

ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Margaux Faure

licenciée ès lettres

diplômée de master en histoire

Rassembler et valoriser les mazarinades

Traiter les numérisations et les métadonnées
d'une collection d'imprimés du XVII^e siècle
pour sa mise à disposition numérique

Mémoire pour le diplôme du master
« Technologies numériques appliquées à l'histoire »

2022

Résumé

Le projet Antonomaz (STIH - Sorbonne Université) entreprend de réunir numériquement le corpus dit des mazarinades, ensemble de plus de 5 000 documents parus pendant la Fronde, au milieu du XVII^e siècle. Trois bibliothèques numériques sont principalement à l'origine de la mise à disposition numérique d'une partie du corpus : Mazarinum (Bibliothèque Mazarine), Gallica (Bibliothèque nationale de France) et Google Books. Après la récupération de ces documents sous format PDF, la production automatique de fichiers en XML-TEI contenant les métadonnées et le texte océrisé de la mazarinade, la problématique s'est posée, dans le cadre de la mise en ligne du projet, de pouvoir lier sur une même page web, chaque fichier XML-TEI à la numérisation correspondante. Une chaîne de traitement des métadonnées et des images a donc du être mise en place pour proposer une solution technique, en portant attention à la qualité des données et aux enjeux d'intéropérabilité. Une réflexion sur le moyen de visionnage des numérisations a été pensée parallèlement à une proposition d'affichage des informations existantes autour de celles-ci et fait l'objet de ce présent mémoire.

Mots-clefs : mazarinades ; XVII^e siècle ; Fronde ; TEI ; ODD ; IIIF ; Python ; TEI Publisher ; collection numérique ; chaîne de traitement.

Informations bibliographiques : Margaux Faure, *Rassembler et valoriser les mazarinades. Traiter les numérisations et les métadonnées d'une collection d'imprimés du XVIIe siècle pour sa mise à disposition numérique*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Thibaut Clérice, École nationale des chartes, 2022.

Remerciements

Je tiens à remercier toute l'équipe du projet Antonomaz pour leur accueil, leur gentillesse et leur disponibilité. Je remercie Karine Abiven et Gaël Lejeune de m'avoir offert la possibilité de prendre part à la réflexion et à la réalisation de ce projet. Merci à Alexandre Bartz pour son accompagnement technique, son écoute constante et tous les moments de sympathie qui ont rythmé le travail. Mes remerciements vont également à Zoé Cappe qui a partagé ces mois de stage avec moi.

Merci à toute l'équipe d'enseignement du master 2 « Technologies numériques appliquées à l'histoire » (2021-2022) pour leur encadrement pédagogique. Merci à Thibaut Clérice d'avoir dirigé ce mémoire.

J'ai une pensée toute particulière pour notre promotion de master, au sein de laquelle a régné une ambiance tant studieuse qu'amicale.

Liste des sigles et abréviations

- API : *Application Programming Interface*
- CNN : *Convolutional Neural Network*
- CSS : *Cascading Style Sheets*
- HTML : *HyperText Markup Language*
- HTR : *Handwritten Text Recognition*
- HTTP(S) : *HyperText Transfer Protocol (Secure)*
- IIIF : *International Image Interoperability Framework*
- ODD : *One Document Does it all*
- OCR : *Optical Character Recognition*
- PDF : *Portable Document Format*
- PNG : *Portable Network Graphics*
- TEI : *Text Encoding Initiative*
- URI : *Uniform Resource Identifier*
- URL : *Uniform Resource Locator*
- XML : *eXtensible Markup Language*
- XSLT : *eXtensible Stylesheet Language Transformations*

Bibliographie

La Fronde et les mazarinades

- ABIVEN (Karine), “Le moment discursif des barricades d’août 1648 : quelle interprétation des récurrences dans le discours sur l’événement ?”, *Cahiers de Narratologie*–35 (sept. 2019), DOI : 10.4000/narratologie.9264.
- BRIÈRE (Nina), *La douceur du roi. Le gouvernement de Louis XIV et la fin des Frondes 1648-1661*, Presses de l’Université Laval, Laval, 2011.
- CARRIER (Hubert), *La presse de la Fronde (1648-1653) : les Mazarinades. Tome 1. La conquête de l’opinion*, Librairie Droz, Genève, 1989.
- *La presse de la Fronde (1648-1653) : les Mazarinades. Tome 2. Les hommes du livre*. Librairie Droz, Genève, 1991.
- JOUHAUD (Christian), *Mazarinades. La Fronde des mots*, Flammarion, Paris, 2009.
- “Frontières des mazarinades, l’Inconnu et l’événement”, dans *Écritures de l’événement : les Mazarinades bordelaises*, dir. Myriam Tsimbidy, 2015, p. 17-25, DOI : 10.4000/books.pub.15678.
- LABADIE (Ernest), *Nouveau supplément à la bibliographie des mazarinades*, Librairie Henri Leclerc, Paris, 1904.
- MADELIN (Louis), *Une révolution manquée : la Fronde*, Plon, Paris, 1931.
- Mazarinum*, URL : <https://mazarinum.bibliotheque-mazarine.fr>.
- MOREAU (Célestin), *Bibliographie des mazarinades : publiée pour la Société de l’histoire de France*, J. Renouard, Paris, 1850.
- “Supplément à la Bibliographie des Mazarinades”, *Bulletin du bibliophile et du bibliothécaire* (, 1862), p. 786-829.
- “Supplément à la Bibliographie des Mazarinades”, *Bulletin du bibliophile et du bibliothécaire* (, 1869), p. 61-81.
- NICOLAS (Jean), *La rébellion française : mouvements populaires et conscience sociale (1661-1789)*, Éditions du Seuil, Paris, 2002.
- PERNOT (Michel), *La Fronde*, Éditions de Fallois, Paris, 1994.

PONCET (Olivier), *Mazarin. L'art de gouverner*, Perrin/Bibliothèque nationale de France, Paris, 2021.

RANUM (Orest), *La Fronde*, Éditions du Seuil, Paris, 1995.

RODIER (Yann), *Les raisons de la haine. Histoire d'une passion dans le France du premier XVIIe siècle*, Champ Vallon, Ceyzérieu, 2019.

SOCARD (Émile), *Supplément à la Bibliographie des Mazarinades*, H. Menu, Paris, 1876.

Les technologies autour du IIIF et du traitement de l'image

BERMÈS (Emmanuelle) et MARTIN (Frédéric), “Le concept de collection numérique”, *Bulletin des bibliothèques de France*–3 (2010), p. 13-17, URL : <https://bbf.enssib.fr/consulter/bbf-2010-03-0013-002>.

Biblissima, URL : <https://projet.biblissima.fr> (visité le 10/09/2022).

Cantaloupe, URL : <https://cantaloupe-project.github.io> (visité le 17/08/2022).

e-editiones, URL : <https://e-editiones.org> (visité le 07/09/2022).

IIIF Online Workshop, URL : <https://training.iiif.io/iiif-online-workshop/> (visité le 29/08/2022).

International Image Interoperability Framework, URL : <https://iiif.io> (visité le 12/08/2022).

Mirador, URL : <http://projectmirador.org> (visité le 28/08/2022).

Mirador version 3, mai 2019, URL : <https://library.stanford.edu/blogs/stanford-libraries-blog/2019/05/introducing-mirador-3-next-generation-image-comparison-viewer> (visité le 26/07/2022).

OpenSeadragon, URL : <https://openseadragon.github.io> (visité le 13/09/2022).

OYALLON (Edouard), *Analyzing and Introducing Structures in Deep Convolutional Neural Networks*, thèse de doct., Paris Sciences et Lettres, 2017, URL : <https://hal.archives-ouvertes.fr/tel-02353134>.

Presentation API Validator, URL : <https://presentation-validator.iiif.io> (visité le 13/09/2022).

ROBINEAU (Régis), *Comprendre IIIF et l'interopérabilité des bibliothèques numériques*, nov. 2016, URL : <https://insula.univ-lille.fr/2016/11/08/comprendre-iiif-interoperabilite-bibliotheques-numeriques/> (visité le 13/08/2022).

VAN ZUNDERT (Joris), “On Not Writing a Review about Mirador : Mirador, IIIF, and the Epistemological Gains of Distributed Digital Scholarly Resources”, *Digital Medievalist*, 11–1 (août 2018), p. 5, DOI : 10.16995/dm.78.

Visualiseur Mirador, URL : <https://doc.biblissima.fr/visualiseur-mirador> (visité le 26/07/2022).

Données et Science ouverte

BERMÈS (Emmanuelle) et POUPEAU (Gautier), “1. Du catalogue de bibliothèque aux données sur le Web : un changement de paradigme du côté de l’usager”, dans *Le Web sémantique en bibliothèque*, Paris, 2013 (Bibliothèques), p. 17-28, URL : <https://www.cairn.info/le-web-semantique-en-bibliotheque--9782765414179-p-17.htm>.

CNRS (Inist -), “DoRANum-Enjeux et bénéfices : Cycle de vie des données, un outil pour améliorer la gestion, la mise en qualité et l’ouverture des données” (, 2021), Publisher : DoRANum, DOI : 10.13143/GDBG-CF63.

“Diffusion et exploitation d’un document numérique : information et mise en garde des usagers”, dans *Manuel de constitution de bibliothèques numériques*, Editions du cercle de la librairie, Paris, 2013.

Dublin Core, URL : <https://www.bnf.fr/fr/dublin-core> (visité le 13/09/2022).

JACQUEMIN (Bernard), SCHÖPFEL (Joachim) et FABRE (Renaud), “Libre accès et données de recherche. De l’utopie à l’idéal réaliste”, *Études de communication*-52 (juin 2019), p. 11-26, DOI : 10.4000/edc.8468.

L’identifiant ARK (Archival Resource Key), URL : <https://www.bnf.fr/fr/lidentifiant-ark-archival-resource-key> (visité le 13/09/2022).

MÉDITERRANNÉE (Urfist), “DoRANum-Enjeux et bénéfices : les principes FAIR” (), Publisher : DoRANum, DOI : 10.13143/Z7S6-ED26.

VANHOLSBECK (Marc), “La notion de Science Ouverte dans l’Espace européen de la recherche : Entre tendances à l’« exotérisation » et à la « gestionnarisation » de la recherche scientifique”, *Revue française des sciences de l’information et de la communication*-11 (août 2017), DOI : 10.4000/rfsic.3241.

L’édition numérique

BURNARD (Lou), “Conclusion : qu’est-ce que la TEI?”, dans *Qu’est-ce que la Text Encoding Initiative ?*, OpenEdition Press, Marseille, 2015, URL : <http://books.openedition.org/oep/1305> (visité le 28/08/2022).

GALLERON (Ioana), DEMONET (Marie-Luce), MEYNARD (Cécile), IDMHAND (Fatiha), PIERAZZO (Elena), WILLIAMS (Geoffrey), BUARD (Pierre-Yves) et ROGERI (Julia), *Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations*, rapp. tech., 2018, URL : <https://halshs.archives-ouvertes.fr/halshs-01932519/document> (visité le 23/08/2022).

HAROLD (Elliotte Rusty), SCOTT MEANS (W) et ENSARGUET (Philippe), *XML en concentré*, O'Reilly, Paris, 2005.

Le consortium "CAHIER" : Corpus d'Auteurs pour les Humanités, URL : <https://cahier.hypotheses.org/le-consortium> (visité le 02/09/2022).

PUREN (Marie) et VERNUS (Pierre), *AGODA (Analyse sémantique et graphes relationnels pour l'ouverture et l'étude des débats à l'assemblée nationale)*, URL : https://hal.archives-ouvertes.fr/hal-03382765/file/puren_vernus_datalab181021.pdf (visité le 13/09/2022).

TEI Guidelines, URL : <https://tei-c.org/guidelines/>.

XPath Tutorial, URL : https://www.w3schools.com/xml/xpath_intro.asp (visité le 13/09/2022).

Développement web

BARCENILLA (Javier) et BASTIEN (Joseph Maurice Christian), "L'acceptabilité des nouvelles technologies : quelles relations avec l'ergonomie, l'utilisabilité et l'expérience utilisateur ?: " *Le travail humain*, Vol. 72-4 (mars 2010), p. 311-331, DOI : 10.3917/th.724.0311.

CHIFFOLEAU (Floriane), *Publication of my digital edition – Working with TEI Publisher*, URL : <https://digitalintellectuals.hypotheses.org/3912> (visité le 01/07/2022).

CHIFFOLEAU (Floriane) et OVIDE (Manon), *Publication of my digital edition – Developing my TEI Publisher application*, URL : <https://digitalintellectuals.hypotheses.org/4173> (visité le 01/07/2022).

Construire une bibliothèque numérique, URL : <https://www.idnum.fr/methodoc/construire-une-bibliotheque-numerique/> (visité le 07/09/2022).

HTTPS (HyperText Transfert Protocol Secure) : définition claire et pratique, URL : <https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203459-https-hypertext-transfert-protocol-secure-definition/> (visité le 13/09/2022).

La face cachée du numérique, nov. 2019, URL : <https://librairie.ademe.fr/cadic/2351/guide-pratique-face-cachee-numerique.pdf?modal=false> (visité le 24/08/2022).

Manifeste des Digital humanities, mars 2011, URL : <https://tcp.hypotheses.org/318> (visité le 10/09/2022).

Manipulating content via JS, URL : <https://unpkg.com/@teipublisher/pb-components@1.38.4/dist/api.html#pb-view.3> (visité le 26/08/2022).

Omeka Classic. Un environnement de recherche pour les éditions scientifiques numériques, URL : <https://ride.i-d-e.de/issues/issue-11/omeka/> (visité le 13/09/2022).

Présentation d'Omeka, URL : <https://omeka.fr/presentation-omeka> (visité le 10/09/2022).

RAVEREAU (Mylène), *L'uniformisation et la pérennité des informations numériques dans les bibliothèques numériques : le cas du logiciel libre Omeka*, mémoire sous la direction de Jean-Baptiste Camps, École nationale des chartes, 2016.

Subscribe, URL : <https://reactivex.io/documentation/operators/subscribe.html> (visité le 13/09/2022).

TEI Publisher, URL : <https://teipublisher.com> (visité le 07/09/2022).

Une API, qu'est-ce que c'est ?, URL : <https://www.redhat.com/fr/topics/api/what-is-a-rest-api> (visité le 11/09/2022).

Introduction

Je ne crois pas m'avancer trop en disant que jusqu'ici on n'avait pas encore étudié les Mazarinades dans leur ensemble ; qu'on s'était contenté d'apprécier isolément celles que l'on avait rencontrées, sans les chercher peut-être ; qu'on s'était borné à quelques anecdotes vérifiées avec peu de soin, à quelques jugements acceptés sans contrôle, et qu'ainsi la bibliographie des pamphlets de la Fronde était un travail à faire en quelque sorte tout entier.¹

Au milieu du XIX^e siècle, Célestin Moreau pointe la nécessité de constituer une liste complète des mazarinades, cet ensemble de plus de 5 000 documents, sorte d'écrits d'actualité parus en grande majorité pendant la Fronde (1648-1653).

Les mazarinades se conçoivent comme un regroupement d'entités textuelles aux genres littéraires hétérogènes qu'il faut pouvoir étudier comme un ensemble. Si des listes de titres de documents sont déjà constituées, la question de l'accessibilité de leur contenu se place au centre d'attention du projet Antonomaz.

Le projet Antonomaz (ANalyse auTOmatique et NumérisatiOn des MAZarinades) est un projet universitaire du laboratoire STIH (Sens-Texte-Informatique-Histoire) de Sorbonne Université². Cette unité de recherche est intégrée à l'École doctorale V « Concepts et Langage » de la Faculté des Lettres et dépend de l'UFR de Langue française. Le projet s'inscrit dans le cadre de l'axe « Histoire des usages linguistiques : approches diachroniques et textuelles » dédiée à l'étude des usages de la langue française, depuis le Moyen Âge³.

L'équipe de ce laboratoire est interdisciplinaire, composée de sociologues, d'informatiens et de chercheurs en littérature. Les spécialités des membres du projet Antonomaz reflètent bien cette dynamique : les deux chefs de projets, Karine Abiven et Gaël Lejeune, sont maîtres de conférences, respectivement en littérature et en informatique. Une troisième personne, Alexandre Bartz, travaille en tant qu'ingénieur sur le projet. Jean-Baptiste Tanguy, doctorant en traitement automatique des langues, est également intervenu dans une étape de traitement.

Antonomaz est un projet universitaire à la rencontre des études littéraires, historiques et informatiques, réalisé à partir de documents numériques sur lesquels un traitement est appliqué. En cela, il s'intègre tout à fait dans le champs des « humanités numé-

1. Célestin Moreau, *Bibliographie des mazarinades : publiée pour la Société de l'histoire de France*, J. Renouard, Paris, 1850, tome 1, p. II.

2. <http://stih-sorbonne-universite.fr> (visité le 28/08/2022)

3. « Elle [l'équipe] comporte un axe autour de la philologie, de l'histoire du livre, de la traduction et des éditions de textes ; un axe autour de la stylistique, de la rhétorique, de la poétique, de l'analyse du discours et des représentations de la langue ; un axe autour de la linguistique diachronique ; et un axe autour des langues de spécialité. Elle aborde notamment la problématique du changement linguistique, y compris dans une perspective comparatiste entre le français et les autres langues européennes. » Voir : <http://stih-sorbonne-universite.fr/research.html>

riques », cette « transdiscipline, porteuse des méthodes, des dispositifs et des perspectives heuristiques liés au numérique dans le domaine des Sciences humaines et sociales⁴ ».

En effet, Antonomaz ne prétend pas constituer une collection complète de mazarinades qui seraient conservées en un seul lieu physique mais plutôt travailler à repérer et rassembler celles disponibles sur le web pour qu'elles puissent être lues depuis un unique site internet.

Permettre de consulter numériquement un document papier nécessite que celui-ci soit numérisé. Cette dématérialisation facilite l'accès à la source qui n'est plus restreinte à son lieu de conservation et traduit ce même document physique en un ensemble de données assimilables et manipulables par l'ordinateur. Cependant, les numérisations qui sont l'objet de notre attention sont actuellement diffusées sur le web par plusieurs entités, propriétaires des documents physiques et/ou de leur format numérique. À l'heure où l'accès distant aux ressources ne cesse de se déployer, l'éparpillement numérique de ces nombreux documents, pourtant considérés comme un ensemble scientifique, rend leur consultation quelque peu fastidieuse pour le chercheur qui, parmi ces 5 à 6 000 entités, ne peut réellement savoir quelles institutions les possèdent et quels éléments de leurs collections sont accessibles sur internet.

Si les mazarinades s'étudient largement comme un ensemble scientifique résultant d'une production cohérente, elles ne correspondent pas à la logique d'un fonds d'archives déterminé par un unique producteur. Elles sont plutôt constituées en petites collections disséminées internationalement. Aussi, le projet Antonomaz propose-t-il de centraliser les mazarinades répertoriées sur plusieurs sites internet pour constituer une collection numérique. Il ne s'agit pas de concentrer toute l'offre disponible sur le web, autrement dit de permettre la consultation de tous les exemplaires d'un même texte, mais plutôt de pouvoir proposer une lecture numérique du texte, *via* la mise à disposition d'un exemplaire, et cela pour tous les documents repérés et qualifiés de « mazarinades ».

De plus, le projet ne souhaite pas seulement proposer un rassemblement d'images mais également produire un enrichissement autour des numérisations : des métadonnées fines donneront des éléments de compréhension pour la consultation et un traitement du texte, alors détaché de l'image, permettra sa fouille. Le tout doit répondre le mieux possible aux principes FAIR (*Findable Accessible Interoperable Reusable*), promouvant la science ouverte. Ces principes « fournissent des lignes directrices pour améliorer la facilité de repérage, l'accessibilité, l'interopérabilité et la réutilisation des ressources numériques. Ces principes sont très axés sur la capacité des machines à gérer des données de façon

4. *Manifeste des Digital humanities*, mars 2011, URL : <https://tcp.hypotheses.org/318> (visité le 10/09/2022).

automatique, avec le minimum d'interventions humaines.⁵ »

Un travail d'édition complète donc la mise à disposition des numérisations. Ces envies conditionnent les moyens techniques à mettre en œuvre devant être efficaces pour débuter une phase de récupération des numérisations puis la mise en place d'une chaîne de traitement permettant de préparer les données en vue de leur visualisation.

Le présent mémoire rend compte d'un stage effectué d'avril à mi-août 2022, dans le cadre de la seconde année du master « Technologies numériques appliquées à l'histoire » de l'École nationale des chartes. Il tentera de présenter les réalisations et les réflexions qui ont occupé les mois passés au sein de l'équipe du projet Antonomaz.

En cours depuis 2019⁶, l'état général du projet était déjà bien avancé à notre arrivée. La plupart des données étaient déjà récupérées, traitées ou en cours de traitement. Situé au moment du basculement entre la fin de préparation des données et la phase de publication, notre principal objectif de stage était de pouvoir produire une page de consultation des documents pour le site web qui accueillera le corpus. Pour cela, tout un travail préparatoire était nécessaire.

Comment donner sens à un ensemble de données issues de la numérisation pour rendre un document numériquement consultable tout en valorisant sa matérialité ? Comment aligner la réutilisation de contenu récupéré et la création d'une nouvelle matière afin de proposer un résultat cohérent, harmonieux et agréable à consulter ? Autrement dit, comment mettre différents éléments en relation pour valoriser des données de recherche tout en respectant les droits de diffusion et de réutilisation ? L'appropriation d'un outil d'édition numérique a été l'occasion de tester sa flexibilité, conçu comme une boîte à outils personnalisable pour répondre aux besoins de l'édition électronique. Quels ont été les défis techniques rencontrés ?

Deux axes principaux peuvent être définis en terme de problématique numérique dans le cadre de nos réalisations de stage. Après la présentation du corpus, de son contexte historique et de l'état de son traitement par le projet à notre arrivée, le premier axe s'articulera autour du travail de traitement de l'image devant permettre d'aligner toutes les numérisations en notre possession selon le principe d'intéropérabilité. Le second se centrera sur la construction de la page web, tentant de faire répondre les possibilités techniques aux attentes du futur utilisateur consultant une mazarinade.

5. Urfist Méditerranée, “DoRANum-Enjeux et bénéfices : les principes FAIR” (), Publisher : DoRANum, DOI : 10.13143/Z7S6-ED26.

6. Le projet reçoit un financement du DIM STCN de la région Île-de-France (Sciences des textes et connaissances nouvelles) (<https://www.dim-humanites-numeriques.fr/projets/antonomaz-analyse-automatique-et-numerisation-des-mazarinades/>) puis en 2020 de l'Institut Universitaire de France.

Première partie

**Le projet Antonomaz et les
mazarinades : présentation du
corpus et de sa chaîne de traitement**

Chapitre 1

Le corpus des mazarinades : l'expression des frondeurs et des anti-frondeurs

1.1 Contexte historique : La Fronde

1.1.1 De la fronde à la Fronde

Avant de désigner une période historique, le terme de « fronde » qualifie un objet, un type de lance-pierre dont la lanière de cuir permet de lancer des cailloux, à des fins de jeux ou d'affrontements armés.

Le terme a été repris pour évoquer les troubles de ce milieu du XVII^e siècle, pouvant s'arrêter à la simple agitation mais également prendre des dimensions plus impressionnantes, laissant des victimes après l'évènement. Aussi, la période de la Fronde renvoie à des années de troubles (1648-1653), un ensemble de « mouvements » dont la qualification saisit le débat historiographique.

De la révolte à la révolution, cette période de contestations est qualifiée à des niveaux d'intensité variables. Sous l'Ancien Régime, le terme de « révolte » est utilisé par les autorités pour qualifier des événements même minimes. Au degré moindre que la révolution mais dépassant la simple escarmouche, la révolte est l'expression d'un mécontentement et provient d'une émotion populaire qui dégénère suite à un engouement, une réaction impulsive commune qui mène au débordement mais dont l'étendue est restreinte¹.

