Le saviez-vous ? Les répertoires de notaires ne sont pas seulement des images numérisées !

Lucas Terriel

Le saviez-vous ? Les répertoires de notaires ne sont pas seulement des images numérisées !

Lucas Terriel*

6 octobre 2020

Résumé

Ce billet de blog brosse un panorama des données associées aux documents du projet coordonnée par INRIA (équipe-projet ALMAnaCH) et les Archives nationales LectAuRep - Lecture automatique de répertoires - qui consiste à appliquer les techniques de reconnaissance optique des écritures manuscrites aux répertoires des notaires. Ce billet s'inscrit dans le cadre d'une réflexion plus large sur la création d'un format pivot TEI pour centraliser les métadonnées associées aux documents et celles générées durant le traitement des images avec la plateforme de transcription eScriptorium.

Abstract

This post provides an overview of the data associated with the documents of the project coordinated by INRIA (team project ALMAnaCH) and the National Archives LectAuRep - Automatic reading of directories - which consists in applying the handwritten text recognition techniques on notaries directories. This post is part of a larger reflection on the creation of a TEI pivot format to centralize metadata associated with documents and those generated during image processing with the eScriptorium transcription platform.

* *Ingénieur stagiaire recherche et développement, Inria, équipe-projet ALMAnaCH; étudiant du Master "Technologies numériques appliquées à l'histoire " de l'École nationale des chartes.* *Ce billet a initialement publié sur le blog hypotheses.org LectAuRep, il a été adapté le 11/02/2021 pour les besoins de la mise en forme LaTeX.*

En 2020, le projet LectAuRep a franchi un cap. Il marque une concrétisation importante des objectifs de départ et des efforts menés jusqu'à maintenant dans le déploiement et l'utilisation importante de la plateforme « couteau suisse » eScriptorium[1] : la segmentation et la transcription automatique des images de répertoires de notaires devient réalisable dans un environnement autonome.

Cependant, le projet est aussi rentré dans sa phase la plus « expérimentale ». Comme dans de nombreux projets menés en mode agile [2], un des principes phares, de ce concept de management de projets numériques consiste à toujours remettre en question et ouvrir des discussions sur les moyens de devenir plus efficace dans les buts que l'on s'est fixés au départ.

Parmi ces réflexions, l'une d'entre elles s'est ouverte sur la garantie de l'accès aux métadonnées aussi bien pour les usagers que pour les administrateurs du projet. Qu'il s'agisse de la récupération ou d'assurer l'enrichissement de ces métadonnées ; en somme la question n'est pas si simple, d'autant que ces dernières sont très nombreuses !

Une galaxie de métadonnées

Commençons par le matériau brut, je veux parler des métadonnées.

Imaginez que vous disposiez de vos dernières photographies de vacances : vous aimeriez peut-être vous souvenir de la date et du lieu des prises de vues, des personnes ou des monuments qui y sont présents. Peut-être encore souhaiteriez-vous pouvoir les classer en fonction des couleurs dans le cadre d'un projet ? En résumé, vous voudriez disposer des informations ou des références liées à vos images à tout instant. Une définition simple des métadonnées consisterait à dire qu'il s'agit « d'une donnée servant à définir ou décrire une autre donnée quel que soit son support (papier, audio, vidéo ou électronique) ».

Cette définition de la métadonnée comme une « donnée sur la donnée » est pourtant loin d'être suffisante dans le cadre d'un projet numérique comme LectAuRep. En effet, les métadonnées répondent à un besoin d'**intégrité** dans le temps (elles ne doivent pas être altérées), à un besoin de **persistance** (stockage et pérennisation des métadonnées), et répondent éventuellement à des modèles d'**interopérabilité** avec le web sémantique ou des modules de visualisations qui correspondent à des standards comme IIIF[3] (*International Image Interoperability Framework*).

Il faut aussi penser à leurs usages futurs et aux droits d'accès que l'on laisse associer à certaines métadonnées : peut-on garantir le même degré d'accès à l'ensemble des informations entre un chercheur en sciences humaines qui réalise des analyses computationnelles sur les répertoires et les opérateurs des Archives nationales ? Enfin les métadonnées doivent être enrichies au fil du temps selon

[1] Pour en savoir plus sur la plateforme eScriptiorium.

