* qui a permis de remonter jusqu'à un fichier JSON contenant les urls des images.

Le format JSON présente une information structurée qui s'apparente à des dictionnaires et des listes pour le langage Python. Celui-ci intègre un module (`json`) permettant de lire les données des fichiers JSON. Dans le cas qui nous intéresse, les url se trouvaient à ce chemin :

```
|----['data']
      |----['brOptions']
            |----['data']
                  |----[index de la double page]
                          |----[index de la page]
                                      |----['uri']
```

Une fois cette information connue, il devenait possible de remplacer dans le fichier XML les chemins locaux par ces urls. Encore fallait-il s'assurer que les deux correspondaient à la même image.

## 3 Le script de substitution

Le script de comparaison des chemins et des urls (fig. 1) requiert trois arguments :

— Un fichier `.csv` contenant une liste des liens vers les fichiers JSON ;
— Un deuxième fichier `.csv` avec une liste de fichiers `.xml` ;
— Le chemin local menant au dossier de ces fichiers.

Dans un premier temps, le script itère sur les urls des JSON et sur les dénominations des fichiers XML. Dans un second temps et à condition que l'identifiant du JSON se trouve dans l'intitulé du fichier XML — et donc si les deux ont bien pour objet le même volume — une requête GET est effectuée vers le JSON via le module `requests`[5] et le fichier XML est ouvert et parsé via la librairie `Beautiful soup`[6].

5. Méthode du protocole HTTP permettant de demander une ressource stockée sur un serveur(documentation).

6. Parser un fichier consiste à le lire et à interpréter son contenu afin d'en extraire certains éléments.

Ensuite, une nouvelle itération est effectuée sur chaque double-page dans le JSON et sur chaque balise `<graphic>` dans le fichier XML. Les valeurs obtenues (url de la page de droite, url de la page de gauche, chemin local de l'attribut `@url` de `<graphic>`) sont stockées et comparées : lorsqu'une correspondance est trouvée, `@url` prend pour nouvelle valeur l'adresse de l'image sur *Internet Archive*.

La comparaison finale est effectuée grâce à une expression régulière fondée sur l'identifiant unique à quatre digits de l'image, précédé par un tiret bas (`_\d{4}`), présent tant dans le chemin local que dans l'url de l'image.

## Conclusion

Le script a effectué 6503 insertions dans 192 fichiers en un peu moins d'une minute.

Il a été écrit pour répondre à un besoin posé par les fichiers XML d'un corpus spécifique ; sa réutilisation est néanmoins envisageable si l'utilisateur peut satisfaire aux arguments du script. Sur les trois requis, les deux premiers sont essentiels pour son bon fonctionnement (deux tableaux listant les JSON et les fichiers XML qui seront comparés). Le troisième, le chemin absolu vers le dossier contenant les XML, peut en revanche être rendu optionnel et converti en une variable intégrée dans l'exécution du script.

En dernier lieu, il faut noter que les urls listées dans les JSON présentent l'avantage de pointer directement vers les images sources, et non vers l'interface de consultation d'Internet Archive. C'est une garantie de pérennité (aucune dépendance vis à vis des mises à jour de la visionneuse). Il devient possible d'utiliser ces images dans une édition en ligne des *Ouvriers des deux mondes*, tout en évitant le coût d'hébergement des images.

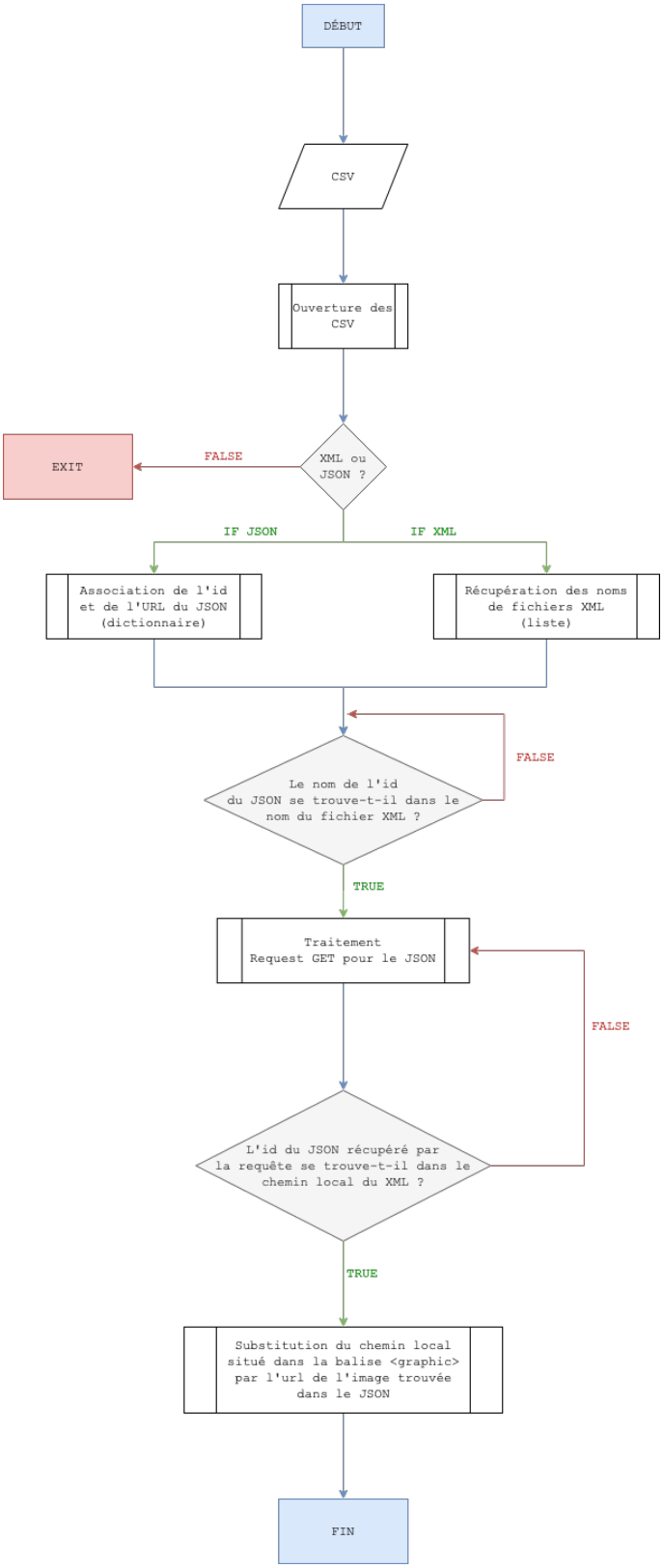


FIGURE 1 – Algorigramme du script