Aussi, Nina Brière exclut-elle le terme de « révolution » en préférant celui de « révolte » en précisant que « la Fronde ne véhicule aucune idéologie et ne monopolise pas

1. Jean Nicolas, *La rebellion française : mouvements populaires et conscience sociale (1661-1789)*, Éditions du Seuil, Paris, 2002.

tous les habitants ni toutes les provinces du royaume de France. Quant à la Fronde des grands, ou petite-fronde, plus précisément, de nombreux grands nobles de France demeurent loyaux au service du roi.² »

Parfois qualifiée de « révolution manquée³ », le fil des évènements identifie, en tout cas, une évolution des acteurs impliqués, touchant les différents pans de la société, ce qui formerait « plutôt une révolte parlementaire et nobiliaire qui a gagné la population de Paris et de certaines grandes villes de province⁴ ». Ce mouvement de contestations ne serait pas tant une volonté de renverser complètement l'ordre des choses que de s'opposer à un abus des détenteurs de l'autorité et à leur manière de l'appliquer : « Aux yeux des Français du XVII^e siècle, la Fronde fut un mouvement de soutien toujours plus large et plus profond à la violation de plus en plus radicale et délibérée des lois du roi.⁵ »

Sujette à débat, la Fronde est, comme le résume Christian Jouhaud, un ensemble de troubles discontinus⁶, suscités dans une période de contestation de la politique du ministre Jules Mazarin dont la position au gouvernement est peu populaire.

1.1.2 La figure de Mazarin

Jules Mazarin (1602-1661) est nommé Premier ministre en 1643 sous la régence d'Anne d'Autriche. Sa présence au gouvernement est critiquée bien avant les évènements de la Fronde : d'une part, sa nationalité italienne est dépréciée⁷, d'autre part les contemporains lui reprochent une position trop dominante dans la politique du royaume, aliénant la régente.

Mazarin est catégorisé comme « l'héritier de la politique de Richelieu⁸ ». Il est largement accusé d'une politique menée par la poigne mais aussi tenu responsable de la continuité de la guerre, notamment de ne pas réussir à mettre fin aux conflits avec l'Espagne, qui épouse les caisses de l'État. Cette situation économique se répercute sur la

2. Nina Brière, *La douceur du roi. Le gouvernement de Louis XIV et la fin des Frondes 1648-1661*, Presses de l'Université Laval, Laval, 2011, p. 8.

3. Louis Madelin, *Une révolution manquée : la Fronde*, Plon, Paris, 1931.

4. Karine Abiven et Gaël Lejeune, “Analyse automatique de documents anciens : tirer parti d'un corpus incomplet, hétérogène et bruité”, ISTE OpenScience 2–1 (2019), URL : <https://hal.archives-ouvertes.fr/view/index/identifiant/hal-02467535>.

5. Orest Ranum, *La Fronde*, Éditions du Seuil, Paris, 1995, p. 12.

6. « La Fronde a duré cinq ans, en une suite compliquée de soubresauts, de combats, de paix, de tensions, d'accalmies. » Christian Jouhaud, *Mazarinades. La Fronde des mots*, Flammarion, Paris, 2009, p. 20.

7. Mazarin est perçu comme « un étranger venu s'enrichir chez eux ». Sa nationalité le positionne comme le parfait bouc émissaire qui ne mériterait pas une telle place de conseil auprès du roi, réservée aux grands du royaume. N. Brière, *La douceur du roi. Le gouvernement de Louis XIV et la fin des Frondes 1648-1661...*, p. 22.

8. Olivier Poncet, *Mazarin. L'art de gouverner*, Perrin/Bibliothèque nationale de France, Paris, 2021, p. 60.

population, subissant une hausse de la fiscalité. Aussi, des institutions comme le parlement de Paris, aspirant à exercer un pouvoir de tempérance sur la politique royale et se positionnant comme défenseur du peuple, prennent une position d'hostilité pour contester la politique du Ministre.

Au cours de la Fronde, son exil est réclamé à plusieurs reprises et la haine de Mazarin formerait, chez les grands nobles du royaume, un terrain d'entente⁹.

1.1.3 Déroulé des évènements

Durant 7 ans, une succession d'évènements trouble le royaume, particulièrement à Paris et dans quelques villes de province comme Bordeaux ou Rouen. Hubert Carrier dresse une chronologie de la Fronde dans son ouvrage sur les mazarinades¹⁰. Il distingue deux phases principales : la « Fronde parlementaire » et la « Fronde des Princes ».

Au début de l'année 1648, Mazarin convoque un lit de justice au cours duquel Omer Talon, premier avocat au Parlement, prononce un discours contestant l'enregistrement de nouveaux édits fiscaux¹¹. Imprimé à plusieurs reprises et diffusé en province, son propos s'inscrit dans un contexte de tensions entre les institutions parlementaires et la monarchie. Se positionnant comme défenseur du peuple et désireux d'un pouvoir de validation des décisions royales, le parlement de Paris s'associe aux autres cours souveraines (Cour des aides, Grand conseil et Chambre des comptes) pour élaborer un programme de réforme, la *Déclaration de la Chambre Saint-Louis*¹², fort d'une volonté d'obtention d'un pouvoir de tempérance de la politique royale.

Les principales réformes demandées dans la *Déclaration* sont validées par la monarchie à la fin du mois de juillet. Cependant, l'arrestation le 26 août du conseiller Pierre Broussel ravive les oppositions. Très populaire dans la capitale, les Parisiens manifestent leur mécontentement : des barricades sont construites dans Paris afin d'éviter le passage des troupes royales et obtenir sa libération. Cet épisode, connu sous le nom de « journées des barricades », souligne l'investissement de la population et est considéré comme le « point de départ le plus visible¹³ » de la Fronde.

9. N. Brière, *La douceur du roi. Le gouvernement de Louis XIV et la fin des Frondes 1648-1661...*, p. 20.

10. Hubert Carrier, *La presse de la Fronde (1648-1653) : les Mazarinades. Tome 2. Les hommes du livre*. Librairie Droz, Genève, 1991, pp. 395-402.

11. Omer Talon, *Harangue faite au roi par monsieur Talon, son avocat général au Parlement de Paris*, Paris, 1649.

12. Le texte est accessible ici : *Journal contenant tout ce qui s'est fait et passé en la cour de Parlement de Paris, toutes les chambres assemblées, sur le sujet des affaires du temps présent*, Paris, Gervais Alliot et Jacques Langlois, 1648.

13. K. Abiven, “Le moment discursif des barricades d'août 1648 : quelle interprétation des récurrences dans le discours sur l'événement?”, *Cahiers de Narratologie*-35 (sept. 2019), DOI : 10.4000/narratologie.9264.

À l'hiver 1649, par crainte d'une prise d'otage de la famille royale, cette dernière fuit avec la Cour vers le château de Saint-Germain dans la nuit du 5 au 6 janvier. Le prince de Condé¹⁴ envoie des troupes pour contenir à la fois l'agitation du Parlement, qui considère toujours ses prérogatives bridées, et le bouillement de la rue. Le 8 janvier, le Parlement déclare Mazarin, « Autheur de tous les desordres de l'Estat » et « Perturbateur du repos public¹⁵ ».

La Conférence de Rueil, tenue en mars 1649, amène à un accord dans lequel le Parlement renonce à désigner le bannissement du Ministre comme solution à la résolution des troubles. La paix est véritablement signée le 1^{er} avril 1649 (paix de Saint-Germain).

Cette signature ne ramène pas pour autant le calme dans le royaume : d'une part les troubles ont atteint les villes de province dont Bordeaux et Aix-en-Provence, d'autre part, le prince de Condé cherche à obtenir mérite de son action militaire à Paris.

L'année 1650 marque le début de la « Fronde des Princes ». Le prince de Condé, le prince de Conti, son frère, et leur beau-frère, le duc de Longueville entendent profiter de l'affaiblissement de la Régence pour obtenir le pouvoir. Les relations de plus en plus tendues entre Condé et Mazarin amènent le Ministre à se rapprocher d'anciens chefs frondeurs pour faire arrêter les trois Princes. Ils sont incarcérés le 18 janvier 1650 au château de Vincennes.

Le bruit de l'affaire gagnant les provinces, des troupes sont envoyées pour freiner les tensions en Guyenne, en Bourgogne ou encore en Normandie. Le 30 janvier 1651, un traité est conclu entre Gaston d'Orléans, les Frondeurs et les partisans des Princes pour leur libération et le bannissement de Mazarin. Le 6 février, Mazarin fuit la capitale après sa rupture avec Gaston d'Orléans. Il est exilé par le parlement de Paris. Les Princes sont libérés et reviennent à Paris dans le mois.

Dès l'été 1651, l'intensité de la Fronde décline. La violence de Condé, nommé gouverneur en Guyenne, inquiète. Avec une situation économique toujours mauvaise, la stabilité se cherche en la personne du Roi, à présent majeur. Louis XIV revient à Paris le 21 octobre 1652 et Mazarin le 3 février 1653. Leurs retours réaffirment l'autorité royale.

Au fil de ces événements, paraissent les mazarinades, des documents qui nous ap-

14. Il s'agit de Louis II, prince de Condé (1621-1686).

15. *Arrest de la cour de Parlement donné toutes les chambres assemblées le 8. jour de janvier 1649. Par lequel il est ordonné que le cardinal Mazarin videra le Royaume, & qu'il sera fait levée de gens de guerre pour la seureté de la Ville, & pour faire amener & apporter seulement & librement les vivres à Paris, Imprimeurs et libraires ordinaires du roi, 1649,* disponible à l'adresse : <https://mazarinum.bibliotheque-mazarine.fr/viewer/15095/?offset=#page=9&viewer=picture&o=bookmark&n=0&q=>.

prennent « au jour le jour, comment les Français du XVII^e siècle ont vécu la Fronde¹⁶ ». Présentons à présent la nature de ces textes.

1.2 Qu'est-ce qu'une mazarinade ?

1.2.1 Origine et signification du mot

Selon Hubert Carrier, la première apparition du terme « mazarinade » se lit dans un triolet de Marigny sur l'échec du siège de Cambrai en juillet 1649¹⁷ avec comme signification :

[...] un tour de farceur, une facétie de bateleur, une singerie de bouffon.
[...] et c'est de ce sens de facétie ou d'attrape de farceur qu'on est passé à celui de mauvais tour, de combine, de fourberie du ministre.¹⁸

Par la suite, le poète Paul Scarron¹⁹ intitule un de ces écrits en reprenant le terme, contribuant largement à sa célébrité :

La Mazarinade est l'épopée de Mazarin comme *L'Iliade* est celle d'Ilion et *La Franciade* de Ronsard celle de Francus : une épopée à la mesure du héros, c'est-à-dire une parodie, une bouffonnerie, une caricature d'épopée, comme Jules n'est qu'une caricature de ministre.²⁰

Aussi ce mot, dont la référence au ministre est transparente, se dote d'une connotation négative assez précise : une mazarinade, c'est un écrit contre Mazarin. Les dictionnaires du XIX^e siècle l'enregistrent largement avec cette signification :

Nom donné à des pamphlets, satires, libelles en prose et en vers que les frondeurs publiaient contre Mazarin [...].²¹

Cependant, dans les faits, des écrits visant d'autres individus, ou même « pro-Mazarin », sont intégrés à ce vaste ensemble, venant élargir la définition. La diversité des écrits qualifiés de « mazarinades » appelle donc à la nuance.

16. H. Carrier, *La presse de la Fronde (1648-1653) : les Mazarinades. Tome 1. La conquête de l'opinion*, Librairie Droz, Genève, 1989, p. 205.

17. *Ibid.*, p. 60.

18. *Ibid.*

19. Paul Scarron (1610-1660) est un écrivain très actif au moment de la Fronde. Grande figure du genre burlesque, son style d'écriture a inspiré de nombreux pamphlets.

20. *Ibid.*, p. 61.

21. Pierre Larousse, *Grand dictionnaire universel du XIXe siècle : français, historique, géographique, mythologique, bibliographique*, Paris, 1866-1877, tome 10, p. 1388

1.2.2 L'hétérogénéité des textes

État général

Pour Christian Jouhaud, se questionner sur la qualification d'un document comme « mazarinade », c'est aussi se questionner sur ce qui n'est pas une mazarinade²². Définir un corpus par ses limites est une preuve de l'hétérogénéité de celui-ci. Sans en questionner les frontières, nous pouvons dresser un portrait général des écrits qualifiés de « mazarinade ».

N'est pas seulement considéré comme une mazarinade un texte écrit contre Mazarin. Bien qu'elle soit majoritairement produite dans cette intention, elle peut prôner diverses positions politiques.

Dans sa forme littéraire, une mazarinade est avant tout un libelle, c'est-à-dire un écrit satirique, diffamatoire, à l'image du pamphlet. Cependant, il est là aussi difficile de chercher une unité à ce corpus à l'échelle du genre littéraire tant les écrits sont variés : pamphlet, discours, chanson, remontrances, poème, lettre, périodique, récit fictionnel, harangue, épitaphe... Des textes officiels comme les arrêts, les ordonnances ou encore les lettres patentes s'inscrivent également dans le corpus. Côté auteurs, ce sont donc des individus ou institutions, identifiés ou anonymes, qui peuvent être frondeurs, anti-frondeurs, pro-Mazarin, anti-Mazarin.

Matériellement, la grande majorité des mazarinades sont des imprimés mais le texte manuscrit n'est pas exclu. Plusieurs chansons répertoriées ne sont conservées que sous une forme manuscrite.

Finalement, il semble que ce n'est pas tant la forme littéraire qui fonde la cohérence de cet ensemble qu'un facteur thématique articulé autour d'un moment de parution, d'un sujet traité ou d'une réception souhaitée. Peut être désigné comme mazarinade tout document traduisant ou éveillant l'opinion publique pendant la Fronde. Hubert Carrier limite la définition des mazarinades aux documents parus pendant les années de la Fronde²³ mais des parutions postérieures cohérentes sur tout point (hormis une date de publication de quelques années plus tardive) peuvent être incorporées.

Corpus ou collection ?

La difficulté à saisir cet ensemble amène des questionnements de la part des chercheurs sur le bon terme à appliquer pour le désigner.

22. C. Jouhaud, "Frontières des mazarinades, l'Inconnu et l'événement", dans *Écritures de l'événement : les Mazarinades bordelaises*, dir. Myriam Tsimbidy, 2015, p. 17-25, DOI : 10.4000/books.pub.15678, pp. 17-25.

23. H. Carrier, *La presse de la Fronde (1648-1653) : les Mazarinades. Tome 1. La conquête de l'opinion...*, pp. 60-69.

Si les mazarinades forment un ensemble de textes à étudier, elles n'ont pas été récoltées dans cette intention mais plutôt dans une logique de rassemblement de documents. Or, un corpus se constitue et se sélectionne avec une cohérence scientifique. C'est pour cela que le projet Antonomaz a plutôt fait le choix de préférer le terme de « collection » :

On entend ainsi conserver la variété documentaire qui en fait l'historicité, mais de la penser comme telle (comme une construction collectionneuse inexploitable comme corpus).²⁴

La définition de collection numérique émise par Emmanuelle Bermès et Frédéric Martin confirme l'application du terme à notre ensemble de documents : « un ensemble cohérent de documents, établi en vue d'un usage précis, faisant l'objet d'une gestion. Chacun des objets qui la composent a plus de valeur dans l'entité collective qu'il n'en aurait individuellement.²⁵ »

À titre personnel, il nous semble que les deux termes peuvent être appliqués selon la considération et le traitement appliqué à l'ensemble. Le mot « corpus » peut être utilisé lorsque l'on évoque les mazarinades comme un ensemble de textes étudiés, avec une approche scientifique (littéraire ou historique). Ce regroupement de textes permet de marquer l'élan et l'intensité des parutions de cette période troublée et en cela, possède une cohérence scientifique. Le terme de collection se justifie également lorsque l'on en vient à considérer le lien matériel (physique ou numérique) entre les documents, issus de plusieurs provenances (contrairement au fonds), regroupés autour d'un thème commun, pensés alors dans une logique de bibliothéconomie. C'est du moins pour ces raisons que nous utiliserons les deux termes au cours de notre travail.

1.2.3 Bibliographies et constitution du corpus

Bien que l'hétérogénéité présentée ci-dessus construise un large éventail de possibilités pour qualifier une mazarinade, la désignation d'un document comme telle est loin d'être évidente ou naturelle. Elle résulte d'un long travail de bibliographie, de considération et de repérage des documents. Un imprimé paru sous la Fronde n'est pas automatiquement assigné à ce terme, de même que cette assignation est dépendante d'un œil humain ayant été confronté à une source croisée et consultée.

24. K. Abiven, G. Lejeune, Alexandre Bartz et Jean-Baptiste Tanguy, “Vers une collection numérique des libelles parus pendant la fronde, ou comment relier des mazarinades”, *Le Verger-bouquet* XXIII (mai 2022), URL : <https://cornucopia16.com/blog/2022/04/24/karine-abiven-alexandre-bartz-gael-lejeune-et-jean-baptiste-tanguy-vers-une-collection-numerique-des-libelles-parus-pendant-la-fronde-ou-comment-relier-des-mazarinades/> (visité le 10/08/2022).

25. Emmanuelle Bermès et Frédéric Martin, “Le concept de collection numérique”, *Bulletin des bibliothèques de France*-3 (2010), p. 13-17, URL : <https://bbf.enssib.fr/consulter/bbf-2010-03-0013-002>.

On doit le premier gros travail de bibliographie à Célestin Moreau (1805-1882), journaliste féru de la littérature du XVII^e siècle, qui établit une *Bibliographie des Mazarinades*²⁶ pour la Société de l'Histoire de France²⁷. Au milieu du XIX^e siècle est donc publiée, en trois volumes, une bibliographie de 4 082 imprimés qualifiés de « mazarinades », suivie d'une liste des imprimeurs-libraires²⁸, d'une proposition de classement chronologique de parution des documents²⁹ et d'une table des noms propres³⁰.

Classées par ordre alphabétique, chaque mazarinade reçoit un numéro, correspondant à son placement dans la liste. Moreau indique son titre et les informations d'édition, liées à l'exemplaire consulté. Il se permet régulièrement d'ajouter des éléments de contexte, de rareté du document mais également des appréciations personnelles sur la qualité littéraire du texte.

3001. Récit véritable d'une action horrible faite dans l'église des pères de l'Oratoire à Paris, le 11^e jour de juin 1649, au grand étonnement d'un chacun qui assistaient (*sic*) à la sainte messe. *Paris*, 1649, 6 pages.

Il s'agit d'un frère de l'Oratoire qui, au moment de la consécration, s'est jeté sur l'officiant et l'a renversé par terre pour que l'hostie soit chue.

Rare et curieux.

B 162 3002. Récit véritable de ce qui fut dit à l'arrivée de messieurs les députés du Parlement de Normandie (à Ruel). *Paris*, Jean Dédin, 1649, 7 pages.

3003. Récit véritable de ce qui s'est passé à Chaliot (*sic*) à l'entrevue de messieurs les princes de Condé, de Conty, de madame de Longueville et autres princes. *Paris*, 1649, 8 pages.

Ce n'est pas un récit, mais une amplification dont il n'y a rien à dire, si ce n'est qu'elle est assez rare.

L'entrevue eut lieu le 5 avril.

FIGURE 1.1 – Extrait de la *Bibliographie des Mazarinades*

Célestin Moreau n'est pas le premier à avoir établi une liste de mazarinades mais il considère que :

[...] ces listes sont toujours fort incomplètes : elles ne contiennent guère que des titres réduits, qui ne peuvent pas aider le travailleur dans ses recherches ;

26. C. Moreau, *Bibliographie des mazarinades : publiée pour la Société de l'histoire de France...*, Les trois volumes sont disponibles sur Gallica, pour le premier tome, voir : <https://gallica.bnf.fr/ark:/12148/bpt6k298570q>.

27. Pour une biographie de Célestin Moreau, voir la notice du fonds éponyme, conservé à la Bibliothèque Mazarine : <https://www.bibliotheque-mazarine.fr/fr/collections/fonds-particuliers/celestinst-moreau>

28. *Ibid.*, tome 3, p. 283-294.

29. *Ibid.*, p. 299-386.

30. *Ibid.*, p. 387-418.

on y trouve à peine quelques renseignements sur les auteurs, sur l'origine et le caractère des pamphlets, sur la pensée politique qui les a dictés, sur les rapports de polémique qui existent entre plusieurs, sur les différentes éditions qui en ont été faites, enfin sur les obstacles que l'action de la justice a opposés à leur publication.³¹

Dans les années 1860, Moreau a lui-même complété sa bibliographie en publiant deux suppléments³². Par la suite, plusieurs spécialistes ont suivi ses pas et ont proposé des listes complémentaires, faisant hériter de nouveaux documents du titre. Trente ans après la publication du travail de Moreau, Émile Socard rédige un *Supplément à la Bibliographie des Mazarinades*³³ composé de 102 nouvelles mazarinades conservées à la Bibliothèque de Troyes³⁴. Pensée dans la continuité de la bibliographie de Moreau, sa publication en reprend la mise en page et son mode de structuration de l'information.

26. — Compagnies (Les) de Picqve-Nicqve ov les charmans effects des Bovrgeois de Paris, avx portes de la ville. Leurs Priuileges et Statuts à eux donnez pour leur conseruation. — A Paris, 1652. In-4° de 7 pages.

A la fin on lit : Approuvé et arresté par les Frères Officiers A. B. C. D. E. F. G. H. I. K. L. M. N. O. P. Q. R. S. T. V. X. Y. Z. et autres dont le nombre est infini.

27. — Complimens (Les) de la place Mavbert, reformez par vne des plvs famevses harangeres de Paris. Avec la harangve qv'elle a faite aux Dames de son Exercice, et la Response qu'elles luy ont faite. En vers byrlesques. — S. l. 1650. In-4° de 7 pages.

Pièce en vers de huit syllabes, du style le plus plat. Elle n'a pour elle que la rareté.

FIGURE 1.2 – Extraits de la *Supplément à la Bibliographie des Mazarinades*, p. 15

31. *Ibid.*, tome 1, p. I.

32. Id., “Supplément à la Bibliographie des Mazarinades”, *Bulletin du bibliophile et du bibliothécaire* (, 1862), p. 786-829 ; Id., “Supplément à la Bibliographie des Mazarinades”, *Bulletin du bibliophile et du bibliothécaire* (, 1869), p. 61-81.

33. Émile Socard, *Supplément à la Bibliographie des Mazarinades*, H. Menu, Paris, 1876.

34. Au moment de ses recherches, Émile Socard est conservateur de la bibliothèque

Au début du XX^e siècle, Ernest Labadie³⁵ publie, avec la même organisation, un *Nouveau supplément à la bibliographie des mazarinades*³⁶.

1.2.4 Bibliographies et identification numérique des mazarinades

Lorsque l'on étudie les mazarinades, ces listes s'imposent véritablement comme des références, repères de base fondateurs de l'unité « mazarinades ». Affirmant une cohérence entre ces différents textes, elles sont devenues un moyen de se référer à ces écrits et ainsi de les nommer. Au XIX^e siècle, Armand d'Artois, conservateur de la Bibliothèque Mazarine, catalogue la collection de la bibliothèque à partir de la bibliographie de Moreau.

C'est donc tout naturellement que le projet Antonomaz a lui aussi repris la numérotation des listes pour attribuer un identifiant à chaque fichier traité pour indiquer facilement quelle est la mazarinade contenue. Les bibliographies nous offrent, en quelque sorte, un identifiant déjà tout prêt, pratique pour gérer les fichiers en interne et également compréhensible pour un spécialiste du sujet en dehors de l'équipe.

Aussi, si nous prenons l'exemple de la 13^e mazarinade de la bibliographie de Moreau, *A un ministre d'État sur les œufs*³⁷, l'identifiant attribué est *Moreau13* et tout fichier renvoyant à cette mazarinade (PDF de la numérisation, fichier contenant le texte brut...) contient cet identifiant dans son nom, en plus de son extension.

Si ces listes font autorité, elles ne restreignent pas pour autant le corpus aux textes déjà repérés. Déjà parce qu'elles sont elles-mêmes rédigées dans une logique de synthèse de plusieurs regroupements existants (la première bibliographie de Moreau) ou dans une logique de complément (les suppléments successifs), ensuite parce que, on l'a vu, la définition autant large que floue d'une mazarinade, ne permet pas de borner strictement à ce qui est déjà annoncé comme tel. Juger pertinente l'intégration d'un texte non classé ne serait pas une opération irrationnelle. Aussi le projet entend-il pouvoir permettre cet ajout au sein de l'ensemble numérique et contribuer, à son tour, à enrichir le corpus. Pour le moment, seuls les documents issus des bibliographies « officielles » sont en cours de traitement. D'autres ont été récupérés pour lesquels il faudra penser à un système d'identification cohérent à la fois pour lui-même et par rapport à celui signalé ci-dessus³⁸.

35. note rapide sur Labadie

36. Ernest Labadie, *Nouveau supplément à la bibliographie des mazarinades*, Librairie Henri Leclerc, Paris, 1904.

37. BnF, département Littérature et art, YE-1803, *A un ministre d'État sur les œufs*, 1649.

38. Ce point est précisé dans le chapitre suivant, au sujet de la récupération automatique des textes.

1.3 La production et circulation en masse d'imprimés

La diversité des textes tient aussi au fait qu'une multitude d'imprimeurs-libraires ont entrepris de publier ces brochures. À ce jour, on dénombre plus de 300 imprimeurs-libraires ayant édités des mazarinades.