[2] La méthode agile est un ensemble de pratiques pour la gestion de projets numériques qui reposent sur le prototypage rapide et des itérations courtes pour en savoir plus.

[3] Lien vers IIIF.

les traitements effectués dans la plateforme eScriptorium pour conserver une trace des opérations.

Le schéma ci-dessous donne une idée de l'ensemble des métadonnées qui gravitent autour d'un répertoire de notaire et auquel un « super-utilisateur » hypothétique devrait avoir accès dans le cadre du projet LectAuRep.

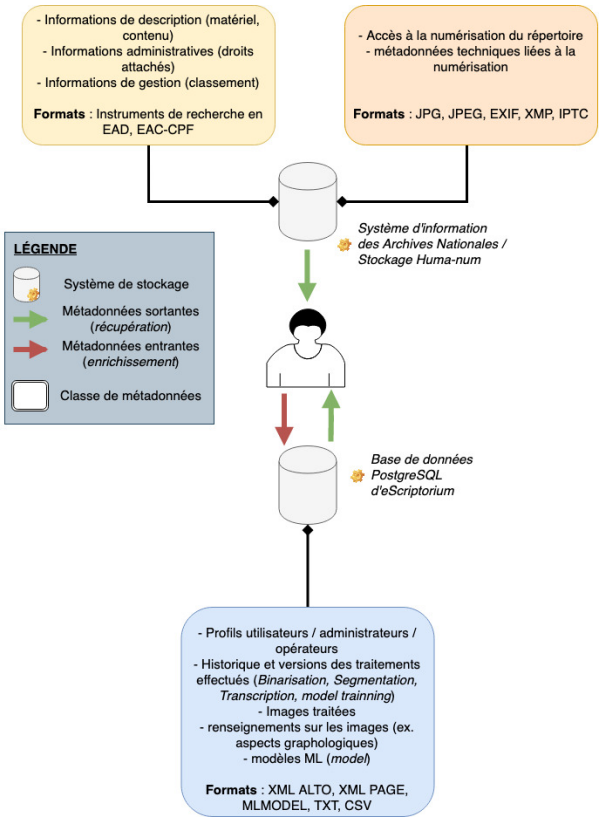

FIGURE 1 – Modélisation des métadonnées associées aux répertoires de notaires dans le cadre du projet LectAuRep

Plusieurs observations sur les métadonnées propres au projet LectAuRep découlent de ce schéma :

1. Il existe des métadonnées « sources » ou inhérentes aux répertoires de notaires (c'est-à-dire celles qui se rattachent à l'archive papier ou à sa copie numérique conservée aux Archives nationales) et des métadonnées « en puissance » (qui seront attachées aux documents pendant ou après leurs utilisations à différents stades de traitements) ;
2. Les métadonnées ne proviennent pas du même silo d'informations (Système d'information des Archives nationales ou base de données SQL [4]) ;

[4]Le langage SQL (*Structured Query Language*) est utilisé pour exploiter les bases de données relationnelles.

3. Les formats de ces métadonnées suivent des schémas très différents, ce qui rend leur interopérabilité difficile car cela suppose de passer par un alignement des différents référentiels ;

4. Toutes les métadonnées ne sont pas forcément consultables, cela dépend des droits attachés à un profil utilisateur.

Des métadonnées liées aux objets physiques

Une première classe [5] de métadonnées correspond aux métadonnées administratives et de gestion (identification des archives, provenance et contexte de production de ces archives, intégrité, droits, etc.) et aux métadonnées descriptives (nom et prénom du notaire, numéro de l'étude, adresse, type de répertoires).

Ce sont généralement des métadonnées qui sont liées à l'objet physique lui-même, c'est-à-dire l'archive papier ou aux copies numériques. Les répertoires de notaires sont associés à des instruments de recherche[6] (abrégé IR) généralement encodés en XML [7] EAD [8] (*Encoded Archival Description*) et en XML EAC-CPF [9] (*Encoded Archival Context - Corporate Bodies, Persons and Families*), rédigés et maintenus par le minutier central des Archives nationales.