Les pages de titre forment une source foisonnante d'informations pour saisir cette effusion d'impressions. Le caractère interdit, clandestin d'une grande partie des mazarinades s'illustre particulièrement avec des pages de titre presque vides, réduites à un titre, une ville et une année. D'autres sont très complètes, mentionnant l'auteur, l'éditeur et son lieu d'activité, permettant de savoir où se rendre pour acquérir le texte. La présence de marques d'imprimeurs et de priviléges d'impression (parfois falsifiés) est également indicatrice des contextes de production.

La période de la Fronde connaît une dynamique nouvelle pour le monde de l'édition avec une intensité des impressions remarquable. On couvre les événements, on publie sur l'opinion publique. Christian Jouhaud parle de « politisation générale de l'imprimé³⁹ ».

Cette intensité de parution définit le rapport au document et leur donne une cohérence si on les considère comme ensemble révélateur de la couverture éditoriale des événements de la période. Quoiqu'éphémères, ces documents ont un lien matériel et thématique. Aussi se sont-ils très bien vendus, chez les imprimeurs-libraires déjà, qui tiennent boutique en plus de leurs ateliers, chez les libraires et dans les rues, au détour du Pont-Neuf notamment et auprès des colporteurs.

Le succès des mazarinades se mesure également à leur circulation. Des textes parisiens connaissent un succès en province. Certains sont réédités plusieurs fois, d'autres sont copiés. Des textes font évènement. C'est le cas du discours d'Omer Tallon en 1648 qui circule largement dans le Royaume.

1.4 La conservation des documents

1.4.1 Aspects matériels

Le format d'impression

Si l'on observe une diversité des contenus, le format des mazarinades est assez classique : la grande majorité sont des in-4°, format économique pour l'organisation du texte et adapté à une lecture rapide après achat.

Bien que ce format de feuilles volantes soit largement majoritaire (98% des do-

39. C. Jouhaud, *Mazarinades. La Fronde des mots...*, p. 26.

cuments selon Hubert Carrier⁴⁰), on enregistre aussi des placards avec gravure, format in-folio, parmi lesquels le célèbre *salut de la France dans les armes de la ville de Paris*⁴¹ ou encore des in-8°⁴².

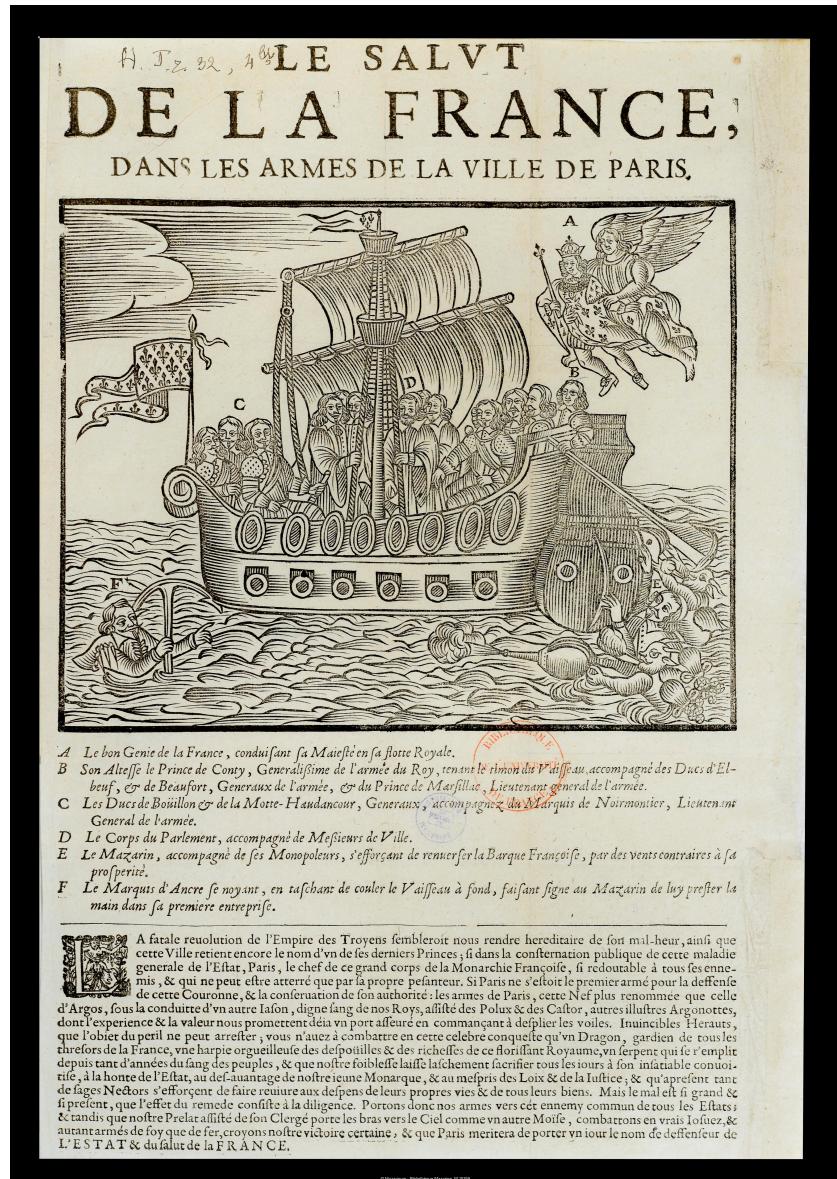


FIGURE 1.3 – *Le salut de la France dans les armes de la ville de Paris*, Paris, 1649.

40. H. Carrier, *La presse de la Fronde (1648-1653) : les Mazarinades. Tome 2. Les hommes du livre....*, p. 192.

41. *Le salut de la France dans les armes de la ville de Paris*, Paris, 1649. Source de l'image : <https://mazarinum.bibliotheque-mazarine.fr/ark:/61562/mz3592>

42. Par exemple, *La Bernarde, comedie.*, Dijon, J. Thilbaut, 1651. <https://mazarinum.bibliotheque-mazarine.fr/records/item/15206-la-bernarde-comedie>

Cependant, ces formats, bien que remarquables et attrayants sont plutôt de l'ordre de l'atypique pour notre corpus d'imprimés, constitué de sorte de brochures devant facilement circuler plutôt que d'affiches à placarder⁴³.

Imprimées dans l'urgence, devant se vendre et se lire rapidement, les mazarinades sont des documents plutôt courts : 72% des imprimés comportent 8 pages ou moins, 20% font entre 9 et 16 pages⁴⁴. De plus, la corporation des colporteurs, acteur majeur de la diffusion de ces papiers, ne possède pas l'autorisation de transporter des documents de plus de 64 pages.

Le format de conservation

Le format commun encourage le regroupement et la conservation en lot. Aussi, ces tracts, brochures, au caractère volant et dispersable ont fréquemment été conservés dans un format englobant : le recueil⁴⁵.

Si l'on consulte une mazarinade aux archives, il est probable de se retrouver face à un livre au sein duquel elle constitue un chapitre, une section artificielle née d'un regroupement, d'une reliure. En effet, dès la période de leur parution, les brochures ont été récoltées et réunies. Elles ont par la suite suscité la naissance d'un véritable marché.

Aussi, les mazarinades ont-elles été conservées en recueils, souvent annuels. En revanche, elles ont été numérisées et mises en ligne individuellement. La conservation en recueil est pourtant révélatrice d'une forme de lien conçu et compris : elle doit proposer une forme de cohérence entre des textes pourtant publiés séparément. Le numérique fait donc perdre cette couche d'informations. Recréer du lien entre les mazarinades est un objectif affirmé du projet Antonomaz :

Ces brochures ont été l'objet, par le passé, de nombreuses formes de recueils et de reliures : on souhaite proposer, avec les moyens numériques, de les *relier* autrement et virtuellement.⁴⁶

43. En plus d'avoir été produits dans une proportion moindre, les placards ont été peu conservés, souvent arrachés par la police. On dénombre une trentaine de placards conservés pour une existence certaine d'au moins le double. Voir Karine Abiven, *Les formats des mazarinades*, disponible à l'adresse https://antonomaz.huma-num.fr/exist/apps/Antonomaz/notices/forme-livre/formats_libelles.xml (visité le 12/09/2022)

44. *Ibid.*

45. Ce format de conservation n'est pas non plus automatique, les mazarinades sont aussi conservés en « simple lot ». Il possible que les institutions possèdent les deux alternatives. Voir par exemple à la bibliothèque de Bordeaux pour un recueil : https://bibliotheque.bordeaux.fr/notice?id=p%3A%3Ausmarcdef_479371&locale=fr et pour un lot : https://bibliotheque.bordeaux.fr/notice?id=p%3A%3Ausmarcdef_487495&locale=fr.

46. K. Abiven, G. Lejeune, A. Bartz, *et al.*, “Vers une collection numérique des libelles parus pendant la fronde, ou comment relier des mazarinades”...

1.4.2 Lieux de conservation des mazarinades

Des exemplaires mondialement dispersés

Pour rassembler les mazarinades, il faut pouvoir les localiser. Si les mazarinades ont déjà largement circulé au temps de la Fronde, à l'échelle nationale et parfois internationale, cette dispersion est d'autant plus criante lorsque l'on en vient à s'intéresser à leurs lieux actuels de conservation. Bien sûr, plusieurs centres d'archives ou bibliothèques françaises possèdent grand nombre d'exemplaires mais il est possible d'en trouver à consulter jusqu'au Japon. On saisit alors bien la potentielle difficulté de dépouillement que l'éparpillement peut soulever pour un chercheur.

En effet, tandis que certaines institutions procèdent à un catalogage à la pièce, d'autres les référencent en lot, rendant alors moins évident le repérage des pièces rares. Certaines les numérisent en partie, d'autres ne l'entreprendront pas pour le moment. Tout cela complique l'estimation du nombre d'exemplaires conservés dans le monde, d'autant plus que, nous l'avons, rien n'interdit finalement un nouvel ajout de documents.

Institutions conservant des exemplaires utilisés pour le projet

Le projet Antonomaz forme sa matière de travail à partir d'exemplaires numérisés, glanés sur le web. Pour cela, trois principaux foyers de mazarinades numériques ont été sélectionnés :

- La Bibliothèque Mazarine *via* sa bibliothèque numérique Mazarinum⁴⁷. 623 numérisations ont ainsi pu être récupérées. L'institution propose également une bibliographie en ligne des mazarinades issues de ses 25 000 exemplaires intégrés dans les collections depuis le XIX^e. collections⁴⁸. Cette ressource, composée de notices individuelles sur les mazarinades, est un apport de connaissances scientifiques précieux autour des textes⁴⁹.
- La Bibliothèque nationale de France qui diffuse ses collections sur sa bibliothèque numérique Gallica⁵⁰. Riche de plus d'un million de livres et de manuscrits numérisés⁵¹, elle a permis à ce jour de récupérer 460 mazarinades numériques. Ce chiffre ne comprend que des exemplaires conservés à la Bibliothèque nationale

47. <https://mazarinum.bibliotheque-mazarine.fr>

48. <https://mazarinades.bibliotheque-mazarine.fr>

49. Voir à titre d'exemple : <https://mazarinum.bibliotheque-mazarine.fr/records/item/17618-arrest-d-absolution-de-me-guy-joly-conseiller-du-roy-au-chastelet-avec-la-permission-de-contoffset=1#page>

50. Depuis sa mise en ligne à la fin des années 1990, Gallica répond à une politique d'accessibilité des collections. Le rythme soutenu des numérisations au fil des ans permet chaque année, la mise en ligne d'une masse considérable de documents. Voir : <https://gallica.bnf.fr/edit/und/a-propos> (visité le 12/09/2022)

51. <https://gallica.bnf.fr/GallicaEnChiffres> (visité le 12/09/2022)

de France et non tout autre ressource pouvant être moissonnée par Gallica (telle Mazarinum).

- Les autres institutions à mentionner s'articulent autour de la bibliothèque numérique Google Books qui a numérisé leurs collections, enrichissant le projet de 1 503 mazarinades :
 - La plus grosse partie est issue de la Bibliothèque municipale de Lyon qui mène une grande campagne de numérisation de ses collections⁵².
 - La British Library, à Londres, qui possèdent plusieurs milliers de mazarinades.
 - La Bibliothèque nationale d'Autriche, à Vienne.
 - La Bibliothèque de Bavière, à Munich.
 - La Bibliothèque centrale de Florence.
 - Dans une proportion moindre, l'Université de Californie, l'Université de Gand ou encore la Bibliothèque nationale de République Tchèque.

Les chiffres indiqués renvoient ici aux documents réellement intégrés dans le corpus traité par le projet : lorsqu'une mazarinade est récupérée plusieurs fois (parce que plusieurs institutions ont fait numériser leur exemplaire), seule une unité est conservée, en priorité celle de la Bibliothèque Mazarine puis de la Bibliothèque nationale de France et enfin de Google Books.

Ainsi, à partir de ces trois bibliothèques numériques, les exemplaires d'une petite dizaine d'institutions sont récoltés et seront connectés sur un site. Au mois d'août, le projet dispose de près de 2 500 mazarinades soit la moitié du corpus complet.

Autres institutions conservant des mazarinades

D'autres institutions possèdent des collections de mazarinades, parmi les plus notables, nous pouvons citer :

- la bibliothèque Sainte-Geneviève,
- la Bibliothèque interuniversitaire de la Sorbonne,
- les archives départementales de la Gironde⁵³ et la bibliothèque de Bordeaux,
- la bibliothèque de l'université de Tokyo,
- la bibliothèque de Saint-Pétersbourg.

A terme, il pourrait être envisageable de chercher à compléter les données du projet

52. *Le projet de numérisation de la Bibliothèque municipale de Lyon et ses partenaires*, URL : <https://numelyo.bm-lyon.fr/projet.php> (visité le 10/10/2022), « L'appel d'offres lancé par la Ville de Lyon pour la numérisation de ses livres imprimés anciens, libres de droits, a été remporté en 2008 par la société Google. La numérisation a commencé fin 2009. Quelque 400 000 ouvrages libres de droits sont en cours de numérisation. »

53. Collection de documents imprimés relatifs à la Fronde, 1649-1651, 4 J 126 à 136.

à partir de ces lieux de conservation qui, pour la plupart, mènent également des campagnes de numérisation.

Maintenant que nous avons donné des précisions de contexte et de compréhension des documents qui soulèvent le besoin de pouvoir lier ces textes, nous pouvons nous atteler à exposer la problématique numérique liée à notre souhait de collection numérique en commençant par faire un état des lieux de la chaîne de traitement à notre arrivée.

Chapitre 2

Préparer et nettoyer les (méta)données par le traitement de masse

La problématique numérique principale du projet Antonomaz est de pouvoir mettre en place une chaîne de traitement s'étalant du repérage des ressources disponibles sur le web (les mazarinades numérisées) à la constitution d'un site internet propre au projet, exposant le travail réalisé sur les documents et les rendant accessibles. Les données issues de la récupération des documents numériques¹ supposent une phase importante de nettoyage et d'alignement pour proposer un ensemble cohérent et propre.

Au moment de notre stage, la chaîne de traitement était déjà entamée : nous présenterons donc ce qui a déjà été mis en place avant notre arrivée afin de mieux comprendre dans quel cadre s'insère notre réalisation.

2.1 Repérer et collecter les documents numérisés

2.1.1 S'adapter à la matière

La volonté de rassembler un ensemble de documents dispersés géographiquement n'est pas une entreprise unique ou inédite dans le monde des humanités numériques. C'est souvent l'objectif des éditions électroniques de correspondance par exemple, puisque fréquemment, les lettres de l'émetteur sont conservées en « lot » selon un lieu de réception

1. « Aujourd'hui, il est intéressant de noter que l'expression "document numérique" renvoie moins à la production originale de l'information dans un environnement informatique en réseau qu'à l'idée de dématérialisation d'un support physique, qui en conserverait les principales caractéristiques informationnelles. » E. Bermès et F. Martin, "Le concept de collection numérique"...

ou un destinataire².

Le repérage des fonds s'effectue par la fouille des inventaires d'archives avant consultation. Cependant, nos mazarinades, imprimées en masse, sont loin d'être des documents uniques : on compte en réalité des millions d'impressions pour nos 5 000 entités textuelles. Or, l'objectif n'est pas de recenser tous les exemplaires mais de proposer la lecture du texte de la mazarinade. Nous avons donc besoin d'un exemplaire par texte.

Si l'on sait que la Bibliothèque Mazarine ou encore la Bibliothèque nationale de France conservent quantité de nos documents, vérifier manuellement l'existence numérique de chaque mazarinade dans leurs collections pour compléter ensuite avec une fouille des sites d'autres institutions serait une entreprise sans fin.

Le projet Antonomaz a donc pris la décision de récupérer automatiquement la matière, avec les avantages et les loupés que cela suppose. Dans cette procédure de constitution du corpus réside une véritable originalité puisque, en général, les projets démarrent avec une source ou un ensemble de sources déjà bien localisées ou identifiées.

2.1.2 Procédure mise en place

La récupération de notre future collection de mazarinades est pensée comme un glanage : on fouille les ressources disponibles et on sélectionne ce qui correspond à nos critères. La procédure est développée dans un article de l'équipe, publié autour de la nécessité de créer du lien entre les mazarinades³. Nous proposons ici d'en reprendre le fil.

Le point de départ, base de notre connaissance des documents identifiés comme mazarinades sont les bibliographies, celle de Moreau en tête car la plus fournie (plus de 4 000 entités). Cependant, lorsque Célestin Moreau mentionne le document, il ne précise pas son lieu de consultation. Aussi, une liste de titres, issus des bibliographies a été formée pour pouvoir lancer des requêtes sur les bibliothèques numériques. En faisant cela, on conditionne ce qui est cherché d'après la liste. Pour requêter Gallica par exemple, l'existence du mot-clé « mazarinades » a permis de cibler plus facilement.

Le fouille de Google Books est particulièrement intéressante en terme de localisation des mazarinades : la bibliothèque numérique agrège de nombreux documents que la société a numérisé pour des institutions. Le requêtage par titre fait émerger celles possédant des mazarinades sans avoir à se pencher sur leurs catalogues individuels.

Ce récolelement automatique a permis d'identifier plus de la moitié des mazarinades comme étant disponibles sur le web avec 2 569 réponses uniques. Pour chacune d'elles,

2. Voir par exemple le projet d'édition de la correspondance de Marc-Michel Rey dont le corpus est conservé dans plusieurs institutions, notamment à Amsterdam et à Genève : <http://rey.huma-num.fr/presentation> (visité le 08/09/2022)

3. K. Abiven, G. Lejeune, A. Bartz, *et al.*, “Vers une collection numérique des libelles parus pendant la fronde, ou comment relier des mazarinades”...

le PDF de la numérisation est récupéré. Le projet prévoit prochainement une nouvelle opération de glane à partir de nouveaux titres pouvant être intégrés au corpus.

2.1.3 Bilan

Ce fonctionnement par requêtage de titres a été perçu comme extrêmement long et coûteux. En effet, comme la machine ne filtre pas d'elle-même ce qui est récupéré, une mazarinade déjà identifiée dans une bibliothèque est à nouveau récupérée. Aussi, pour obtenir nos 2 569 entités de mazarinades, 2 970 éditions ont été collectées. Près de 400 doublons doivent être triés.

Cependant, cela a permis dans certains cas de délaisser une numérisation à la qualité médiocre pour une autre plus appréciable. Certains textes sont en effet peu lisibles à cause de l'altération du papier qui, on l'a évoqué, était de basse qualité. Certaines numérisations de Google ne sont pas optimales, laissant une impression de traitement en chaîne et à la « va-vite »⁴. Des exemples sont disponibles dans l'annexe B.

Autre point à noter sur cette expérimentation : les requêtes ont renvoyé des résultats qualifiés de « faux positifs ». De plus, elles étaient lancées d'après une liste diplomatisée du XIX^e siècle, là où les titres étaient parfois transcrit comme ceux du XVII^e siècle. Il a donc fallu harmoniser le tout pour relancer l'opération.

En revanche, ces résultats, bien que parfois erronés, ont aussi permis de faire émerger des titres non catalogués dans les bibliographies mais qui finalement, après jugement de leur contenu, pourraient en faire partie. Ces documents ont été conservés comme « mazarinades sans identifiants », en attente de la fin du traitement de ceux inscrits sur les listes officielles.

2.1.4 Nommer les fichiers

Avec plus de 2 500 entités identifiées à l'automne 2021, il fallait un moyen clair de se référer à la mazarinade contenu dans le fichier. Nous avons déjà évoqué l'identification liée à la bibliographie d'origine, celle-ci est complétée par la mention du lieu de glane. Selon sa bibliothèque d'origine, un fichier peut ainsi être nommé :

- Moreau3316_GBOOKS
- Moreau3316_GALL
- Moreau3316_MAIZ

“Moreau” correspond donc à l’individu ayant identifié le document comme une mazarinade. On peut également trouver un fichier débutant par ”Carrier”, ”Labadie” ou ”Socard”. ”3316” correspond au numéro attribué par le bibliographe (3316^e de la liste).

4. Voir notamment le chapitre 3 à ce sujet.

Si le document est un supplément Moreau, il se nomme ainsi : *Moreau1suppl12_GALL*, "1suppl" correspondant au premier supplément.

Aussi, notre PDF du *Moreau3316*, émanant de Google Books, se nomme *Moreau3316_GBOOKS.pdf*. Ce système d'identification est très pratique, puisque pour tout fichier créé autour de la mazarinade, il peut être repris avec la précision du nouveau format : *Moreau3316_GBOOKS.xml*. Voyons de suite à quoi ce fichier renvoie.

2.2 La TEI et le premier objectif scientifique du projet : analyser automatiquement le texte

2.2.1 L'océrisation des textes

Rassembler les mazarinades ne vise pas seulement à proposer un ensemble de numérisations mais aussi à permettre leur étude textuelle. Gaël Lejeune travaille particulièrement sur la linguistique computationnelle et Karine Abiven s'intéresse à la construction du genre burlesque et du discours d'actualité dans les mazarinades. Extraire le texte de son image est utile pour leurs recherches respectives et, dans le même temps, produit une matière supplémentaire à mettre à disposition des chercheurs.

Le texte sera alors proposé sur la page de consultation de la mazarinade, ce qui est pratique pour en récupérer un extrait ou sa totalité sans avoir à le recopier (si la qualité en est satisfaisante) mais également procéder à la reconnaissance d'entités nommées. Les études sur la langue par le traitement informatique ainsi qu'un enrichissement de la matière proposé sur le futur site web motivent donc cette récupération.

Définition

Un processus d'OCR (*Optical Character Recognition*) a été appliqué pour obtenir un texte brut issu des pages. Un logiciel a effectué une reconnaissance de caractères sur les images.

Pour en quelque sorte, « reconnaître » les caractères imprimés, un apprentissage machine (*machine learning*)⁵ est nécessaire, par la constitution de données d'entraînement ou l'utilisation d'un modèle déjà entraîné, proche du type de la source à transcrire. Antonomaz a utilisé un modèle existant, entraînés sur un ensemble d'imprimés du XVII^e siècle⁶.

5. L'apprentissage machine est part intégrante des études en intelligence artificielle. Il consiste à réussir à faire « apprendre » à l'ordinateur à partir d'un ensemble de données.

6. Simon Gabay, Thibaut Clérice et Christian Reul, *OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more)*, 2020.

Tous les documents, même ceux provenant de Gallica qui propose fréquemment un texte océrisé, ont été (à nouveau) océrisé. En effet, le texte de Gallica résulte de plusieurs campagnes d'océrisation, rendant impossible la comparaison de deux sorties OCR⁷.

Résultats

La qualité du texte obtenu par l'OCR est variable⁸. Rappelons aussi que nos mazarinades sont imprimées sur des papiers de faible qualité par souci d'économie. La transparence de certaines pages par exemple, entâche largement la possibilité d'une bonne reconnaissance.

Les deux exemples (voir page 30) montrent bien l'écart de qualité obtenue. Le second illustre la conséquence de la faible qualité du papier tandis que le premier illustre un résultat satisfaisant, peu bruité⁹.

Pour la suite, l'équipe songe à revenir sur l'OCR obtenu pour passer cette fois, les images dans le logiciel eScriptorium qui utilise la technique de l'HTR (*Handwritten text recognition*)¹⁰.

2.2.2 Où stocker le résultat ?

Le texte obtenu est stocké dans un fichier écrit dans un langage structuré, le langage XML (*eXtensible Markup Language*) qui permet de baliser et d'encoder un texte. Crée en 1996, ce langage est à la fois lisible de la machine et de l'humain, par un système de contenant et de conteneur : un élément contient du texte ou un autre élément. On structure par l'imbrication.

La TEI (*Text Encoding Initiative*) reprend cette structuration XML pour proposer un ensemble de recommandations pensées pour l'édition numérique de sources historiques¹¹. La TEI désigne à la fois une technique d'encodage des textes développée depuis

7. K. Abiven, G. Lejeune, A. Bartz, *et al.*, “Vers une collection numérique des libelles parus pendant la fronde, ou comment relier des mazarinades”...

8. Voir notamment l'article suivant : K. Abiven, J.B. Tanguy et G. Lejeune, “Exploiter un corpus de données textuelles sans post-traitement : l'écriture burlesque de la Fronde”, *Humanités numériques*–4 (déc. 2021), DOI : 10.4000/revuehn.2355.

9. Jean-Baptise Tanguy, ayant réalisé l'OCR de nos données, propose d'ailleurs dans son travail de thèse une étude sur les corpus bruités par la reconnaissance de caractères. Voir <https://jbtanguy.github.io>

10. Relativement proche et souvent confondu avec l'OCR, on pourrait distinguer les deux techniques en expliquant que l'OCR fonctionne uniquement par une reconnaissance de caractère en caractère tandis que l'HTR traite généralement une ligne par une ligne.

11. « *The TEI Guidelines for Electronic Text Encoding and Interchange define and document a markup language for representing the structural, renditional, and conceptual features of texts. They focus (though not exclusively) on the encoding of documents in the humanities and social sciences, and in particular on the representation of primary source materials for research and analysis.*

Voir : *TEI Guidelines*, URL : <https://tei-c.org/guidelines/>.

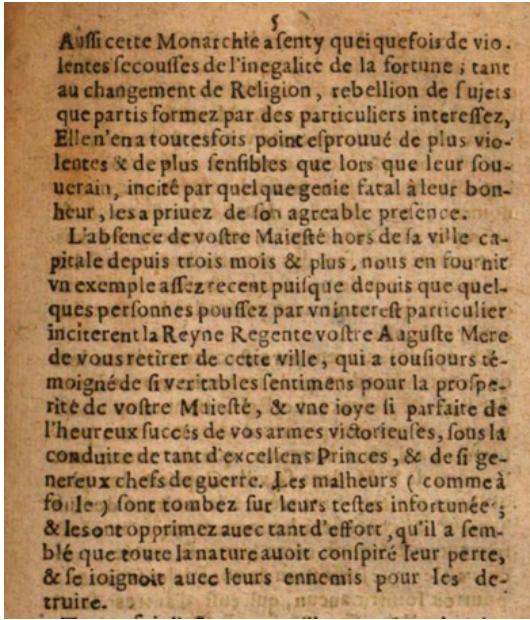


FIGURE 2.1 – Extrait de la numérisation et de l'océrisation du Moreau3707

<lb>Auffi cette onarchie a fenty quaquefois de vio-
<lb>lentes fecoufes de l'inegalite de la fortune; tant
<lb>au changement de Religion, rebellion de fujes
<lb>quepartis formez par des particuliers intereffez,
<lb>Elleu'en a toutesfois point eprouué de plus vio-
<lb>lentes & de plus sensibles que lors que leur sou-
<lb>uerain, incité par quelque genie fatal à leur bon-
<lb>heur, les a priuez de son agreeable prefence.
<lb>L'abfeace de vostre Maiefte hors de fa ville ca-
<lb>paitale deguis crois mois & plus. nous en fournis-
<lb>n exaple afezrecent puifque depuis que quel-
<lb>ques perfones pouuez par vnintereft particulier
<lb>inciterent la Reyne Regente vostre Auguſte Mere
<lb>de vous retirer. de cette ville, qui a toufiours té-
<lb>pigé de fi veriables fentimens pour la prospé-
<lb>rity de vostre Maifeé, & vne ioye fi parfaite de
<lb>l'heureux fuceés de vos armes victorieufes, fous la
<lb>coaduite de tant d'excelfns Princes, & de fi ge-
<lb>nereut chefs de guerre. Lea malheurs (comme à
<lb>foule; font tabez fur leurs teftes infortunées;
<lb>& le sont opprineaz avec tant d'effort, qu'il a fem-
<lb>blé que toute la nature auoit confpiré leur perte,
<lb>& fe ioignoit avec leurs ennemis pour les de-
<lb>truire.

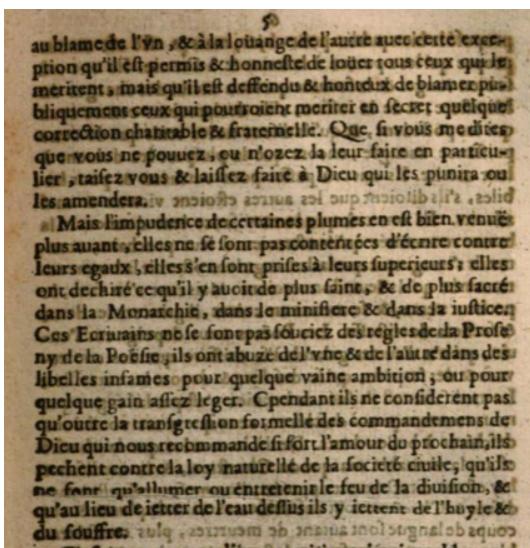


FIGURE 2.2 – Extrait de la numérisation et de l'océrisation du Moreau3780

<lb>aaldaaßa B oàdaloudgede laoanhbqaoide eaqa
<lb>pionql epaais banualde las youar ph
<lb>ariagt odabquflbat dalaadlaaûobde blaiopu
<lb>tqeasaoqaiaihat aadai fbet ainq8
<lb>araoûaaaibolaôaaaisi: Qe, osgadip
<lb>qareasouap q ionea da leuraien paatûeu-j
<lb> aana vilauieuie iu quis lts puaim,
<lb>lâaua:
<lb>doapimdaeoauanan
<lb>plhauatellanaaf fonpaoauapedéretas
<lb>lmesaggade ,aitlesiefan pnifeslàalemp fuperuess llan
<lb>at dedié'eqpail uoindpds. ain dophraan-
<lb>daas la: Mamaia, daodlmianibaieilaad iulo4
<lb>Cas'Eraiooble aapasuacie diblojloodedo'anf
<lb>rdela ofials ot ahaudl'fadet' alhaéldaide-
<lb>&nl les infanbes puulqud, vaisa ambiû ;ar pue
<lb>qgalqae gai affeai leger. Cpeilanils ne cofiderepasl
<lb>qioae da taofgeefiaoforellddab éoumapadcmad
<lb>Biaqaidbapmananideafedrl auabouodu péoalaa.10
<lb>padhaeaaroda leg naaellbde aacidé oiuite;lquie
<lb>aa nr aqolanaernantâüninlefeue la diuifit,!
<lb>eaolauaïde y i de roy
<lb>dhaafa.

plus de 30 ans et la communauté d'utilisateurs investis dans les réflexions autour de l'édition numérique¹². Un ensemble de balises est prédéfini pour pouvoir décrire sémantiquement et scientifiquement le texte. Le tout est englobé dans une sorte de schéma général, structurant l'utilisation de ces balises afin que l'édition réalisée à partir de ces recommandations soit uniforme.

Le format XML-TEI contient deux éléments intéressants pour le projet Antonomaz :

- l'élément **teiHeader** où peuvent être renseignées les métadonnées renvoyant à l'océrisation du texte, à la source de la numérisation, aux informations de publication du document physique... Autrement dit, c'est la gestion de la source qui est développée à cet endroit.
- l'élément **body** qui contient le texte et donc le résultat de l'océrisation, avec un balisage respectant la mise en forme du document physique.

Aussi, le fichier XML-TEI devient une sorte de résumé de la mazarinade numérique : il référence le lien vers la numérisation, une description bibliographique de l'exemplaire traité, des indications de classification du texte (mots-clefs renvoyant au genre, à la thématique abordée, note descriptive...), une description de l'objectif du projet, une mention des personnes ayant participé à ce travail, et le texte dont l'océrisation représente une forme d'édition... Un exemple d'encodage est disponible en annexe (page 99).

2.3 Assurer la qualité des métadonnées

2.3.1 Générer semi-automatiquement les fichiers

Chaque texte de mazarinade et ses métadonnées sont donc stockés dans un fichier XML-TEI individuel. Pour obtenir ce résultat, la création des fichiers a été semi-automatique : un script a construit un fichier pré-rempli avec la structure générale des balises et l'ajout de certaines métadonnées : le lien vers la numérisation, les informations extraites des bibliographiques comme le titre, la date et le lieu de publication, le nombre de pages ou encore le lien vers une notice de la Bibliothèque Mazarine correspondant à la mazarinade... Le texte océrisé est aussi ajouté avec un marquage des changements de ligne et un balisage des lignes.

Les choix d'encodage ne poursuivent pas le but de « proposer une édition fine des textes en XML-TEI, mais bien de disposer d'un encodage minimal du texte qui réponde

12. « Cependant, au cours de la dernière décennie, il est apparu de plus en plus clairement que la TEI fait partie de ce qui rend possible les humanités numériques : elle est devenue une partie intégrante de l'infrastructure à laquelle tout le monde a affaire, techniquement et socialement, dès que l'on commence à réfléchir sur le texte ou sur d'autres formes de ressources culturelles sous forme numérique. » Lou Burnard, “Conclusion : qu'est-ce que la TEI ?”, dans *Qu'est-ce que la Text Encoding Initiative ?*, OpenEdition Press, Marseille, 2015, URL : <http://books.openedition.org/oep/1305> (visité le 28/08/2022).

aux ambitions du projet : disposer de métadonnées riches et d'une version océrisée du texte que l'on peut exploiter¹³».

Aussi, après l'enregistrement des fichiers, chacun est relu ou plutôt complété par un enrichissement de métadonnées. Il convient de renseigner l'auteur et l'imprimeur, et de renseigner leurs identifiants *idref* ou *viaf* afin de désambiguïser l'information et de l'aligner sur des référentiels¹⁴. Les notices de la Bibliothèque Mazarine sont porteuses d'analyses intéressantes, notamment des commentaires d'Hubert Carrier qui a proposé des datations précises pour beaucoup de brochures. La date ou l'intervalle de dates est alors ajoutée. Des mots-clefs sur la thématique du texte peuvent être renseignés, de même que le genre littéraire dans lequel il s'inscrit.

Cependant, tous ces ajouts manuels sont sources d'erreurs ou d'oublis, il faut penser à les limiter au maximum par une manipulation technique.

2.3.2 Comment s'assurer de la qualité de l'encodage ?

L'encodage est régi par un schéma de validation, vérifiant que tous les fichiers sont conformes et balisés selon le cadre défini. Autrement dit, le projet a fait des choix d'encodage en sélectionnant des balises et un ordre d'apparition. Le respect de tout cela est vérifié par ce schéma qui informe des erreurs lorsqu'un élément n'est pas conforme.

Un document permet de maintenir la cohérence de l'encodage : c'est l'ODD (*One Document Does it all*). Il comprend un schéma RELAX NG qui détermine les balises utilisables, leur ordre et contexte d'utilisation, leur répétition ou non, quels attributs ou valeurs peuvent être attachés aux éléments. L'ODD est aussi le fichier contenant une documentation venant justifier les choix d'encodage¹⁵.

Aussi, l'une des premières missions de notre stage a été de relire l'ODD déjà rédigé pour proposer de nouvelles règles vérifiant la qualité des informations saisies et limitant les erreurs ou oublis de saisie manuelle.

Par exemple, l'extrait de code ci-dessous vérifie que, si un lien vers une notice de la Bibliothèque Mazarine est renseigné, le numéro de notice doit être ajouté. Tant que cela n'est pas fait, un message reporte l'erreur et en informe le lecteur.

13. K. Abiven, G. Lejeune, A. Bartz, *et al.*, “Vers une collection numérique des libelles parus pendant la fronde, ou comment relier des mazarinades” ...

14. Ces précisions permettront par la suite, de créer un lien simple entre les documents et notamment, de permettre une recherche des textes par auteur ou imprimeur.

15. «Un projet scientifique d'édition numérique est défini par la qualité de sa documentation (...). En fait, une édition qui n'expose pas sa question de recherche et ne déclare pas ses critères de numérisation et de gestion des sources, n'est pas une entreprise scientifique. » Ioana Galleron, Marie-Luce Demonet, Cécile Meynard, Fatiha Idmhand, Elena Pierazzo, Geoffrey Williams, Pierre-Yves Buard et Julia Rogeri, *Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations*, rapp. tech., 2018, URL : <https://halshs.archives-ouvertes.fr/halshs-01932519/document> (visité le 23/08/2022), p. 10.

```

<constraintSpec ident="subtype" scheme="schematron">
    <constraint>
        <sch:rule context="tei:relatedItem[@target/starts-with(., 'https://mazarinades.bibliotheque-mazarine.fr/ark')]">
            <sch:report test=".[@subtype='none']">Si le @target contient un lien vers une notice de la Bibliothèque Mazarine, l'identifiant de la notice doit être indiquée dans le @subtype. Le @subtype ne peut pas être 'none'.</sch:report>
        </sch:rule>
    </constraint>
</constraintSpec>

```

Nous ne développerons pas tout ce travail mais insisterons plutôt sur l'intérêt de produire un schéma de validation le plus contraignant possible¹⁶.

2.3.3 Pourquoi contraindre ?

Dans notre cas, ce n'est pas tant la qualité de l'encodage du texte (celui n'a en effet pas vocation à être modifié après la génération des fichiers) que celle des métadonnées qui nous a occupé. Définies comme « un ensemble structuré d'informations permettant de décrire la ressource, de la classer, de l'organiser et de caractériser des données ou du contenu¹⁷ », les métadonnées identifient et donnent un contexte au document.

Aussi, la mazarinade ne se considère pas seulement par son texte mais prend également sens avec les informations de description (bibliographiques, matérielles...) qui lui sont liées. De même, le contexte de diffusion du document doit être clair : des informations techniques sur la production du fichier, son objectif scientifique et ses conditions de réutilisation sont indispensables. Ces deux derniers éléments sont produits automatiquement par l'encodage automatique, ce sont surtout les informations liées au document qui nous viennent à être précisées.

La qualité des métadonnées construit la qualité scientifique du projet puisqu'elle produit une forme d'analyse de la source mise à disposition. Toute personne ouvrant le fichier XML peut alors identifier l'origine du texte édité.

Les métadonnées du **teiHeader** seront largement utiles pour la mise en ligne des documents, si ce n'est fondamentales. L'outil utilisé pour construire le site permet de

16. L'ODD complet est disponible avec sa documentation dans les livrables ou à l'adresse suivante : <https://github.com/Antonomaz/ODD> (visité le 15/09/2022).

17. *Ibid.*, p. 7.

piocher des éléments directement dans les documents XML-TEI et d'y appliquer un rendu web. Elles seront donc exposées à tout visiteur du site et par conséquent, doivent être rigoureusement saisies. Toute faute de frappe n'empêche souvent pas la compréhension mais perturbe la qualité de lecture.

En interne, ce travail de métadonnées dans les fichiers XML-TEI est également exploitable et réutilisable. Nous verrons dans le chapitre 4 comment nous avons intégré automatiquement ces données dans des fichiers liés à la diffusion des images. Aussi, les informations saisies doivent-elles être propres pour ne pas reproduire des métadonnées erronées. De même, une métadonnée non renseignée dans le fichier TEI conduit à une perte d'information qui se reproduit lors d'une réutilisation ou de l'affichage.

Nous pouvons à présent donner un exemple de leur intégration dans notre chaîne de travail en nous penchant sur un autre enjeu de propriété dans notre ensemble de données : celui des images.

Deuxième partie

**Réalisation technique autour du
traitement de l'image : l'opportunité
IIIF**

Chapitre 3

Le nettoyage des PDFs des numérisations Google BOOKS

Les PDFs récupérés nous offrent une qualité variable de numérisations. Si celles effectuées par la Bibliothèque Mazarine et la Bibliothèque nationale de France sont utilisables en l'état, celles de Google Books questionnent sur l'application d'un traitement avant leur diffusion sur le site. En effet, bien que la plupart soient propres, une partie d'entre elles a de nombreuses pages inutilement répétées, parfois une vingtaine de fois (voir en annexe page 103). Ces doublons se constatent lorsque la page a été numérisée plusieurs fois et que chaque image obtenue a été conservée.

3.1 Pourquoi nettoyer ?

Penser à un moyen de nettoyage des PDFs signifie rajouter une étape dans la chaîne de traitement pour des données qui, à la fin de la procédure, ne seront pas enrichies mais plus propres. L'objectif n'est là, pas de tenter de compléter certains PDFs aux pages de titre manquantes, mais au contraire d'élaguer les pages en surplus.

Si l'on considère l'utilisation du PDF du point de vue d'un chercheur, la répétition inutile de pages rend certes, la consultation moins confortable, mais n'empêche pas l'accès au contenu littéraire. Cependant, fournir au lecteur un document au déroulement cohérent est déjà un point à noter en faveur de l'ajout d'une étape de traitement. Le document numérique doit pouvoir correspondre le plus fidèlement à la version physique qu'il doit reproduire.

Du point de vue technique, pensons au coût de stockage inutile que peut entraîner la conservation de pages doublonnées au sein du corpus. Si les procédures sont automatisées et que ces pages supplémentaires ne provoquent pas un temps de traitement conséquent pour l'avancée du projet, leur existence est loin d'être nécessaire et les conserver serait

même une attitude contraire à la vigilance portée à la pollution environnementale générée par l'informatique.

L'omniprésence du numérique dans notre quotidien est, nous le savons, source de pollution, responsable d'une émission carbone¹. Aussi, faut-il garder à l'esprit que cette problématique concerne largement les humanités numériques dont les projets nécessitent un matériel au coût énergétique certain². Optimiser le stockage des données relève d'une bonne pratique : ne pas héberger sur le serveur une image en double, laisse de la place pour une image inédite.

Un nettoyage s'envisage donc comme une étape non superflue de notre chaîne de traitement.

3.2 Comment nettoyer ?

3.2.1 La librairie python imagededup

Pour essayer de nettoyer nos PDFs, nous avions deux options : travailler avec la répétition du texte ou la répétition de l'image. Gaël Lejeune avait rédigé un script vérifiant les répétitions, dans l'OCR, des chaînes de caractères, permettant d'isoler le PDF si certaines pages avaient un texte commun³. Si cette technique a l'avantage d'être très fiable pour repérer les pages de texte répétées, elle ne pouvait pas prendre en compte la répétition des pages vides de caractères. Aussi, nous avons opter pour travailler avec un outil de comparaison d'images.

Présentation générale

*imagededup*⁴ est une librairie python⁵ conçue pour repérer automatiquement des images dupliquées dans un lot d'images défini. Elle procède, par un algorithme de *hashing* ou un réseau de neurones convolutifs (CNN), à une comparaison d'images pour évaluer un taux de ressemblance.

L'utilisation de la librairie est relativement facile d'appropriation : on choisit son outil de comparaison, un dossier d'images à traiter sur laquelle on applique une méthode

1. Le numérique serait à l'origine de 4% des émissions carbone mondiale. Voir : *La face cachée du numérique*, nov. 2019, URL : <https://librairie.ademe.fr/cadic/2351/guide-pratique-face-cachee-numerique.pdf?modal=false> (visité le 24/08/2022), p. 4.

2. Voir notamment le manifeste suivant, écrit pour sensibiliser à ce sujet : *Manifeste des Digital humanities...*

3. Le script est disponible ici : https://github.com/Antonomaz/tools/blob/master/check_txt.py (visité le 12/09/2022).

4. *Imagededup*, URL : <https://github.com/idealo/imagededup> (visité le 02/09/2022).

5. En programmation, une librairie est un ensemble de fonctions, de modules contenant des outils de travail autour d'un thème.

*find_duplicates*⁶, à laquelle on peut préciser un degré minimal de ressemblance recherché.

Nous n'avons pas pu trouver de projets ayant eu recours à un dispositif de ce genre pour des sources semblables (numérisations de documents anciens). Si l'outil fonctionne très bien pour retrouver des images de type photographique qui seraient stockées plusieurs fois dans un même dossier, il n'a pas été pensé pour comparer des pages de texte. Aussi, ce développement rendra compte de l'expérimentation effectuée.

Le traitement par CNN

imagededup intègre une méthode de comparaison par CNN (*Convolutional Neural Networks*). Un CNN est un ensemble de neurones⁷ artificiels, inspiré de la vision animale.

Cette technique est beaucoup utilisé pour le *deep learning*, forme d'apprentissage machine fonctionnant avec un apprentissage par couche de neurones. La technologie HTR a recours à ce procédé pour comparer les caractères. Les traitements automatiques des langues s'en emparent aussi.

C'est la méthode pour laquelle nous avons penché dans notre expérience de dé-doublonnage d'images, afin de comparer les taux de ressemblance. Un autre méthode⁸, le traitement par *hash*, est proposée par la librairie : la fonction attribue un identifiant numérique sous forme de chaîne de caractères à un ensemble défini qui fait office d'empreinte. C'est une méthode fréquemment utilisée pour enregistrer les mots de passe afin d'éviter de les conserver en clair. Ces empreintes peuvent ensuite être comparées pour établir un niveau de ressemblance. Cependant, cette technique n'a donné pour notre part, aucun résultat concluant.

3.2.2 Logique du script

La librairie prend en entrée des images individuelles : il faut donc séparer les pages des PDFs et les convertir en PNG, format recommandé dans la documentation. Pour chaque PDF, le script crée un dossier où sont glissées les images fraîchement découpées. Chaque image est nommée d'après son numéro de page de PDF. Puis, on donne en entrée ce dossier pour traitement par réseau (le CNN). Les images au fort taux de ressemblance sont sélectionnées et, pour chaque duo, la seconde image est supprimée. On supprime

6. https://idealo.github.io/imagededup/user_guide/finding_duplicates/ (visité le 11/09/2022)

7. Un neurone est à considérer comme une unité artificielle elle-même connectée à d'autres neurones, capable de réaliser des opérations. Cf Edouard Oyallon, *Analyzing and Introducing Structures in Deep Convolutional Neural Networks*, thèse de doct., Paris Sciences et Lettres, 2017, URL : <https://hal.archives-ouvertes.fr/tel-02353134>, p. 2.

8. Les différentes méthodes sont disponibles ici : <https://idealo.github.io/imagededup/> (visité le 12/09/2022)

toujours l'image au numéro de page (de PDF) le plus élevé puisque dans la quasi-totalité des cas, l'image est répétée après son positionnement correct⁹.

Si la comparaison des pages ne repère pas de pages similaires, le PDF d'origine est conservé. Sinon, un nouveau PDF est créé à partir des images PNG restantes dans le dossier. L'ensemble se résume avec le schéma ci-dessous :

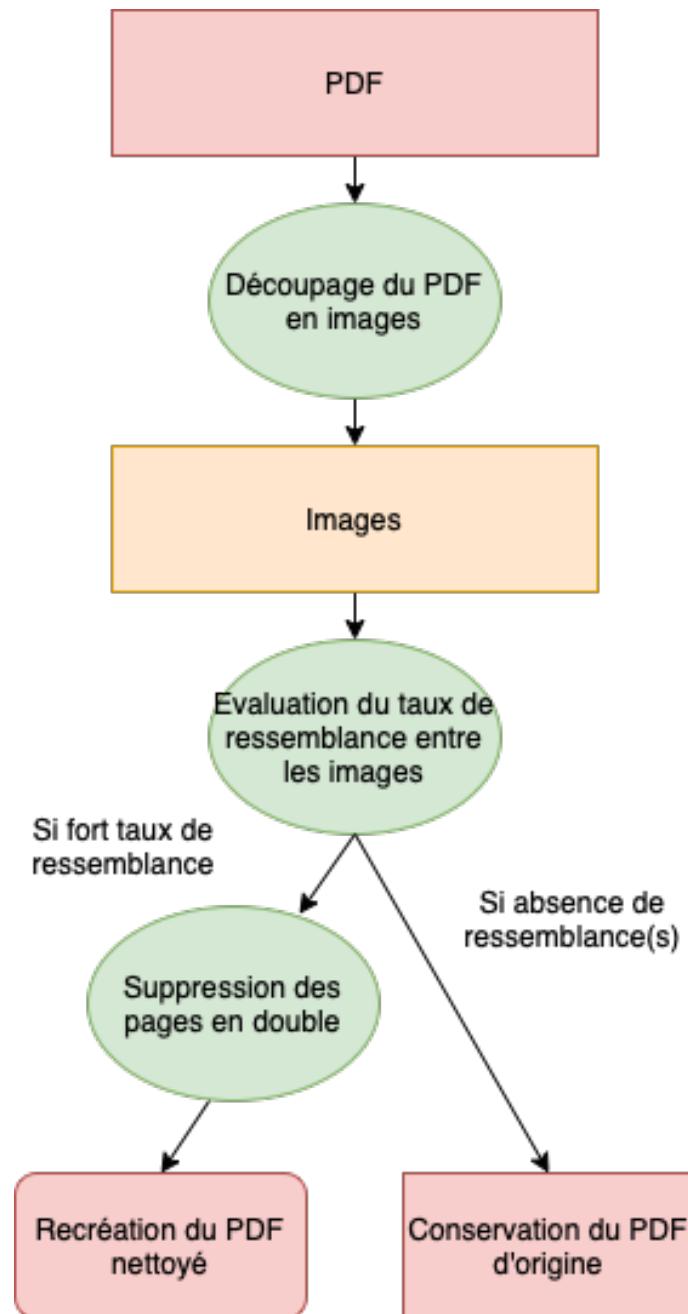


FIGURE 3.1 – Schéma du traitement opéré sur l'image

9. Nous n'avons croisé qu'un seul PDF où une page de texte doublonnée était dupliquée entre la page de titre et la première page de texte.