On compte deux IR EAD associés à chaque répertoire de notaires : le premier relatif aux images du (ou des) répertoire(s) d'un notaire qui se présente sous la forme d'un inventaire spécifique pour des images numériques du répertoire de tel ou tel notaire et un deuxième relatif aux « Minutes et répertoires d'un notaire » qui constitue un inventaire des actes accompagnés de descriptions minutieuses de chacune des minutes de notaires et des répertoires eux-mêmes. Pour un notaire, nous disposons donc de trois fichiers XML en tout (deux XML EAD et un XML EAC-CPF), contenant des balises XML similaires mais des structurations d'arborescence différentes et parfois des informations redondantes.

[5] Je parle de classes de métadonnées afin de ne pas sous-entendre une hiérarchie par niveaux, ce qui laisserait supposer que certaines métadonnées ont une importance supérieure à d'autres dans le projet.

[6] « Un instrument de recherche est un outil papier ou informatisé énumérant ou décrivant un ensemble de documents d'archives de manière à les faire connaitre aux lecteurs. » Définition extraite de Lydiane Gueit-Montchal (Dir.), *Abrégé d'archivistique. Principes et pratiques du métier d'archiviste 4e édition revue et augmentée*, AAF, 2020

[7] *Extensible Markup Language* (« langage de balisage extensible »). « XML est un langage qui permet d'encoder des informations. Il est largement utilisé pour l'encodage de ressources textuelles de par son système de balise. Un paragraphe peut être noté `<p>Texte</p>` », définition extraite de Thibault Clérice, "« Les outils CapiTainS, l'édition numérique et l'exploitation des textes » ", *Médiévales. Langues, Textes, Histoire*, 73–73 (2017), URL : http://journals.openedition.org/medievales/8211 (visité le 01/06/2020)

[8] « EAD ou *Encoded Archival Description* est un modèle pour la production en XML d'instruments de recherche archivistiques conforme à la norme internationale ISAD(G) », L. Gueit-Montchal (Dir.), *Abrégé d'archivistique. Principes et pratiques du métier d'archiviste 4e édition revue et augmentée...*

[9] « EAC ou *Encoded Archival Context - Corporate Bodies, Persons and Families* est un modèle de production en XML de notices d'autorité de personnes physiques , de familles ou de personnes morales conformes à la norme internationale ISAAR(CPF) », *Ibid.*

Au total le minutier central a recensé plus de 2000 registres de répertoires laissés par le notariat parisien sur la période allant de 1803 à 1944, je laisse donc au lecteur la liberté de mesurer la quantité de fichiers XML associés aux répertoires. Pour avoir un aperçu de ces instruments de recherche vous pouvez vous rendre sur la Salle des inventaires virtuels[10] et télécharger l'inventaire *.xml* pour le visualiser dans un éditeur XML.

Des métadonnées techniques liées aux numérisations

La numérisation des répertoires de notaires est associée à une seconde classe de métadonnées : les métadonnées techniques. Parmi celles-ci les métadonnées EXIF (*Exchangeable image file format*) [11] qui fournissent un ensemble de renseignements sur l'image numérisée : les réglages et la marque de l'appareil, les dimensions de l'image, la date et l'heure de la numérisation, ou encore des informations sur les droits d'auteur. Ces métadonnées peuvent être structurées sous la forme d'un fichier XML en respectant l'espace de nom[12] et ne sont accessibles que par l'intermédiaire d'un logiciel de traitement photographique (XnView, GIMP ou ExifTool) ou en passant par des scripts. Avec le langage Python, on peut utiliser les bibliothèques Pillow ou PyExifTool, par exemple.