3.3 Premiers résultats sur le traitement de numérisations d'imprimés anciens

3.3.1 Adapter le script aux résultats : le découpage en zones

La traduction technique du déroulement du script a supposé d'adapter certaines étapes, notamment pour optimiser les résultats de comparaison.

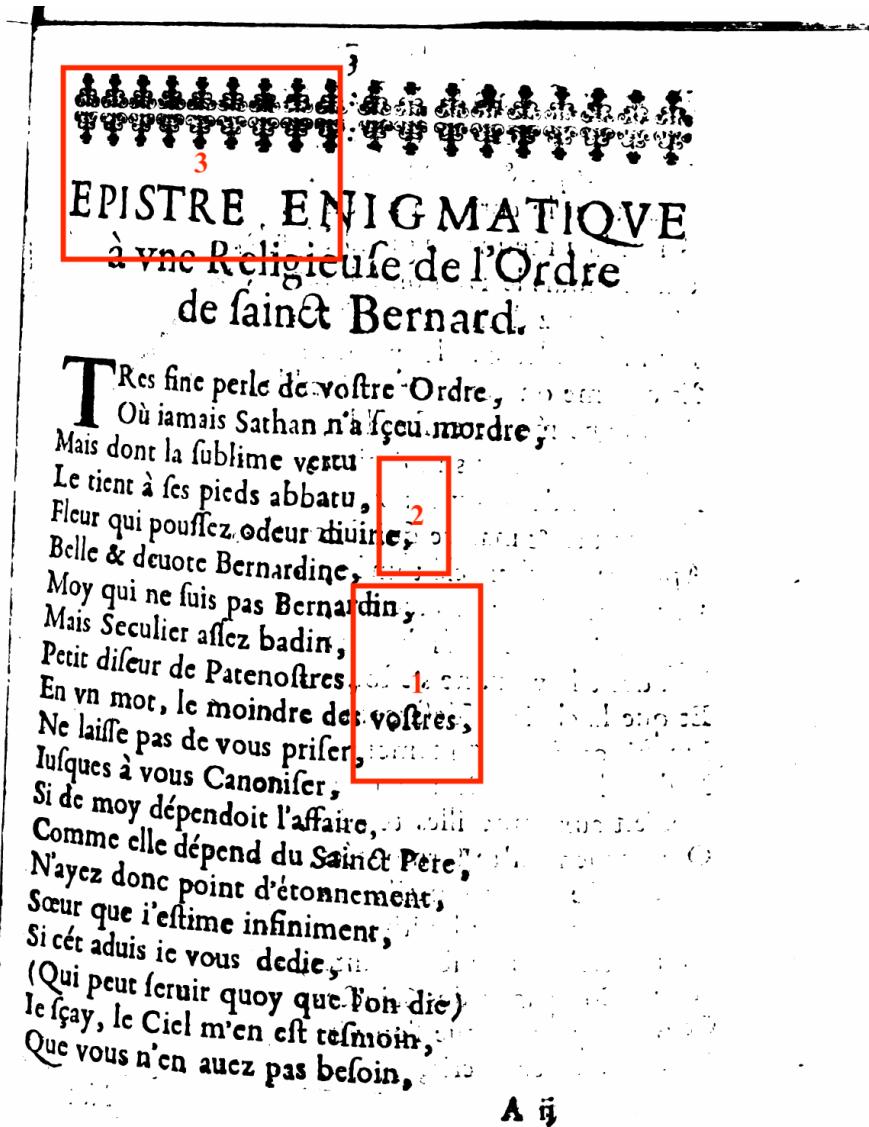
Pour être comparées, les images sont réduites à une taille commune. Cependant, en prenant en considération l'entièreté de la page, l'encodage des images est moins performant : les pages sont vite perçues comme similaires, résumées à des lignes noires sur fond blanc. Pourtant, la détection de similarité, pour être efficace, ne doit pas se faire au niveau de la mise en page extrêmement commune (tant par document qu'à l'échelle du corpus) mais des caractères imprimés.

Nous avons donc pensé à sélectionner seulement une partie de l'image à comparer afin de conserver au maximum les divergences dans les lignes de texte. Le CNN reçoit alors une zone sélectionnée qu'il réduit pareillement que s'il avait reçu l'image entière et traite donc un ensemble de pixels plus nuancé et moins « flou ».

Chaque page est comparée aux autres pages par l'intermédiaire de trois zones. Autrement dit, chaque page est comparée trois fois mais à trois endroits différents. La zone 1 sera donc comparée avec la zone 1 des autres images, etc. Si au moins deux zones font ressortir une forte ressemblance avec une autre page, nous considérons que la page est dupliquée. Comme, nous allons le voir, le degré de similarité pour déterminer si l'image est doublonnée ou non est très sensible. Vérifier le seuil sur deux ou trois zones est une barrière de sécurité limitant les risques de mauvais traitement.

Pour définir la taille des zones, nous avons testé plusieurs formes (en carré ou en bande) de plusieurs dimensions. Les résultats les plus optimaux sont obtenus lorsque les zones sont des zones rectangulaires, sélectionnant une partie assez précise du document. Nous avons également essayé de les placer de manière stratégique afin de sélectionner des endroits « clefs » du document : en évitant les zones vides pour les pages de texte, en positionnant une zone de manière à comparer les fins de ligne pour les textes en prose...

Aussi, les trois zones traitées pour chaque page sont situées aux endroits suivants :



Digitized by Google

FIGURE 3.2 – Sélection des zones sur une page de document

3.3.2 Trouver le bon degré de similarité

Outre le traitement en zone, il a fallu calibrer le bon degré de similarité indiquant si les pages doivent être considérées comme identiques. Pour préciser le degré souhaité, il suffit de le préciser dans la fonction `find_duplicates()`, avec l'argument `min_similarity_threshold=`. Ce qui donne par exemple pour comparer les zones 1 :

```
duplicates_zone1 = cnn.find_duplicates(encoding_map=
    encodings_zone1, min_similarity_threshold=[degré de
```

similarité])

La similarité minimale se précise par un chiffre compris entre 0 et 1, 0 traduisant l'absence totale de ressemblance et 1 des images complètement identiques.

Avec nos données, nous ne pouvons appliquer un tel partage puisque nos images ont de nombreux pixels en commun. Le taux de similarité entre deux images peut atteindre des scores élevés, avoisinant les 0.99 mais le score de deux images de pages différentes se situe souvent entre 0.90 et 0.95. Le risque est donc de supprimer des pages uniques au cours de la procédure, d'où la comparaison sur plusieurs zones pour limiter cette possibilité.

La difficulté dans le repérage des pages de document répétées réside dans le fait que les images ne sont pas strictement identiques : une même numérisation n'est pas enregistrée deux fois mais c'est plutôt la page de document qui a été numérisée deux fois. Nous avons donc deux (ou plusieurs) numérisations de la même page. Numériquement, ce ne sont pas les mêmes images mais c'est pourtant ce que nous devons faire comprendre à la machine.

Le script nous a permis de nettoyer 400 PDFs (sur 2 010 en entrée). Cependant, si notre méthode a suffit pour le nettoyage de notre collection de PDFs, plusieurs points négatifs ou défauts sont à souligner.

3.4 Limites de la procédure

Bien que l'objectif de dédoublonnage ait été atteint, la technique utilisée présente des limites et des défauts qui n'ont pas pu être améliorés dans le cadre du stage.

1. Le coût de la procédure est assez lourd : le découpage du PDF en images avec la librairie *pdf2image* est long, ce qu'une autre librairie comme *imagemagick* aurait peut-être fait plus légèrement.
2. La librairie *imagededup* comporte d'importantes difficultés d'installation. Si nous avons pu l'utiliser sur notre machine, la tentative d'installation sur le serveur est restée vaine, avec des messages d'erreur signalant l'incompatibilité des versions de certains paquets à télécharger. Cette situation se répète ou non en fonction des machines, sans que nous ayons pu en comprendre la raison technique. Cela limite pourtant les possibilités de réutilisation du script.
3. Si les numérisations d'une même page de texte sont trop éloignées (numérisation de la page plus inclinée par exemple), le script a tendance à ne pas les signaler comme doublons.

4. En revanche, sur les documents où le texte est très concentré, on remarque une tendance à évaluer trop largement les pages qui seraient dupliquées, ce qui cause des suppressions erronées.
5. Les PDFs des documents issus de la bibliothèque de Vienne sont fréquemment traités « inutilement » du fait de la présence de pages blanches en début et fin de document.

Les dernières remarques ont souligné le besoin d'une vérification manuelle des PDFs nettoyés par le script : déjà pour vérifier que les pages en surplus avaient toutes été supprimées mais surtout pour s'assurer que des pages « uniques » n'avaient pas été éliminées, causant un trou dans le document.

Cette démarche nous a permis de corriger certaines erreurs dans le nommage des fichiers comme le cas de 2 PDFs au même numéro Moreau mais contenant pourtant deux textes différents.

Maintenant que la totalité de nos numérisations Google Books ont été vérifiées, nous pouvons nous pencher sur l'étape suivante qui consiste à aligner leur hébergement sur un standard de diffusion d'images pensé pour répondre aux besoins d'interopérabilité qui dynamisent les humanités numériques.

Chapitre 4

Le standard IIIF pour l'hébergement des images

4.1 Standards et intéropérabilité de l'hébergement image

4.1.1 Qu'est-ce que le IIIF ?

Le *International Image Interoperability Framework*¹ (ou IIIF) est formé d'un ensemble de standards proposant des directives pour la mise en ligne et la valorisation d'images, de leur hébergement à leur visualisation. Il tend à mettre à disposition des images selon un processus technique commun, soucieux d'alléger les coûts de stockage des fichiers (en ce sens, plus respectueux du facteur environnemental) et imposant un ensemble de pratiques communes autour du partage d'images sur le web, utiles à la fois pour les développeurs et les chercheurs. Dans notre cas, les dites images sont plutôt des numérisations de documents physiques anciens mais une telle logique fonctionne également avec les sources nativement numériques.

Comme l'écrit Régis Robineau, l'objectif de IIIF est de « créer un cadre technique commun grâce auquel les bibliothèques numériques peuvent délivrer leurs contenus de manière standardisée sur le Web afin de les rendre consultables, manipulables et annotables par n'importe quelle application ou logiciel compatible.² » Autrement dit, chaque possesseur d'images, s'il suit le standard IIIF, permet, à toute application d'accéder aux fichiers hébergés sur un serveur. En se connectant à ce contenu, il est alors possible de

1. *International Image Interoperability Framework*, URL : <https://iiif.io> (visité le 12/08/2022).

2. Régis Robineau, *Comprendre IIIF et l'intéropérabilité des bibliothèques numériques*, nov. 2016, URL : <https://insula.univ-lille.fr/2016/11/08/comprendre-iiif-interoperabilite-bibliotheques-numeriques/> (visité le 13/08/2022).

réutiliser ces images sans les ré-héberger puisque leur chargement émane directement du serveur émetteur et court-circuite ainsi la nécessite de stocker à nouveau le fichier pour le proposer sur un site web différent de celui d'origine.

Deux avantages percent la toile : une entité (institution patrimoniale, projet d'édition numérique...) souhaitant récupérer l'image pour la diffuser ne se heurte ni à la question trublionne des droits de réutilisation (puisque tout contenu proposé avec le standard IIIF reste propriété de son hébergeur), ni à la problématique de l'espace de stockage inhérent à tout projet numérique. D'où qu'elle soit visualisée, l'image est chargée depuis ce serveur distant, ce qui incombe bien sûr la possibilité de se pourvoir d'un tel logiciel de service.

Aussi, le maître mot de ce standard est l'intéropérabilité, c'est-à-dire la capacité à fournir un cadre de partage et de diffusion. Dans notre cas, ce sont des images sur le web. Une telle possibilité s'exécute grâce aux API (*Application Programming Interface*), services web permettant le requêtage par une application³. La version la plus récente des API (version 3) permet maintenant de prendre en charge les fichiers audios et vidéos.

Ce requêtage d'images par l'API offre l'avantage de pouvoir personnaliser l'affichage de l'image disponible. Le serveur héberge l'image dans son entièreté, telle qu'elle a été déposée initialement. Par la suite, cela ne restreint pas sa réutilisation à ce seul état : par une URL (*Uniform Resource Locator*), il est possible de manipuler l'image. Peuvent alors être obtenus un cadrage sur une zone particulière, un pivot à un degré défini, une colorisation en noir et blanc ou encore un affichage à une taille réduite. Ce jeu sur les critères d'affichage permet obtenir une visualisation dynamique, prête à servir les centres d'intérêts d'un utilisateur, capable de proposer uniquement une portion de l'image sans la tirer définitivement de son berceau. Tout cela se gère par des précisions dans l'URL ci-dessous :

`http(s) ://{domaine}/{identifiant}/{zone}/{taille}/{rotation}/{qualité}.{format}`

Prenons un exemple pour expliciter cette syntaxe qui définit l'API Image de IIIF. L'image IIIF contenant la page de titre de la mazarinade *Requeste presentee au roy Pluton par Conchino Conchini*⁴ est hébergée par la Bibliothèque nationale de France à l'adresse suivante :

3. Les API permettent de diffuser les données de manière automatique. Une API est « parfois considérée comme un contrat entre un fournisseur d'informations et un utilisateur d'informations, qui permet de définir le contenu demandé au consommateur (l'appel) et le contenu demandé au producteur (la réponse). » *Une API, qu'est-ce que c'est ?, URL : https://www.redhat.com/fr/topics/api/what-is-a-rest-api* (visité le 11/09/2022).

4. *Requeste presentee au roy Pluton par Conchino Conchini, contre Mazarin & les Partisans*, Paris, 1649.

[https://gallica.bnf.fr/iiif/ark:
/12148/bpt6k855568v/f5/full/full/0/native.jpg](https://gallica.bnf.fr/iiif/ark:/12148/bpt6k855568v/f5/full/full/0/native.jpg)

Cette URL affiche l'image originale : la zone et la taille sont *full*, la rotation est nulle et la qualité *native*. Imaginons que nous souhaitons faire apparaître sur notre site internet, seulement le titre du document, en noir et blanc. Cela se définit aussi au niveau de l'URL :

[https://gallica.bnf.fr/iiif/ark:
/12148/bpt6k855568v/f5/100,100,3000,1500/full/0/grey.jpg](https://gallica.bnf.fr/iiif/ark:/12148/bpt6k855568v/f5/100,100,3000,1500/full/0/grey.jpg)

Pour ces deux URL, l'affichage de l'image sera le suivant :

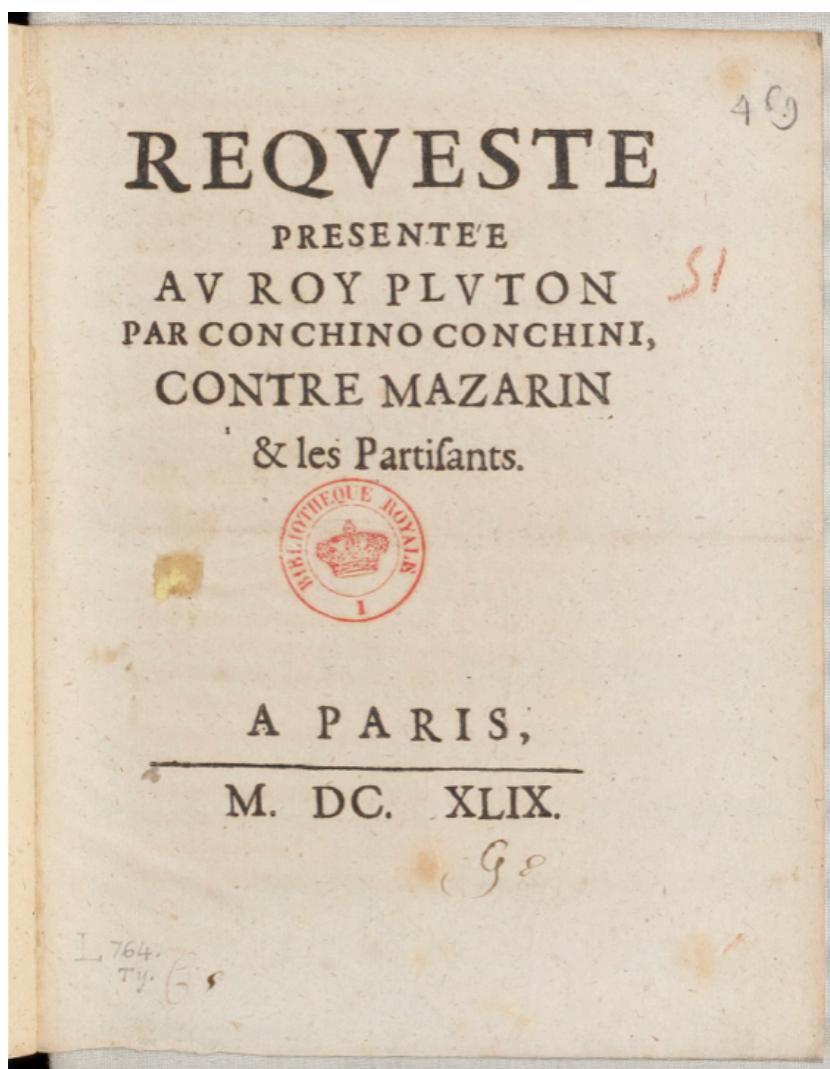


FIGURE 4.1 – Image IIIF affichée dans son entiereté

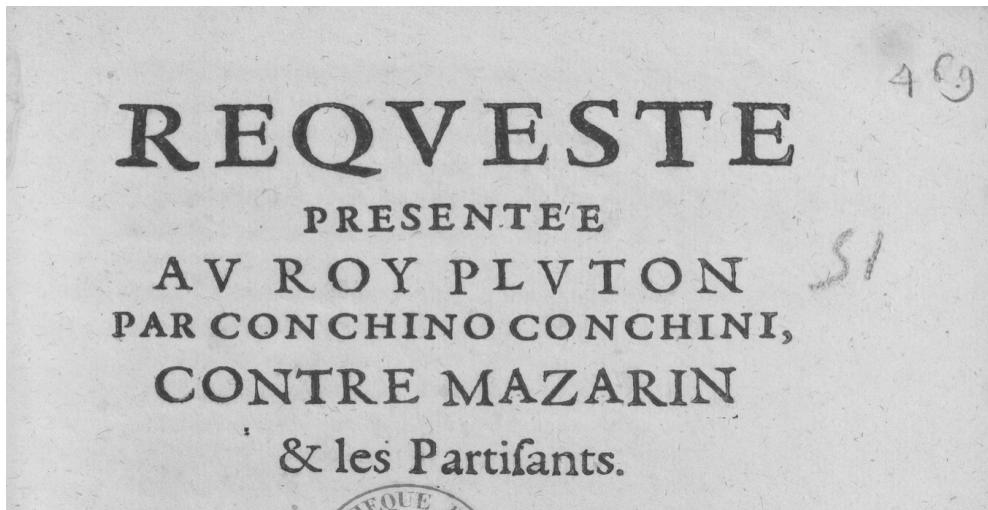


FIGURE 4.2 – Image IIIF zoomée sur un point d'intérêt

L'autre API intéressante, du point de vue d'un chercheur ou ingénieur, pour la diffusion des images est l'API Présentation qui s'utilise pour « délivrer de manière standardisée des informations de présentation et de structure d'un objet numérique⁵ ». Autrement dit, cette API gère les métadonnées présentes à l'échelle d'une image ou d'une suite d'images⁶.

Présentons à présent le fonctionnement technique de IIIF.

4.1.2 Présentation de la chaîne technique de IIIF

Si l'on comprend bien que l'intérêt du IIIF est de contribuer à la diffusion d'images sur le web en les stockant individuellement, il est également pensé pour concevoir des « corpus » d'images liées par une interprétation humaine traduite, techniquement dans un fichier numérique. Autrement dit, à l'échelle de la machine, toute image se résume à un fichier, ensemble de pixels dont la forme se conserve. L'étape de la numérisation aboutit à ce stade : les pages du document perdent leur reliure et chacune d'entre elles devient un fichier image. Mais, à l'échelle humaine, ces unités doivent pouvoir être proposées comme un ensemble cohérent : un lecteur d'un document numérisé souhaite pouvoir accéder à celui-ci dans sa forme numérique complète et non fouiller pour reconstituer lui-même sa suite.

Jusqu'à présent, nous avons évoqué l'hébergement image en IIIF. Celui-ci s'intègre dans une structuration plus vaste, le manifeste, unité constituant véritablement ce qui peut être appelé « objet IIIF ». En effet, contenant à la fois la ou les images à afficher et les métadonnées de celles-ci (titre, date de création, taille de l'image, auteur...), il est l'unité de base utilisée pour visualiser. Une image ne se visionne pas en récupérant son

5. Id., *Comprendre IIIF et l'interopérabilité des bibliothèques numériques...*

6. Nous insisterons particulièrement sur cet aspect dans la partie suivante.

URL mais depuis un manifeste qui lui, se connecte au serveur l'hébergeant pour la charger tout en affichant à l'écran les informations de contexte sur cette image (métadonnées) qui lui sont attachées. Aussi, chaque image est contenue dans un canevas, qui forme son cadre de diffusion. Les canevas sont eux, articulés dans une séquence définissant l'ordre d'affichage. Enfin, la séquence s'intègre dans le manifeste qui peut en contenir plusieurs. Des structures plus complexes peuvent être mises en forme, incluant notamment des annotations. Le schéma ci-dessous résume ce que nous venons d'écrire et avons utilisé dans le cadre de ce stage.

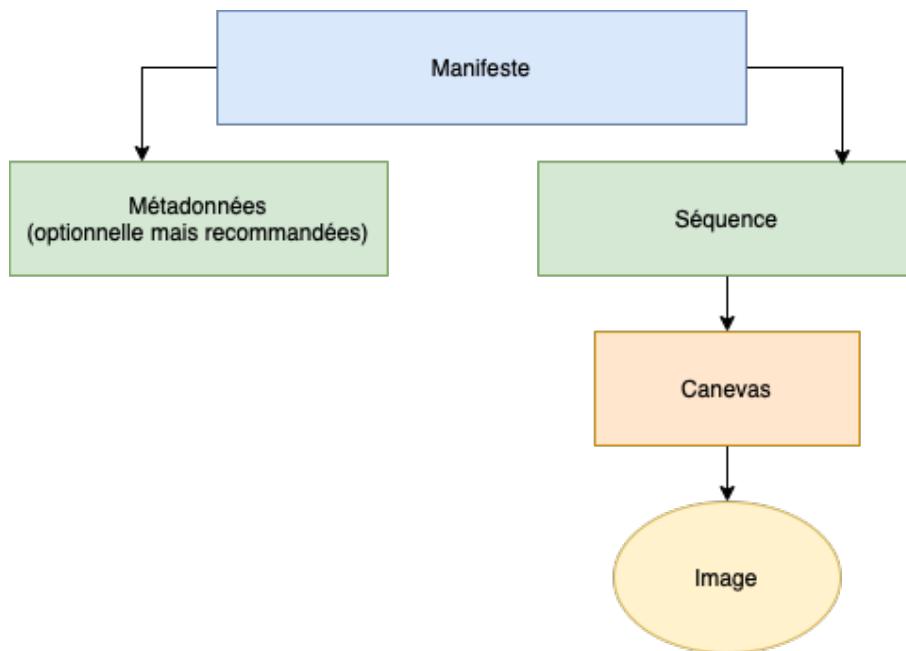


FIGURE 4.3 – Structuration du manifeste IIIF

Cette structure technique est pensée pour être adaptée au mieux à la nature des documents à diffuser, en multipliant ou non les séquences et les canevas. Une photographie (recto seul) peut alors être contenue dans un manifeste au sein d'un unique canevas et d'une unique séquence tandis qu'on pourrait envisager celui d'un livre ancien numérisé, pour respecter sa structure interne, avec la création d'une nouvelle séquence à chaque chapitre, au sein de laquelle chaque page serait un canevas. Dans notre cas, les mazarinades formant, dans l'immense majorité des cas, une brochure de quelques pages, une séquence contenant un canevas par page nous a paru, nous le verrons, le plus approprié.

Cette flexibilité dans la structuration des manifestes façonne tout l'intérêt pour le IIIF qui se positionne comme une solution perméable et adaptable à qui veut diffuser des images, favorisant alors l'adoption de ce standard.

4.1.3 La communauté IIIF

Sur le web, le choix du moyen de diffusion d'un contenu n'est jamais anodin. Choisir de suivre le standard IIIF, c'est s'intégrer dans une communauté scientifique et technique.

Portant le même nom que le standard, la communauté IIIF se forme autour d'un *consortium* d'institutions patrimoniales et universitaires (archives, bibliothèques, musées...). Fondé en 2015, le IIIF-Consortium (IIIF-C) recense 11 membres actifs, parmi lesquels les université d'Oxford et de Stanford, la British Library ou encore la Bibliothèque nationale de France. Aujourd'hui, cette association internationale a séduit plus de 60 membres⁷.

Ce groupe travaille à proposer et améliorer l'offre de service autour du IIIF en faisant évoluer le standard (nouvelles versions d'API), en assurant l'interopérabilité du standard qu'il utilise pour fournir des images, audios ou vidéos et en contribuant au développement technique des outils compatibles. Régulièrement, des sessions d'échanges sont organisées pour échanger autour du standard⁸. En parallèle, des formations en ligne et gratuites sont proposées pour se former techniquement⁹.

En France, plusieurs institutions ont adopté ce standard pour valoriser leurs fonds et collections. Les deux plus grands contributeurs sont, on l'a vu, la Bibliothèque nationale de France et l'infrastructure de recherche et de service Biblissima¹⁰. Progressivement, de nouveaux acteurs s'alignent sur le standard. C'est notamment le cas de la Bibliothèque interuniversitaire de la Sorbonne, dont la bibliothèque numérique Nubis, sans avoir été pensée initialement avec le standard, met à disposition les collections patrimoniales de la bibliothèque¹¹.

L'interopérabilité de IIIF facilite assurément la création d'une communauté travaillant avec des pratiques communes, encourageant en ce sens, l'adoption de ce standard, assurant l'éloignement de l'isolement très rapidement arrivé lors du développement technique. De même, il standardise le rapport à l'image pour le chercheur, en mesure de la récupérer ou de l'utiliser habilement et avec habitude.

Concernant notre corpus, la moitié de celui-ci est déjà diffusée sur le web via IIIF : la Bibliothèque nationale de France ainsi que la Bibliothèque Mazarine suivent le stan-

7. <https://iiif.io/community/consortium/> (visité le 12/09/2022)

8. <https://iiif.io/community/groups/> (visité le 12/09/2022)

9. Indiquons particulièrement celle qui nous a éclairé : *IIIF Online Workshop*, URL : <https://training.iiif.io/iiif-online-workshop/> (visité le 29/08/2022).

10. Constitué de 17 établissements, Biblissima s'attelle à soutenir la recherche sur la transmission des textes anciens, de l'Antiquité à la Renaissance. Voir : *Biblissima*, URL : <https://projet.biblissima.fr> (visité le 10/09/2022).

11. Sur la construction de la bibliothèque numérique Nubis, voir le travail réalisé par Mylène Ravereau. Les images sont hébergées au format JPEG, sans mention de IIIF. Mylène Ravereau, *L'uniformisation et la pérennité des informations numériques dans les bibliothèques numériques : le cas du logiciel libre Omeka*, mémoire sous la direction de Jean-Baptiste Camps, École nationale des chartes, 2016.

dard pour la diffusion de leurs numérisations. En revanche, les mazarinades récupérées de Google Books sont seulement disponibles au format PDF. Si quelques mazarinades numérisées par Google ont été mises en ligne en IIIF par les institutions propriétaires (telles la British Library), rares sont-elles. La Bibliothèque municipale de Lyon, d'où émanent la plupart des PDFs, ne propose pas ces collections en IIIF. Aussi, pour aligner ces numérisations avec le reste du corpus et se mettre en accord avec ce qui vient d'être développé en faveur de l'utilisation du standard IIIF, décision a été prise de réaliser un hébergement IIIF pour les documents concernés.

4.2 Production des manifestes à partir des métadonnées des XML-TEI

L'enjeu principal de cette étape a été, d'un point de vue technique, de proposer une méthode permettant de créer automatiquement les manifestes IIIF lorsque les numérisations émanaient de la bibliothèque numérique de Google. À ce stade, nous disposions des métadonnées contenues dans le **teiHeader** des fichiers XML-TEI et, bien sûr, des numérisations au format PDF.

4.2.1 Où héberger les images ?

Le fondement du manifeste IIIF est de regrouper en un seul fichier une suite d'images IIIF pour leur donner un contexte d'existence et une interprétation. Si l'on choisit le cas d'un manuscrit médiéval, chaque page numérisée est hébergée individuellement et c'est le manifeste IIIF qui a la charge de recréer numériquement le manuscrit, proposant à la visualisation une suite d'images en un ordre défini.

Créer un manifeste IIIF nécessite donc de disposer d'images IIIF, premier point d'intention puisque nos mazarinades sont téléchargées depuis Google Books en documents PDF. Pour cela, plusieurs alternatives existent : par exemple, Internet Archive propose d'héberger des images gratuitement. Si cela peut constituer une solution, cela signifie aussi ne pas avoir la main sur le serveur hébergeur¹². Sinon, pour être autonome, des serveurs d'images dynamiques proposent un hébergement en haute résolution, compatible avec les API IIIF. Au sein de Sorbonne Université, l'unité de service CERES dispose d'un tel outil en ayant installé un serveur Cantaloupe¹³.

Le travail autour du IIIF a donc été réalisée en collaboration avec cette unité qui fait

12. Voir <https://archive.org/create/>. Cependant, plusieurs problèmes d'affichage sont signalés par les utilisateurs de ce service.

13. *Cantaloupe*, URL : <https://cantaloupe-project.github.io> (visité le 17/08/2022).

bénéficier au projet Antonomaz d'un espace serveur pour l'étape dont il est actuellement question. C'est plus particulièrement Thomas Bottini, membre du comité de pilotage de CERES, qui s'est chargé de stocker les images sur le serveur¹⁴. Il faut ici souligner l'intérêt de faire héberger les images par CERES, notamment pour la question de la maintenance du serveur qui est assurée annuellement par l'équipe du service.

Bien que nous n'ayons pas réalisé ce travail avec le serveur pendant le stage, nous pouvons tout de même présenter la mise à disposition avec les images. Chaque PDF a été découpé en un lot d'images avec la librairie *imagemagick*. Déposées sur le serveur, elles reçoivent un identifiant sous la forme d'un URI (*Uniform Resource Identifier*) qui traduit leur emplacement sur le serveur et donc leur chemin d'accès. Prenons l'exemple de la mazarinade intitulée *Les actions de grâces des bourgeois et habitants de la ville de Paris faictes au roy, à la reyne et aux princes après l'heureux retour de Sa Majesté en sa bonne ville de Paris*¹⁵, identifiée comme Moreau29. L'URI de sa page de titre est la suivante : https://ceres.huma-num.fr/iiif/3/antonomaz--antonomaz--Moreau29_GBOOKS-001

- Après le protocole HTTPS, "ceres.huma-num.fr" indique le domaine, ici le serveur human-num de CERES.
- "/iiif/3/" précise que le fichier est situé sur le serveur d'images Cantaloupe, accessible à la version 3 de l'API IIIF.
- Le dernier morceau du chemin personnalise l'identifiant, mentionnant le nom du projet, le numéro Moreau identifiant le document et le numéro de page (pouvant par sécurité atteindre 999, même si, à ce stade, aucun document n'a cette envergure). Cela signifie que l'image est stockée selon le chemin "antonomaz/antonomaz/Moreau29_001".

Comme expliqué précédemment (page 45), l'affichage de l'image originelle est disponible si l'on complète cette URI avec les critères de visualisation : https://ceres.huma-num.fr/iiif/3/antonomaz--Moreau29_GBOOKS-001/full/max/0/default.jpg.

L'entièreté des mazarinades issues de Google Books est donc disponible de cette manière, en modifiant dans l'URI, le numéro Moreau et le numéro de page, correspondant à la place de la page dans le document et non à la pagination du texte (même si cette dernière est très souvent basée sur la première). C'est donc à partir de ces identifiants qu'il faut composer pour recréer la mazarinade.

Si l'on analyse la signification de l'identifiant de l'image hébergée, nous remarquons

14. Au moment de la rédaction de ce mémoire, seuls deux PDFs à valeur de test ont entièrement été hébergés. Cependant, ce que nous présentons ici sera aussi appliqué pour les suivants.

15. *Les actions de grâces des bourgeois et habitants de la ville de Paris faictes au roy, à la reyne et aux princes après l'heureux retour de Sa Majesté en sa bonne ville de Paris*, Paris, 1649.

que celle-ci est parfaitement compréhensible par un humain... ayant travaillé sur le projet ou connaissant le travail de bibliographie entrepris par Moreau. La pratique courante et qui se veut standard, est plutôt d'attribuer des identifiants ARK (*Archival Resource Key*). Créé en 2001, ce système d'identification permet d'obtenir un identifiant unique lié à un objet en dehors de tout contexte d'interprétation. Constitué d'une chaîne de caractères aléatoire, ce numéro d'immatriculation se veut une solution pérenne puisque dénuée de sens.

Une institution désireuse d'adopter ce système peut obtenir un numéro d'autorité nommante (le NAAN : *Name Assigning Authority Number*) la référençant comme entité de gestion des ARK qu'elle attribuera¹⁶. Ce numéro apparaît alors dans l'URI juste avant l'identifiant ARK attribué au document.

ark:/12148/bpt6k107371t

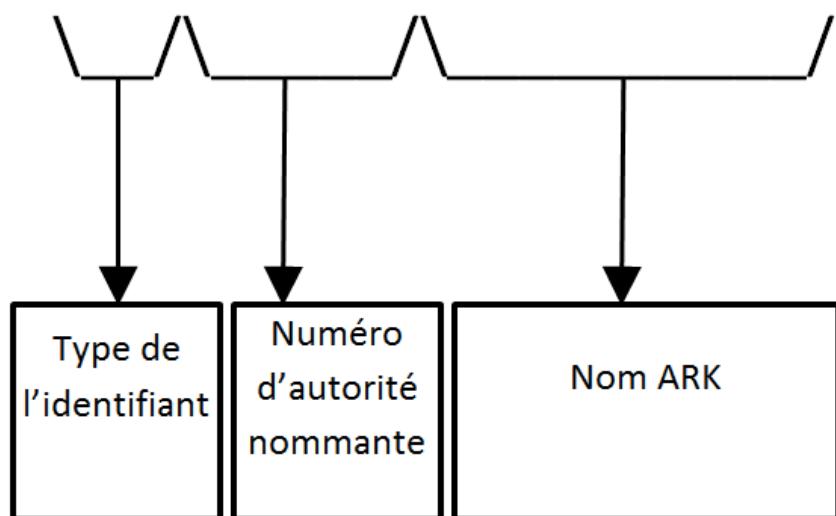


FIGURE 4.4 – Schéma du chemin ARK dans l'URI

Le schéma ci-dessus, tiré du site internet de la Bibliothèque nationale de France¹⁷, s'intéresse uniquement à la composante ARK du chemin d'accès. Dans une URI complète, l'ARK s'intègre ainsi : <https://gallica.bnf.fr/ark:/12148/bpt6k55764575>.

De nombreuses institutions patrimoniales françaises font le choix d'utiliser ce système. Tout l'hébergement IIIF de la Bibliothèque nationale de France consultable sur

16. La demande se fait gratuitement auprès de la communauté ARK : <https://arks.org/about/getting-started-implementing-arks/>.

17. Source du schéma : *L'identifiant ARK (Archival Resource Key)*, URL : <https://www.bnf.fr/fr/lidentifiant-ark-archival-resource-key> (visité le 13/09/2022).

Gallica suit ce principe. La Bibliothèque Mazarine où nous trouvons bonne part du corpus le fait également¹⁸. Des bibliothèques universitaires possédant une bibliothèque numérique suivent aussi la démarche à l'exemple de la Bibliothèque interuniversitaire de la Sorbonne¹⁹. Au delà de l'identification d'une image IIIF, l'identification ARK permet de désigner tout document numérique comme les notices bibliographiques mais aussi des ressources physiques (produits éditoriaux) ou mêmes immatérielles (concepts).

Aussi, il aurait fallu que l'unité de service CERES se déclare comme une entité nommante, ce qui pourrait être un système d'identification intéressant étant donné l'investissement dans le serveur Cantaloupe pour ce service universitaire. Cependant, la décision de nommer les images selon leur numéro Moreau facilite grandement le lien entre images, manifestes et fichier XML-TEI pour chaque mazarinade et elle a notamment été faite dans cette perspective. Ce point est expliqué dans la sous-partie suivante et le chapitre 5. L'identifiant, bien qu'interne au projet, remplit tout à fait son rôle d'URI.

4.2.2 Ce que doit contenir le manifeste

Informations générales

Après l'hébergement des images sur le serveur IIIF, nous pouvons nous pencher sur la question de leur mise en relation, nécessaire pour composer la mazarinade numérique.

Un manifeste IIIF se traduit techniquement par un fichier JSON (*JavaScript Object Notation*), ensemble de données structurées sous forme de « clef/valeur », faciles d'échange et appréciées pour leur compatibilité, très lisibles pour une machine. Ce format est lu par les API IIIF.

```
▼ metadata:  
  ▼ 0:  
    label: "Repository"  
    value: "Bibliothèque nationale de France"  
  ▼ 1:  
    label: "Digitised by"  
    value: "Bibliothèque nationale de France"  
  ▼ 2:  
    label: "Source Images"  
    value: "https://gallica.bnf.fr/ark:/12148/bpt6k855568v"
```

FIGURE 4.5 – Extrait d'un fichier au format JSON

Source : <https://gallica.bnf.fr/iiif/ark:/12148/bpt6k855568v/manifest.json>

18. Par exemple : <https://mazarinum.bibliotheque-mazarine.fr/ark:/61562/mz17668> (visité le 11/09/2022)

19. <https://nubis.univ-paris1.fr> (visité le 11/09/2022)

Nous l'avons déjà évoqué, le manifeste contient à la fois des liens vers les images à afficher et des métadonnées propres à son existence, donnant un contexte d'interprétation aux images. Aussi, celui-ci doit rendre contenir des informations bibliographiques, en mesure de signifier clairement une description de la mazarinade numérisée à laquelle il donne un cadre.

Des informations techniques sont obligatoires pour la bonne validité et la lecture du manifeste par une application tierce :

- le manifeste doit contenir une mention de la version de l'API utilisée, permettant de préciser à la machine comment analyser le fichier.
- il doit obligatoirement posséder un identifiant unique, résultant de son chemin d'accès depuis le serveur et permettant son accès distant. Cela sous-entend donc de réfléchir en amont au lieu de stockage du fichier afin d'éviter toute modification qui peut être coûteuse en temps.
- il compte au moins une image IIIF à renvoyer lorsqu'il est chargé pour visualisation.

Les données supplémentaires favorisent la compréhension humaine du corpus d'images mais ne contribuent pas à sa bonne lecture. Aussi, l'ajout de métadonnées n'est techniquement pas obligatoire mais pourtant nécessaire pour décrire nos documents, sources qu'il faut pouvoir identifier pour le travail d'historien.

La considération des métadonnées

Le champs *metadata* d'un manifeste peut être visualisé, puisqu'il l'illustre, dans le même temps que le contenu image. L'ajout se fait à l'échelle non de l'image mais du corpus d'images « recréant numériquement » le document physique. Que souhaitons-nous faire apparaître au moment de sa consultation ?

Si cette question peut paraître anodine, elle permet de faire surgir une question sous-jacente : comment adapter la description d'un document physique à sa version numérique ? Il faut là penser la mise à disposition d'informations sur la production du document en vue de sa consultation mais également à qualifier la version numérisée.

Tout individu consultant les métadonnées d'un fichier résultant de la numérisation d'un document patrimonial s'attend à trouver des informations archivistiques ou bibliographiques, traduction de sa recherche dans un catalogue ou un inventaire. Aussi, nos manifestes doivent contenir, quand elles sont disponibles, les indications traditionnelles que sont le titre du document imprimé, l'auteur, la date de publication, l'imprimeur, le format... Tout cela permet d'identifier le contenu intellectuel de notre corpus d'images.

Nos mazarinades sont conservées dans des institutions patrimoniales. Les métadonnées doivent donc rendre compte d'indications sur la conservation de l'exemplaire numé-

risé. S’instaure alors une distinction intéressante entre la notion de série et d’exemplaire : la mise en ligne d’une numérisation suppose de considérer la diffusion d’un contenu et de son contenant, résultant de la dématérialisation d’un exemplaire particulier possédé, dans notre cas, par une bibliothèque distincte. Cet élément doit apparaître dans les métadonnées.

De même, une couche supplémentaire d’informations sur l’environnement de création et de conservation des numérisations concernant la forme numérique du document physique se précise. Il tient là de mentionner les acteurs du document numérique : qui numérise ? qui est propriétaire des images ? sous quelle licence sont-elles placées ? Nous devons penser à créer une métadonnée précisant la source initiale des images présentées²⁰.

En cela, la rédaction des métadonnées diffère de l’unique description physique et dépasse le suivi stricte des normes de description bibliographique ou archivistique. En réalité, IIIF permet une grande souplesse à cette partie du manifeste, tout texte peut y être ajouté, sans contrainte de champs. Cependant, traiter ces documents patrimoniaux suppose de se tenir à un certain cadre cohérent de description, liant à la fois un contexte de production physique et numérique.

L’établissement des métadonnées est donc un moment fondamental dans la création de nos manifestes IIIF. Cependant, avec nos 2 000 PDFs Google Books, les rédiger manuellement aurait été trop coûteux en temps. Nous avons donc dû trouver une solution technique pour automatiser leur écriture.

4.2.3 Créer un script python écrivant les manifestes

L’enjeu de cette étape était de pouvoir automatiser par un script, dans une base de fichier JSON, l’écriture des métadonnées et la création d’un canevas à chaque nouvelle page de document pour que, à la fin de la procédure, un manifeste IIIF complet soit enregistré. Nous devions donc composer à la fois avec une partie des données présentes dans les fichiers XML-TEI et sur un serveur distant.

Attribution d’un identifiant au manifeste

L’identifiant du manifeste, son *id*, est indiqué dans le fichier JSON. Comme l’URI de l’image IIIF, il correspond à son emplacement de stockage, mis en accès distant depuis un serveur. C’est une propriété technique du manifeste qui identifie la ressource. Elle est donc primordiale pour accéder au contenu du manifeste.

Son renseignement suppose donc de savoir, au moment de l’écriture des fichiers, où

20. La licence de Google Books autorise la réutilisation des images mises en ligne par la société tant que les informations de provenance sont précisées.

le manifeste sera stocké, sinon il ne sera pas accessible. Si les images sont stockées sur le serveur Cantaloupe de CERES, décision a été prise, afin de limiter les intermédiaires, d'héberger les manifestes sur le serveur huma-num d'Antonomaz, où sera également déposée l'application finale. L'hébergement distinct des images et des manifestes souligne ici un autre avantage indébiable de IIIF : le bon chargement des images est possible sans altération suivant le lieu d'hébergement. Le nommage des manifestes suit le numéro Moreau attribué à la mazarinade, ce qui facilitera, nous le verrons, la connexion entre notre manifeste et le fichier XML-TEI de chaque mazarinade. Sur le serveur, ils sont stockés dans un répertoire *manifests*. L'identifiant pour nos manifestes se construit donc ainsi :

```
"id ":"https://antonomaz.huma-num.fr/manifests/Moreau{numero}.json"
```

Intégration des métadonnées dans le manifeste

Puisqu'un travail conséquent sur les métadonnées est effectué depuis le début du projet, celui-ci nous sert de base évidente pour construire celles des manifestes. Le plus logique est de récupérer celles dans les fichiers XML-TEI des mazarinades puisque relues et complétées pour chaque document. Elles constituent les informations les plus complètes et les plus certaines, où le risque d'erreur ou de bruit est réduit.

Dans un premier temps, il faut donc pouvoir accéder à ces métadonnées XML-TEI, ce que la librairie *lxml* permet. En donnant le chemin XPath²¹ de l'élément à récupérer, nous pouvons le sélectionner et le stocker dans un dictionnaire où il est assigné à une clef, pratique pour être injecté par la suite à l'endroit souhaité dans le manifeste. Prenons par exemple, l'élément **idno** qui précise la cote attribuée par le lieu de conservation du document. Il nous servira d'illustration tout au long de ce développement²².

```
# si la cote est renseignée, on la récupère dans une variable
if doc.xpath("//tei:TEI//tei:msDesc/tei:msIdentifier/tei:idno/text()", namespaces=ns):
    cote = doc.xpath("//tei:TEI//tei:msDesc/tei:msIdentifier/tei:idno/text()", namespaces=ns)[0]

# on la stocke dans un dictionnaire nommé data
```

21. Développé par le W3C, XPath désigne le *XML Path Language*, un langage conçu pour sélectionner les éléments dans un fichier XML. Il permet de requêter le fichier en indiquant un chemin dans l'arbre XML. Voir à ce sujet : https://www.w3schools.com/xml/xpath_intro.asp (visité le 02/09/2022)

22. L'entièreté du script est disponible dans les livrables. Il est ici légèrement adapté pour l'accompagnement du propos

```
data[ "cote" ] = cote
```

Maintenant l'élément stocké, nous pouvons l'injecter dans une métadonnée que nous nommons *Shelfmark* :

```
if 'cote' in data:
    manifeste_json[ "metadata" ].append({
        "label": { "en": [ "Shelfmark" ] },
        "value": { "fr": [ f'{data[ 'cote' ]}' ] }
    })
```

Deux points sont à souligner par rapport à cet extrait de code :

- La métadonnée *Sherlmark* n'est créée dans le manifeste que si la cote est renseignée. Ce *if* permet de gérer les absences de données dans le XML-TEI sans générer une erreur interrompant le script mais également de ne pas créer un champs vide, et donc peu utile.
- L'API 3 de IIIF impose nouvellement la précision de la langue dans laquelle est inscrit le texte. Bien que cela n'ait que peu d'intérêt pour notre cote, cette fonctionnalité est très pratique pour indiquer une information en plusieurs langues : titre original et traduit, description...

Pour le Moreau²⁹ conservé à la Bibliothèque municipale de Lyon, nous possédons cette métadonnée informant de la cote dans son fichier XML-TEI :

```
<idno type="cote">SJ IF 247/172, 44</idno>
```

Nous obtenons ce résultat dans le manifeste :

```
{
    "label": {
        "en": [
            "Shelfmark"
        ]
    },
    "value": {
        "fr": [
            "SJ IF 247/172, 44"
        ]
    }
}
```

Les autres métadonnées ont été fabriquées avec le même procédé. La traduction des métadonnées XML-TEI vers des métadonnées IIIF soulève toutefois un questionnement : au-delà du format de conservation, quelles différences existent-ils entre les métadonnées XML-TEI et IIIF ? Quelles métadonnées faut-il conserver ou adapter ? Pourquoi ne pas sélectionner toutes les informations encodées en TEI ?

Puisque le format XML répond aux exigences de l'édition numérique et le JSON du IIIF à celui de la diffusion d'images, il est certain que, même si nos métadonnées doivent décrire le même document, qu'il soit au format texte ou image, les informations nécessaires à la contexte de celui-ci diffèrent.