Un exemple de métadonnées EXIF extraites à l'aide d'un script utilisant PyExifTool :

```
EXIF:ImageDescription => 74 - Main frame
EXIF:Make => i2S, Corp.
EXIF:Model => SupraScanII [SN: 283910] - Cam7600RGB [SN: 283910]
EXIF:Orientation => 1
EXIF:XResolution => 300
EXIF:YResolution => 300
EXIF:ResolutionUnit => 2
EXIF:ModifyDate => 2015:09:07 09:45:39
EXIF:YCbCrPositioning => 1
EXIF:ExifVersion => 0230
EXIF:CreateDate => 2015:09:07 09:45:39
EXIF:ComponentsConfiguration => 1 2 3 0
EXIF:FlashpixVersion => 0100
EXIF:ColorSpace => 65535
```

[10] Lien vers la SIV (Salle des inventaires virtuels)

[11] EXIF ou *Exchangeable image file format* est une spécification de format de fichier pour les images utilisées par les appareils photographiques numériques. Cette spécification repose sur des formats existants tels que le JPEG, JPG, TIFF (mais pas le JPEG 2000 ni le PNG). Elles sont généralement conservées automatiquement avec chaque photographie. Des logiciels de traitement d'images permettent aisément de consulter ces métadonnées.

[12] Pour en savoir plus sur l'espace de nom EXIF.

Des métadonnées sur les traitements réalisés avec eScriptorium

Pour finir, la plateforme eScriptorium permet un certain nombre d'opérations comme la binarisation, l'entrainement de modèles, la segmentation, ou la transcription à partir des images de répertoires. Toutes les opérations génèrent autant de métadonnées de traitement propres au projet LectAuRep. Celles-ci impliquent que l'image est modifiée par un opérateur et/ou un relecteur (qui dispose d'un compte utilisateur stocké dans une base de données SQL, qui conserve l'historique des révisions effectuées sur les documents, eux-mêmes conservés en base).

Dans eScriptorium, il possible pour un utilisateur de réaliser des exports au format XML ALTO ou PAGE [13] pour récupérer des informations relatives à la mise en page physique et à la structure logique d'un texte transcrit par reconnaissance optique de caractères (OCR) comprenant les coordonnées des segments, le taux de confiance en la reconnaissance ou encore des éléments de mise en forme (polices ou graphies).

Extrait d'un export XML ALTO après segmentation et transcription réalisée dans eScriptorium :

```
<TextLine ID="eSc_line_31857" BASELINE="1349 1186 1944 1186 2503 1197"
    HPOS="1349" VPOS="1096" WIDTH="1154" HEIGHT="134">
          <Shape>
            <Polygon POINTS="1349 1186 1360 1150 1406 1132 1453 1147
                1526 1132 1544 1150 1565 1132 1728 1154 1807 1143 1829
                1161 1919 1136 1937 1154 2002 1147 2063 1103 2114 1096
                2243 1114 2492 1107 2503 1197 2485 1230 2409 1212 2240
                1215 2160 1194 2124 1208 1926 1201 1912 1215 1619 1201
                1569 1219 1547 1201 1518 1222 1504 1208 1378 1212 1352
                1186"/>
          </Shape>
      <String CONTENT="Murel (après décès de Gabrielle Jeanne Coste Ve
          de François Samuel Louis)" HPOS="1349" VPOS="1096" WIDTH="
          1154" HEIGHT="134"></String>
</TextLine>
```

[13] Les formats XML ALTO et PAGE XML sont adaptés à la conservation à long terme des données issues de la conversion comme l'*Optical character recognition* (OCR) ou l'*Handwritten text recognition* (HTR).

Conclusion

Cet état des lieux des métadonnées brosse le tableau d'un paysage désorganisé d'informations, certes hétérogènes, mais essentielles pour les utilisateurs finaux du projet. Un point commun, peut-être, entre toutes ces métadonnées : le format XML. Les prochaines étapes consisteront donc à opérer des choix dans ces métadonnées et trouver le moyen d'agréger celles-ci au sein d'un format pivot TEI pour les rendre facilement interrogeables.

« Un format pour les gouverner toutes. Un format pour les trouver. Un format pour les amener toutes et dans eScriptorium les lier », avis aux cinéphiles.