1. Les champs de métadonnées IIIF se proposent de suivre le Dublin Core, d'où leur mention en anglais. Il s'agit d'un format constitué pour proposer une manière commune de décrire les documents, comprenant 15 éléments articulés autour de 3 champs descriptifs : contenu, propriété intellectuelle et instanciation²³.
2. Il n'est pas nécessaire pour un manifeste IIIF de contenir les informations propres à l'édition numérique renvoyant aux responsables du projet, aux dates de modification du fichier ou encore à l'encodage du texte... Un certain tri est donc opéré pour n'utiliser que les métadonnées descriptives de la source.
3. Au sein même de ces données bibliographiques, certaines sont légèrement adaptées à la visée de la métadonnées IIIF. C'est particulièrement le cas de l'élément **date**. Cet élément peut être multiplié dans le fichier XML, permettant ainsi de préciser les différentes datations proposées en mentionnant les responsables. Précisons de suite que, pour notre corpus, les différences de datation s'effectuent à l'échelle du mois ou du jour, et non de l'année qui -jusqu'à preuve du contraire- fait toujours consensus. Cependant, cette possibilité de répétition de l'élément est la traduction d'un travail d'édition scientifique, visant à analyser le document et mettre sur le même plan diverses propositions. Nous avons jugé que ces divergences sont moins cohérentes pour les métadonnées de la numérisation et avons donc fait le choix de ne conserver que l'année d'impression pour situer historiquement le document. Répéter les différentes datations et leurs auteurs pourrait être source d'incompréhension pour qui visualise la mazarinade en dehors de notre projet d'édition.

23. *Dublin Core*, URL : <https://www.bnf.fr/fr/dublin-core> (visité le 13/09/2022).

Intégration des liens vers les images

Après le traitement des métadonnées du manifeste, le même script considère la création des canevas contenant les images. Il s'agit pour cette étape d'articuler la récupération des URI des images IIIF disponibles sur le serveur CERES avec l'écriture du manifeste.

Pour réaliser cela, il faut pouvoir requêter le serveur pour récupérer les liens. Python intègre le paquet *requests*²⁴ qui permet la connexion à un protocole HTTP(S)²⁵. Rappelons que l'URI de nos images contient le numéro Moreau de la mazarinade et un numéro de page. L'idée est donc, pour chaque document et dans l'ordre croissant, de vérifier s'il existe un URI dont le chemin comprend le numéro Moreau correspondant à la mazarinade traitée et un numéro de page en commençant à la page 1.

```
# on vérifie ici l'existence des 9 premières pages
for i in range(1,10) :
    url = f"https://ceres.huma-num.fr/iiif/3/antonomaz--antonomaz--{info['numero_Moreau']}_GBOOKS-00{i}/full/max/0/default.jpg"
    lien = requests.get(url)
```

Si la requête renvoie le code 200, la connexion est établie. Autrement dit, la page existe. Nous pouvons donc enregistrer le lien et vérifier si une page suivante existe. On en profite également pour récupérer les informations de taille de l'image qui seront aussi affichée dans le canevas. Dès qu'une requête ne reçoit pas une réponse 200, la boucle est cassée, signifiant que toutes les pages de la mazarinades ont été récupérées.

Chaque page trouvée déclenche la création de son canevas. L'identifiant du canevas correspond à l'URI de l'image IIIF. Pour la première page de notre Moreau29, c'est donc :

```
https://ceres.huma-num.fr/iiif/3/antonomaz--antonomaz--Moreau29_GBOOKS-001
```

L'affichage de l'image se gère plutôt dans le **body** du canevas, dans un nouvel *id* où est indiqué un lien plus complet qui, comme nous l'avons vu au début de ce chapitre, précise comment l'image doit être affichée. Puisque notre objectif est de proposer la numérisation du document, nous affichons chaque image dans son entièreté.

24. <https://pypi.org/project/requests/> (visité le 02/09/2022)

25. Ce protocole *HyperText Transfert Protocol (Secure)* est l'élément clef permettant la communication entre le serveur et le client souhaitant s'y connecter. Le HTTPS est la version sécurisée du HTTP et donc à privilégier. Voir *HTTPS (HyperText Transfert Protocol Secure) : définition claire et pratique*, URL : <https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203459-https-hypertext-transfert-protocol-secure-definition/> (visité le 13/09/2022).

```

" id ":" https://ceres.huma-num.fr/iiif/3/antonomaz—antonomaz—
oreau29_GBOOKS-001",
[ . . . ]
" body ":{

    " id ":" https://ceres.huma-num.fr/iiif/3/antonomaz—
Moreau29_GBOOKS-001/full/max/0/default.jpg",
    " type ":"Image",
    " format ":"image/jpeg",
    " height ":480,
    " width ":323
}

```

Vérification de la validité du fichier

Pour chaque fichier XML-TEI donné en entrée, la fin du script enregistre le manifeste en le nommant au numéro Moreau correspondant. Bien que nous ayons à présent une matière faisant office de manifeste, il nous paraissait nécessaire de vérifier la validité de ce qui venait d'être écrit.

Un outil très pratique pour réaliser cela est la librairie python *tripoli* dont le paquet *IIIFValidator*²⁶ valide le document ou imprime les erreurs de structure dans la console.

Cependant, cette librairie ne vérifie que les manifestes écrits sous l'API 2, dont la structure diffère suffisamment pour ne pas être employable avec nos manifestes. Nous avons donc, après avoir mis en ligne notre manifeste sur le serveur d'Antonomaz, tester notre lien avec un IIIF Validator compatible avec l'API Présentation 3²⁷.

4.3 La valorisation et la réutilisation des données et métadonnées IIIF

IIIF produit un contenu intéropérable et donc largement compatible avec l'objectif de valoriser numériquement le corpus. Si la diffusion constitue en elle-même une valorisation, elle est aussi l'occasion pour l'équipe du projet de posséder une nouvelle forme de contenu à développer et de proposer une matière supplémentaire, qui ne se limite pas à l'unique exposition des mazarinades dans la forme renseignée par les bibliographies.

26. <https://github.com/DDMAL/tripoli>

27. *Presentation API Validator*, URL : <https://presentation-validator.iiif.io> (visité le 13/09/2022).

4.3.1 Regrouper numériquement le corpus

L'étape précédente a permis d'aligner le futur traitement des images issues des PDFs Google Books avec celles de la Bibliothèque nationale de France et de la Bibliothèque Mazarine : tout notre corpus est à présent disponible en IIIF. Nous pouvons donc le diffuser entièrement sur un seul site internet en ajoutant, pour chaque mazarinade, son texte numérisé.

Cette accessibilité concentrée à un même endroit représente un intérêt pour un chercheur en sciences humaines. Déjà parce qu'elle constitue un gain de temps : l'accès distant facilite grandement la consultation des documents, moins chronophage qu'un déplacement en bibliothèque. Bien sûr, le numérique fait perdre le rapport à la matérialité, mais se pose en bonne position pour offrir l'accès au contenu et toute forme de traitement numérique.

De plus, cette concentration du corpus limite la navigation sur le web : plus besoin de naviguer entre les différentes bibliothèques numériques pour trouver la mazarinade désirée.

Cette mise en ligne participe donc à rendre accessible, non seulement un texte, mais également un texte compréhensible dans un ensemble littéraire dont la publication à l'intensité nouvelle est une clef de lecture fondamentale pour l'étude de la Fronde.

Revenons sur un point de conservation évoqué au début de ce travail : la plupart des mazarinades ont été conservées en recueil mais numérisées individuellement. Puisqu'il est possible de créer plusieurs séquences d'images dans un manifeste, nous pourrions imaginer, reformer numériquement ces recueils, révélant une logique de conservation qui, pour le moment, n'est saisissable qu'avec le déplacement en centre d'archives. Ce format pourrait constituer un autre moyen de fouiller le corpus ou d'analyser sa conservation.

En tout cela, notre démarche contribue à la constitution d'une « hyper-collection ²⁸ », qui, en cassant la barrière de la consultation physique, forme une collection numérique réunissant les fonds patrimoniaux de plusieurs bibliothèques européennes. Est alors créé un ensemble de documents anciens, dont le groupement surpassé la contrainte matérielle bien que, rappelons-le, le bouquet numérique dépends des documents numérisés et ne corresponde pas à la totalité des mazarinades disponibles à la consultation physique. Il est alors à penser comme une masse évolutive, nécessitant des moissonnages réguliers, glanant les ajouts dans les bibliothèques numériques respectives ²⁹.

28. E. Bermès et F. Martin, “Le concept de collection numérique”...

29. Les campagnes de numérisation poursuivies par les institutions patrimoniales au fil des financements conditionnent un ajout progressif, voire au compte-goutte, des documents dont l'état est propice à l'étape de la numérisation.

4.3.2 Exploiter le corpus

Diffuser le corpus en IIIF donne diverses opportunités en terme d'exploitation numérique.

Pour l'équipe du projet

Du point de vue technique pour le projet, IIIF contribue à développer le contenu scientifique du site, ajoutant une valeur supplémentaire au travail de référencement et de mise en ligne. IIIF s'articule particulièrement bien avec le montage d'expositions virtuelles, réalisations attractives pour maintenir un certain dynamisme du site web et renforcer la légitimité scientifique du projet : une réflexion dépassant le travail d'édition à l'échelle du document individuel est diffusable, illustrée directement avec les images IIIF.

Le standard prévoit des outils pour ce genre d'entreprise, facilement utilisable puisque la forme des données sont déjà adaptées au fonctionnement de l'outil. Aussi, il est possible d'utiliser *Exhibit*³⁰, pensé pour gérer ce type de mise en place avec un système de navigation guidée.

Des images ou zones d'images pourraient alors être réutilisées, sans être doublement hébergées, pour constituer de nouveaux manifestes aux thématiques définies. Puisque le projet a entamé un travail conséquent sur les imprimeurs et imprimeuses de mazarinades, une exposition des marques d'imprimeurs pourrait être envisagée, mettant en avant l'identité du travail. Des publications illicites aux impressions royales, les pages de titre font jaillir des éléments de contexte de production loin d'être anodins pour l'analyse et la compréhension globale de ce corpus.

Pour les visiteurs

Puisque chaque manifeste est structuré selon un standard, tous les fichiers sont donc requêtables de manière commune *via* le format JSON, réputé pour sa facilité de lecture humaine et informatique. D'un point de vue utilisateur, toutes les données IIIF sont facilement accessibles, du moins lisibles³¹.

L'image pouvant être visionnée en dehors du site, tout autre projet en humanités numériques pourra en bénéficier, sans l'obstacle de l'hébergement. Elle peut servir de base pour tout individu souhaitant entraîner un modèle d'HTR : le logiciel eScriptorium accepte l'entrée de manifestes IIIF, récupérant à la fois les images et les métadonnées.

Concernant la consultation des manifestes, IIIF valorise largement la visibilité des métadonnées, fondamentales à la compréhension humaine. Aussi, les visionneuses de do-

30. <https://www.exhibit.so> (visité le 13/09/2022)

31. Il existe une API IIIF permettant de limiter la réutilisation des données, il s'agit de l'API Authentification.

cuments IIIF intègrent-*t*-elles une exposition des métadonnées renseignées, pensée pour être intuitive et consultée par l'utilisateur.

Dans le même temps, l'individu accède à une image de haute qualité, sur laquelle il peut se déplacer et zoomer pour accéder aux détails l'intéressant. Certaines visionneuses permettent également la consultation multiple de documents dans une même fenêtre, option pratique pour comparer des documents étudiées alors qu'il est rarement possible de faire de même avec les archives physiques.

Puisque notre propos nous porte à considérer les avantages de la visualisation des manifestes IIIF, nous pouvons à présent nous justifier sur les choix de visionnage effectués afin de rendre nos mazarinades consultables.

Troisième partie

**La construction d'une page de
consultation pour chaque
mazarinade**

Chapitre 5

Mettre en ligne le corpus avec TEI Publisher

5.1 TEI Publisher : un outil pour l'édition numérique

5.1.1 Qu'est-ce que TEI Publisher ?

Pour tout projet de valorisation numérique, se pose la question de la structure web qui accueillera le projet et le mettra à disposition des utilisateurs. S'il est possible de construire entièrement un site internet, un outil a été pensé par le développeur allemand Wolfgang Meier permettant de générer une application « pré-fabriquée » pour la mise en ligne de corpus. C'est ce que propose l'outil TEI Publisher¹ :

The motivation behind TEI Publisher was to provide a tool which enables scholars and editors to publish their materials without becoming programmers, but also does not force them into a one-size-fits-all framework. Experienced developers will benefit as well by writing less code, avoiding redundancy, improve maintenance and interoperability - to just name a few. TEI Publisher is all about standards, modularity, reusability and sustainability !²

La logique du fonctionnement de l'outil est de bâtir un site internet autour de fichiers XML qui forment alors autant la source du contenu intellectuel de l'édition que la base technique de la publication électronique. Autrement dit, TEI Publisher travaille à partir des fichiers XML et permet de les manipuler en tant que base d'une application web qui permet de partager les textes à la communauté scientifique. À partir des XML,

1. *TEI Publisher*, URL : <https://teipublisher.com> (visité le 07/09/2022), En développement depuis 2015, la version 7 est sortie en décembre 2020.

2. <https://teipublisher.com/exist/apps/tei-publisher/doc/documentation.xml?odd=docbook.odd> (visité le 08/09/2022)

différentes sorties web peuvent être produites, c'est par ce biais que sont générés les différentes pages HTML³ qui constituent le site de l'édition. Chacune de ces pages puise des informations dans le XML et les fait apparaître : c'est comme cela que le texte du document source ou ses métadonnées peuvent être affichés. L'outil propose également des exports en différents formats (dont le PDF) et intègre le langage Markdown⁴ pour les pages ou éléments HTML statiques, qui ne dépendent pas du contenu d'un fichier XML à charger (page de remerciements, de bibliographie...). C'est toute l'application qui est construite autour du corpus que l'on souhaite mettre en ligne.

L'outil fonctionne avec le TEI Processing Model⁵, conçu pour préfigurer les actions et les comportements des éléments TEI selon un résultat attendu : par ce système, on détermine si un élément TEI doit se comporter dans le HTML comme un lien, une cellule de tableau, un titre de page ou encore une métadonnée.

Ce fonctionnement autorise un affichage différent d'un même élément TEI en fonction d'un contexte défini : un même élément <**persName**> peut avoir un rendu différent selon s'il est positionné dans le corps du texte ou s'il est contenu dans la métadonnée précisant le responsable du projet.

Pour réaliser cela, des paramètres sont à assigner à l'élément travaillé pour pouvoir être appelés à l'endroit où son apparition est souhaité sur la page HTML. Avec cette possibilité, c'est tout l'affichage d'un contenu qui est personnalisable : un titre centré en gras, tous les noms des individus en italique ou seulement ceux dont l'élément <**persName**> contient un certain attribut...

TEI Publisher génère donc des pages web à partir de fichiers TEI stockés dans exist-db⁶, une base de données documents que l'outil requête pour injecter les éléments des fichiers dans le document HTML.

Pour enrichir la mise en ligne du corpus, des composants peuvent être implémentés. Ceux-ci contiennent des fonctionnalités de l'application : page dynamique d'index ou encore visionneuse pour les numérisations. Il est possible d'utiliser directement ceux présents par défaut dans l'application mais également de créer les siens. Là encore, il est cohérent d'insister sur le fait que TEI Publisher est un outil qui est conçu pour faire une mise en ligne dès l'ajout des fichiers XML, sans compétences avancées en programmation ou dé-

3. Le *HyperText Markup Language* est le langage à balise conçu pour afficher les pages web.

4. Crée en 2003, ce langage permet une mise en forme très rapide et facile à lire et écrire. Voir <https://www.markdownguide.org>

5. Le TEI Processing Model est un framework contenant les règles de transformation de la TEI vers un autre langage, comme dans notre cas présent, vers le HTML, sans passer par le langage XSLT. Il a été intégré aux recommandations TEI en 2016. Voir <https://tei-c.org/release/doc/tei-p5-doc/fr/html/TD.html#TDPMPM> (visité le 05/09/2022)

6. Créée par le même développeur que TEI Publisher, exist-db est une base de données propre au format XML.

veloppement web mais qui est aussi pensé pour être complètement personnalisable et qui peut donc, techniquement parlant, être largement approprié. Utiliser cet outil, même avec une grande personnalisation, c'est avoir une structure et une logique de fonctionnement communes à d'autres projets d'édition numérique.

5.1.2 Une communauté grandissante

Dans le domaine des humanités numériques, l'édition de corpus prend une place conséquente. Nombreux sont les projets qui portent cette ambition. Après l'encodage des textes motivant l'édition, la mise à disposition de ceux-ci sur le web est un enjeu à part entière et doit être pensé comme un élément clef du travail.

La collaboration pour penser, améliorer et préciser l'encodage des textes est un fait bien avéré dans le monde de l'édition numérique, déjà parce que la TEI est pensée comme un ensemble de recommandations qui prend poids de standard dans le domaine, ensuite parce que, pour faire autorité, elle doit être envisagée dans un cadre suffisamment large pour que son degré de précision puisse convenir à tous les types de sources. Aussi, tous les ans, un congrès international est réuni pour discuter de potentiels changements dans les recommandations et, dans le même temps, des groupes de travail se forment pour proposer des guides d'encodage propres aux différents types de sources⁷.

Aussi, si l'on constate une dynamique solide autour des questions d'encodage, celle-ci est plus faible lorsqu'il s'agit de s'intéresser à la mise en ligne du travail encodé. TEI Publisher vient quelque part combler ce manque et se présente comme une solution qui pourrait être fédératrice pour mettre en ligne les corpus TEI.

L'ensemble du code de l'outil étant disponible sur github⁸, il hérite d'une accessibilité favorisant son utilisation. En ce sens, il peut intéresser les équipes ne bénéficiant pas particulièrement de budget pour le développement applicatif. Il est aussi conçu pour être accessible sans la nécessité de maîtrise avancée de la programmation. La mise à disposition du code sur github encourage les échanges entre les équipes : le système d'*issues* offre un lieu de questionnement et de débat sur les problèmes techniques rencontrés mais aussi de proposition d'innovations ou de modifications. Chacun peut donc prendre part à la discussion et contribuer à l'amélioration de l'outil. Il est également possible de rejoindre un fil de discussion hébergé sur une plateforme de communication.

La principale organisation promouvant TEI Publisher et forgeant une communauté

7. C'est notamment le cas du *consortium CAHIER* qui s'intéresse principalement aux problématiques d'édition numérique. Voir <https://cahier.hypotheses.org/1e-consortium>. Il propose par exemple un guide d'encodage pour l'édition numérique de correspondances : <https://cahier.hypotheses.org/guides/guide-correspondance>.

8. <https://github.com/eeditions/tei-publisher-app> (visité le 02/09/2022)

autour de l'outil se nomme *e-editiones*⁹, une association suisse s'intéressant à l'édition numérique. D'une part, elle propose de centraliser les projets utilisant l'outil : ces derniers peuvent contacter l'organisation pour se déclarer¹⁰. En cela, elle offre une visibilité très utile, à la fois pour découvrir et suivre le développement des applications renseignées, et pour faciliter les prises de contact entre équipes des projets, ayant alors existence de telle ou telle initiative pour échanger sur un problème technique ou partager une fonctionnalité intéressante.

Depuis avril 2021, *e-editiones* a mis en place des *community meetings* prévus pour être mensuels et se pencher, à chaque réunion, sur un thème ou un projet lié à TEI Publisher¹¹.

5.1.3 La personnalisation de l'affichage des éléments XML-TEI

Nous pouvons à présent présenter la gestion de l'affichage des éléments XML-TEI sur la page HTML. La personnalisation s'effectue à plusieurs échelons et nous a paru quelque peu particulière, nécessitant un temps de prise en main. Aussi, pour illustrer notre propos, nous prendrons l'exemple de la gestion de l'affichage des commentaires additionnels sous les métadonnées descriptives. Dans notre fichier XML-TEI, ils sont renseignés dans l'élément **<abstract>** du **<teiHeader>** :

```
<profileDesc>
  <langUsage>
    <language ident="fra">Document en français</language>
  </langUsage>
  <abstract>
    <p source="Mazarine">Relatif au traité de Münster de mai 1648. Autre émission (BM01474) avec
      le même titre : Paris, Claude Boudeville, 1649, 15-[1 bl.] p., in-4. Adresse restituée
      d'après l'autre émission. D'après Hubert Carrier, texte paru à l'automne 1649
      (Bibliothèque Mazarine, Ms. 4682-3, f. 36v).</p>
    <p source="Antonomaz">La page de titre ainsi que la dernière page sont manquantes dans la
      numérisation.</p>
    <p source="Moreau">1556. Harangue de M. Servient (sic), faite aux Hollan dois, sur le sujet
      de leur traité de paix avec l'Espagnol. S. l., 1649, 15 pages. Le traité n'était pas
      conclu encore. C'est une des pièces qui ont été publiées pour justifier Mazarin du
      reproche de n'avoir pas voulu faire la paix avec l'Espagne à Munster.</p>
  </abstract>
```

FIGURE 5.1 – Notes additionnelles dans un fichier XML-TEI

Chacun des commentaires, issus des notices de la bibliothèque Mazarine, de la bibliographie de Moreau ou de l'équipe du projet, est inséré dans un élément **p** avec pour attribut l'origine du commentaire.

Du côté de TEI Publisher, nous travaillons avec une interface permettant de gérer un ODD spécifiant la transformation des données XML-TEI vers le rendu HTML. Cet

9. *e-editiones*, URL : <https://e-editiones.org> (visité le 07/09/2022).

10. Une fois le contact établi, le projet est renseigné sur une carte et est accessible depuis une page registre. Voir : <https://e-editiones.org/map/>.

11. <https://e-editiones.org/community-meetings>

ODD est donc bien à différencier de celui constitué pour le schéma d'encodage des fichiers.

Nous commençons par sélectionner l'élément **p** dans lequel sont inscrits les commentaires. Seulement, l'élément **p** peut se trouver ailleurs que dans notre **abstract** qui nous préoccupe actuellement. On définit alors ce qu'on appelle un *model* auquel on attribue un *predicate* soit un contexte de lecture pour la machine :

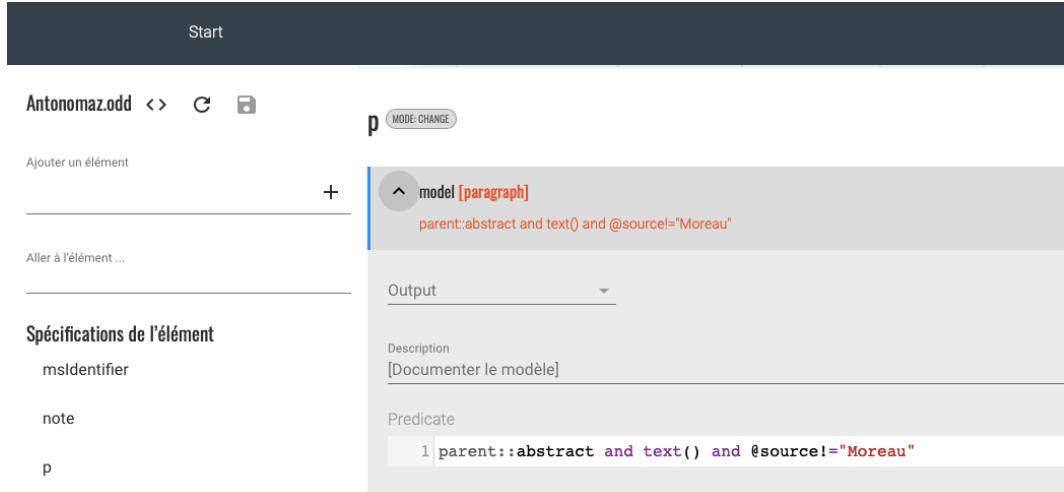


FIGURE 5.2 – Attribution d'un prédicat dans TEI Publisher

On indique ici que l'élément doit se trouver dans un élément **abstract**, contenir du texte (on évite ainsi un problème d'affichage avec un élément vide) et ne pas être un commentaire Moreau que l'on ne souhaite pas afficher sur la page (bruités, ils sont surtout utiles en interne pour faciliter les relectures). Si l'élément répond aux conditions, il est sélectionné sinon il n'est pas considéré. Quand l'élément renvoie à un commentaire de la Bibliothèque Mazarine ou de l'équipe Antonomaz, on lui attribue un modèle précisant quel sera son affichage (*template*) :



FIGURE 5.3 – Définition d'un *template* dans TEI Publisher

Ici, nous demandons dans le *template* à ce qu'apparaissent sur le site un texte contenant la source du commentaire et le contenu du commentaire. Pour cela, nous écrivons le rendu texte en utilisant deux variables :

`[[source]]` qui pioche le contenu de l'attribut `@source`

`[[content]]` qui, par défaut, récupère le contenu texte d'un élément

La variable `[[source]]` est personnalisée dans la section *Parameters* qui permet de préciser un comportement ou un contexte. Ici, par une requête XQuery¹², nous faisons en sorte d'avoir un rendu commode, lisible et agréable pour le visiteur. Par exemple, si la source renvoie à 'Mazarine', nous n'afficherons pas « Commentaire Mazarine : » mais plutôt « Commentaire de la Bibliothèque Mazarine : ».

Enfin, la partie *Renditions* permet d'appliquer un style CSS (*Cascading Style Sheets*) pour gérer la mise en forme. Nous demandons ici que le texte soit justifié.

Cependant, nous devons maintenant inclure ce rendu HTML dans un contexte d'affichage plus large. Notre élément **p** est situé dans un **abstract** qui sera affiché, sur la page web, dans les métadonnées du document. Nous lui attribuons alors le mode défini pour l'affichage des métadonnées et indiquons que, pour chaque **p** ayant matché avec le contexte défini ci-dessus, le *template* sera placé dans un élément HTML désignant le paragraphe.



FIGURE 5.4 – Définition du mode d'affichage dans TEI Publisher

Nous précisons en dernier comment doivent s'intégrer nos éléments **p** dans les métadonnées, à savoir en dernière position :

12. XQuery est un langage de requête pour les documents XML. Il est utilisé pour extraire l'information dans les bases de données document.

```

Template
1 <h3 style="margin-bottom: 50px;">Métadonnées</h3>
2
3 <div style="width:100%; overflow:auto; font-size:19px;">
4   <div style="text-align:left; position:relative; float:left; width:72%;">
5     <table>
6       [[bibl]]
7       [[msIdentifier]]
8     </table>
9   </div>
10
11   <div style="font-size:smaller; text-align:left; position:relative; float: right; width:>[[textClass]]
12   </div>
13 </div>
14
15 <div>[[abstract]]</div> ←
16
17

```

FIGURE 5.5 – Intégration de l'*abstract* pour l'affichage des métadonnées

Nos deux balises XML-TEI contenant les commentaires à afficher seront bien transformées en deux éléments HTML et apparaîtront en dernière position sur la page web¹³.

```

▼ <p class="tei-p">
  Commentaire de la Bibliothèque Mazarine : Relatif au traité de Münster de mai 1648. Autre émission (BM01474)
  avec le même titre : Paris, Claude Boudeville, 1649, 15-[1 bl.] p., in-4. Adresse restituée d'après l'autre
  émission. D'après Hubert Carrier, texte paru à l'automne 1649 (Bibliothèque Mazarine, Ms. 4682-3, f. 36v).
</p>
▼ <p class="tei-p">
  Commentaire Antonomaz : La page de titre ainsi que la dernière page sont manquantes dans la numérisation.
</p>

```

FIGURE 5.6 – Résultat HTML des éléments sélectionnés

A chaque enregistrement de l'ODD, des fichiers de transformation actualisent le résultatat HTML. Toute cette mécanique confirme la flexibilité de l'outil TEI Publisher que nous allons également tester dans le chapitre suivant sur un autre point : l'intégration d'une visionneuse IIIF.

5.2 Choisir son outil de mise en ligne

Après cette présentation de l'outil et de son succès croissant, il tient de rappeler que TEI Publisher n'est qu'une des options possibles pour mettre en ligne une édition numérique parmi lesquelles il constitue une alternative de plus en plus séduisante pour les projets.

13. Le résultat est développé dans le chapitre suivant.

5.2.1 Les options possibles

Publier une édition numérique, c'est avant tout publier un site web, c'est sous-entendre construire un site web. Cette dernière étape, fondamentale pour la mise à disposition du travail scientifique, doit être portée par un développement applicatif bien envisagé. En effet, face à ce choix technique, plusieurs langages sont candidats. En sciences humaines, les plus considérés sont le langage de programmation Python, le langage de transformation XSL ou bien le système de publication Omeka S.

Python

Le langage Python propose une librairie Flask qui est très utile pour monter une application web¹⁴. Le site se construit alors par la création de routes, une par page HTML souhaitée. Connectée à une base de données, la librairie est capable d'aller y puiser des informations pour les injecter dans le contenu des pages HTML qui fonctionnent également sur une logique de *templates*, il s'agit d'un modèle structuré à partir duquel Flask automatise la création de la page web. L'installation de la librairie *lxml* est nécessaire pour naviguer dans les fichiers XML-TEI et les requêter.

Cependant, si l'efficacité de Flask est démontrée pour les petites applications web, elle n'assure pas un volume de données trop conséquent. Dans ces cas-là, l'utilisation de Django est préférable mais nécessite une place de stockage conséquente.

XSLT (*eXtensible Stylesheet Language Transformations*)

Le langage XSLT est « est un langage de programmation fonctionnel utilisé pour spécifier comment un document XML est transformé en un autre document qui peut, mais qui n'est pas nécessairement, un autre document XML. Un processeur XSLT lit un arbre XML en entrée et une feuille de style XSL et produit un arbre résultat en sortie.¹⁵ »

Si le recours à XSLT est fréquent pour modifier un fichier XML-TEI et, en quelque sorte, le réécrire en restructurant son contenu, il propose également une transformation vers le langage HTML. Chaque élément TEI peut être sélectionné pour se voir attribuer un rendu HTML. Cette feuille de style XSL pourrait tout à fait être intégré à une application web (Flask par exemple), et fonctionnerait comme un *template*, générant automatiquement la page web au clic de l'utilisateur.

14. Miguel Grinberg, *Flask Web Development : Developing Web Applications with Python*, O'Reilly Media, 2018.

15. Elliotte Rusty Harold, W Scott Means et Philippe Ensarguet, *XML en concentré*, O'Reilly, Paris, 2005, p. 519.

Omeka

Omeka (Omeka Classic ou Omeka S dans sa plus récente version) est présenté comme un « système de publication web spécialisé dans l'édition de collections muséales, de bibliothèques numériques et d'éditions savantes en ligne se situe à la croisée du système de gestion de contenus (CMS) de la gestion de collections patrimoniale ainsi que de l'édition d'archives numériques.¹⁶ » Omeka est un *Content Managing System* (CMS) personnalisable avec l'ajout de modules. Des projets d'édition numérique optent pour cet outil, considéré comme « un environnement de recherche pour les éditions scientifiques numériques¹⁷ ». Par exemple, le projet EMAN, « outil de publication numérique pour la diffusion et l'exploitation de manuscrits et de fonds d'archives modernes¹⁸ », est une plateforme basée sur Omeka, conçue pour héberger tous types de projets et intègre des transcriptions en TEI.

Mais, qu'est-ce qui séduit et démarque TEI Publisher de ces alternatives ? Nous pouvons à présent justifier l'utilisation de cet outil pour le projet Antonomaz.

5.2.2 Antonomaz et le choix de TEI Publisher

Si TEI Publisher s'est placé en bonne alternative pour le projet Antonomaz, c'est bien que certaines de ces caractéristiques ont fait pencher la balance en sa faveur.

Qu'est-ce qui a séduit ?

L'idée de boîte à outils, et TEI Publisher se présente ainsi (*The Instant Publishing Toolbox*), est particulièrement intéressante. Chacun part d'une structure opérationnelle et y ajoute les compléments qu'il souhaite, eux aussi déjà prêts ou constructibles.

Débuter la partie de mise en ligne avec cette base à disposition est un gain conséquent de temps, précieux pour les projets de recherche aux financements souvent serrés, permettant aussi de consacrer plus d'énergie à la disposition des éléments nouveaux et à l'habillage du site. La logique de travail avec cet outil émane directement du besoin de manipuler les données d'un fichier XML-TEI : on peut éclater facilement l'information et l'adapter à ses envies d'affichage.

De plus, les composants intégrés sont pensés pour répondre aux besoins d'une édition numérique, à savoir un index des textes avec filtre selon plusieurs critères, une page de consultation individuelle des documents, la possibilité de lier la numérisation ou d'afficher plusieurs versions du texte (texte original et texte traduit par exemple)... Cela permet

16. Présentation d'Omeka, URL : <https://omeka.fr/presentation-omeka> (visité le 10/09/2022).

17. Omeka Classic. Un environnement de recherche pour les éditions scientifiques numériques, URL : <https://ride.i-d-e.de/issues/issue-11/omeka/> (visité le 13/09/2022).

18. <https://odhn.ens.psl.eu/article/eman-edition-de-manuscrits-et-archives-numeriques> (visité le 13/09/2022)

une économie certaine de code à écrire et donc de pouvoir rapidement mettre un résultat en ligne, point très avantageux pour les projets qui doivent souvent rendre compte de leur avancée aux financeurs : sans être achevé, il est possible de présenter quelque chose de concret.

Le fonctionnement en page dynamique permet de générer les pages individuelles au moment du clic de l'utilisateur et donc de ne pas les héberger individuellement. Cela aurait été également possible avec une application Flask mais elle demande une écriture de code conséquente, bien que ce *framework* intègre de nombreuses fonctions pour la génération de page, la structure est à bâtir. XSL permet très bien les transformations vers le HTML mais ne dispose pas de structure pré-construite.

Enfin, dernier point qui n'est pas le plus anodin, utiliser TEI Publisher, c'est utiliser un outil utilisé par une communauté qui, on l'a vu, grandit. Autrement dit, c'est avoir des exemples d'utilisation de l'outil *via* des projets d'édition numérique (et dont les besoins techniques sont similaires). C'est également disposer d'une documentation bien fournie (du moins dans le cas de cet outil) et c'est pouvoir contacter des équipes pour échanger sur un quelconque point technique puisque les projets utilisent un code commun. En cela, la communauté TEI Publisher forme une prolongation de la communauté TEI qui se retrouve autour d'une problématique de diffusion de ces textes.

Si, pour l'instant, la grande partie de la mise en ligne consiste à la mise à disposition de la numérisation et des métadonnées pour donner du contexte à celle-ci, nous pouvons convenir que cette entreprise se rapproche fortement d'une mise en page de bibliothèque numérique. Cela aura pu conditionner l'utilisation d'autres outils comme l'instance Oméka S qui remplit tout à fait cet objectif, également à partir de modules déjà construits.

Cependant, l'objectif à terme est bel et bien d'intégrer la fouille de texte, ce pour quoi Oméka n'est pas conçu. Il faut également penser à la rigidité de la structure d'Omeka, moins flexible que celle de TEI Publisher et donc moins personnalisable. De plus, gardons à l'esprit que le travail d'édition a été fait dans des fichiers XML-TEI, que les numérisations accompagnent, et non l'inverse. Si Omeka possède un module proposant de gérer la TEI, l'investissement dans son développement est moindre que celui sur TEI Publisher. Il est donc plus cohérent de pencher pour un outil pensé pour le travail avec ce format de fichier et basé sur une base de données XML, point fort pour l'indexation des fichiers TEI.

Maintenant que nous avons choisi un outil de mise en ligne, nous pouvons présenter le travail de développement web fait sur celui-ci dans le cadre de la mise en ligne du corpus.

Chapitre 6

Développement applicatif : la construction de la page HTML dédiée à l'affichage du document

6.1 L'implémentation de la visionneuse IIIF Mirador sur TEI Publisher

6.1.1 Qu'est-ce que Mirador ?

Mirador (<http://projectmirador.org>) est une visionneuse (ou visualiseur) capable de charger et afficher en grande résolution des images hébergées avec le standard IIIF. Elle est un outil en *open source*, développé principalement par deux universités américaines : l'université de Stanford et l'université d'Harvard¹. Depuis 2019, une version 3 de la visionneuse est disponible².

Si l'on cherche à se renseigner sur la définition d'une visionneuse de ce type, la première phrase de présentation insiste quasi-toujours sur la possibilité de zoom « profond » qui ne détériore pas la qualité de l'image³. Mirador encapsule la visionneuse OpenSeaDra-

1. A la tête du développement, se trouvent deux chercheurs : Rashmi Singhal (Harvard Arts & Humanities Research Computing) et Drew Winget (Stanford University Libraries).

2. *Mirador version 3*, mai 2019, URL : <https://library.stanford.edu/blogs/stanford-libraries-blog/2019/05/introducing-mirador-3-next-generation-image-comparison-viewer> (visité le 26/07/2022), Elle propose notamment davantage d'options pour personnaliser la fenêtre de consultation comme la possibilité de modifier l'affichage des miniatures, de nouveaux modes d'affichage des documents mais aussi une nouvelle gestion du multi-langue.

3. *Visualiseur Mirador*, URL : <https://doc.biblissima.fr/visualiseur-mirador> (visité le 26/07/2022), Biblissima présente Mirador comme « visualiseur web qui offre des fonctionnalités avancées de zoom, de comparaison et d'annotation d'images en haute résolution, indépendamment du type de document ou de la bibliothèque numérique qui les héberge ».

gon⁴ qui propose les fonctionnalités de zoom et d'affichage. Les deux images ci-dessous illustrent la capacité de zoom de l'outil (voir 6.1).

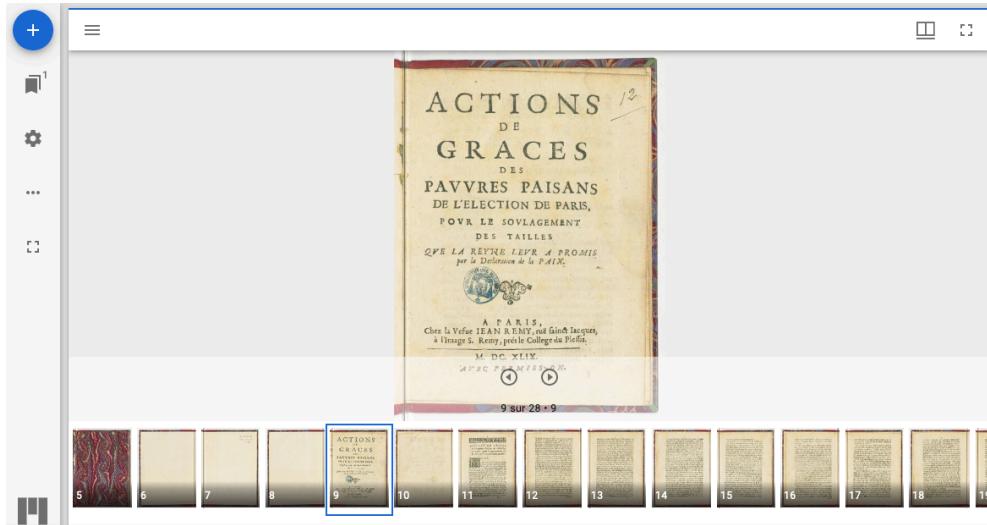


FIGURE 6.1 – Interface au chargement de la visionneuse IIIF

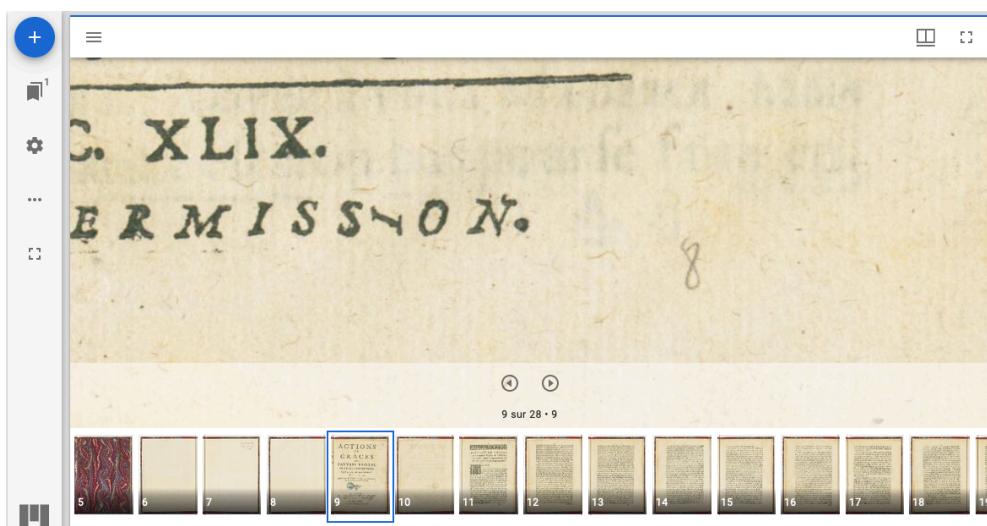


FIGURE 6.2 – Zoom de la visionneuse IIIF sur le bas droit du document

Nous pouvons voir ici la facilité de lecture d'un petit élément manuscrit que peut permettre la visionneuse. Dans notre cas, ce ne sont pas tant les rares ajouts manuscrits qui intéressent particulièrement avec l'utilisation de l'outil, le texte imprimé ne nécessitant pas un tel zoom, que la possibilité de naviguer à son aise sur une page et de s'arrêter sur

4. *OpenSeadragon*, URL : <https://openseadragon.github.io> (visité le 13/09/2022) ; Joris Van Zundert, "On Not Writing a Review about Mirador : Mirador, IIIF, and the Epistemological Gains of Distributed Digital Scholarly Resources", *Digital Medievalist*, 11–1 (août 2018), p. 5, DOI : 10.16995/dm.78, La construction technique de Mirador, fondée sur une logique d'encapsulation de librairies, est développée dans cet article.

une marque d'imprimeur, de se concentrer sur un extrait de la page ou encore d'adapter l'affichage du texte à sa convenance de lecture. Cela dit, il ne faut pas pour autant oublier la part manuscrite -bien que minime- du corpus des mazarinades qui sera potentiellement ajouté dans l'application une fois leur numérisation réalisée⁵.

Pour afficher les images, la visionneuse requête les API IIIF, récupère les métadonnées à la fois à l'échelle du manifeste et à celle du canvas, et charge les images depuis leurs URI. C'est pour cela qu'il n'est pas superflu de rappeler la nécessité de produire des manifestes conformes à la version de l'API IIIF choisie.

6.1.2 Pourquoi Mirador ?

Il est tout à fait légitime de se questionner à propos de ce choix d'implémentation, qui cause un coût supplémentaire, notamment en matière de chargement de la page⁶ sur un outil qui propose déjà une visualisation d'images.

De plus, Mirador n'est pas la seule visionneuse IIIF disponible en *open source* (Universal Viewer par exemple) mais elle est celle en vogue et parmi les plus utilisées, ce qui induit l'existence d'une communauté facilitant la coopération et la recherche de solutions face aux difficultés techniques.

D'ailleurs, le projet n'est pas le seul à s'intéresser à cette accessibilité de Mirador depuis TEI Publisher. Le projet *Démêler le cordel*⁷ n'a lui, pas directement intégré la visionneuse et utilise celle de TEI Publisher mais propose, depuis la page de visualisation, un lien cliquable qui ouvre le document consulté dans Mirador afin de pouvoir comparer celui-ci avec un autre document de son choix (Voir Annexe D.3).

Mirador se positionne comme une visionneuse plus riche que celle proposée par TEI Publisher. Facilement configurable, elle est aussi facilement personnalisable. Une liste d'options est disponible⁸ et il suffit de préciser les paramètres à faire charger.

Mirador offre notamment la possibilité de consulter plusieurs documents dans une même fenêtre ce qui peut être particulièrement intéressant pour proposer une comparaison de mazarinades rééditées ou copiées. Elle permet également l'annotation des textes, est visuellement plus agréable.

En n'utilisant pas la visionneuse par défaut, le projet marque en quelque sorte la

5. À ce jour et à notre connaissance, aucune procédure de numérisation n'est en cours pour cette partie du corpus.

6. TEI Publisher est un outil faisant appel à de nombreux éléments en JavaScript, un langage dynamique pour les pages web dont la lourdeur est parfois reprochée. En ajoutant la visionneuse Mirador, on injecte un code JavaScript supplémentaire.

7. Ce projet propose une bibliothèque numérique des documents de littérature de colportage conservés à l'Université de Genève. Voir : <https://desenrollandoelcordel.unige.ch/inicio.html>

8. <https://github.com/ProjectMirador/mirador/blob/master/src/config/settings.js> (visité le 09/09/2022)

volonté de se détacher du principe d’alignement d’une page de texte et de sa numérisation, ce qui est encouragé par la possession de textes lisibles sans compétences paléographiques.

6.1.3 Mise en place sur TEI Publisher

L’outil Mirador est développé en JavaScript, code qu’il faut donc nécessairement implémenter si l’on veut voir la visionneuse apparaître sur son projet.

Puisque l’on travaille sur une page HTML, la solution la plus évidente est d’ajouter directement le code JavaScript proposé par les développeurs de Mirador pour intégrer leur outil⁹.

```
<script type="text/javascript">
var mirador = Mirador.viewer({
  "id": "my-mirador",
  "manifests": {
    "https://iiif.lib.harvard.edu/manifests/drs:48309543": {
      "provider": "Harvard University"
    }
  },
  "windows": [
    {
      "loadedManifest": "https://iiif.lib.harvard.edu/
manifests/drs:48309543",
      "canvasIndex": 2,
      "thumbnailNavigationPosition": 'far-bottom'
    }
  ]
}) ;
</script>
```

Cependant, on se confronte à l’impossibilité de pouvoir fournir automatiquement le manifeste IIIF correspondant à la mazarinade affichée. En effet, le script attend l’URI du manifeste en chaîne de caractère :

```
"manifests": {
  "https://iiif.lib.harvard.edu/manifests/drs:48309543"
```

9. Le code est disponible ici : <https://github.com/ProjectMirador/mirador/wiki/M3-Embedding-in-Another-Environment#in-an-html-document-with-javascript> (visité le 03/09/2022)

}

Or, il faudrait pouvoir récupérer le lien du manifeste qui serait présent dans le fichier XML-TEI ou pouvoir construire ce lien depuis l'ODD de l'application Antonomaz. Dans les deux cas, c'est donc une variable issue du chargement du document à afficher qui doit être récupérée par le code JavaScript ci-dessus.

Si l'on traduit ce besoin pour correspondre au fonctionnement de TEI Publisher, c'est un paramètre qu'il faut configurer dans l'ODD et injecter dans le HTML par le moyen d'un `<pb-param name="" value="">`. Cependant, cet élément serait à injecter directement dans le JavaScript mais causerait un conflit d'interprétation et serait illisible pour la machine. L'idée a donc été de créer tout de même ce paramètre et de l'injecter dans le HTML afin que le JavaScript récupère la chaîne de caractères qui forme le lien du manifeste.

Pour construire le paramètre contenant le lien du manifeste, la première étape est de pouvoir récupérer ce lien. Pour cela, nous le construisons directement depuis TEI Publisher. Pour éclairer le propos, nous pouvons prendre l'exemple des manifestes hébergées sur le serveur de la Bibliothèque Mazarine.



FIGURE 6.3 – Modèle défini pour reconstruire les liens des manifestes IIIF hébergés par la Bibliothèque Mazarine

Le prédictat indiqué restreint les paramètres définis en dessous aux seuls fichiers XML-TEI dont l'identifiant se termine par '_MAZ'. Il est ensuite possible de définir un paramètre nommé *manifest* qui construit le lien souhaité à partir de celui pointant vers le document dans la bibliothèque numérique Mazarinum. Les deux liens sont plutôt similaires :

1. le lien de la page de notice du document (également renseigné dans le fichier XML-TEI) : <https://mazarinum.bibliothèque-mazarine.fr/ark:/61562/mz18009>

2. le lien du manifeste IIIF : <https://mazarinum.bibliotheque-mazarine.fr/iiif/18009/manifest>

Le paramètre *manifest* récupère donc le lien situé dans l’attribut de l’élément **ref** et le modifie légèrement le chemin d’accès de l’URL. Si l’on traduit la requête de la fonction *replace*, on conserve l’URL jusqu’à l’extension du nom de domaine, on y insère la chaîne ”iiif/” puis l’identifiant situé après le ”mz” qui est commun aux deux liens avant de terminer l’URL par ”/manifest”.

Le résultat est ensuite injecté dans le *template* qui prend la forme d’un **span**, élément HTML qui n’a pas de cohérence ou d’incidence particulière sur la structure d’une page. Il faut veiller à ce que cet élément soit bien créé avant la tentative de chargement de la visionneuse sinon le lien du manifeste ne pourra pas être récupéré. Aussi, il est nécessaire d’insérer dans le code JavaScript existant dans une fonction exigeant le chargement d’un élément défini pour déclencher son exécution, ce qui est possible avec la fonction *.subscribe()*¹⁰. L’ensemble du code s’exécute alors de la manière suivante¹¹ :

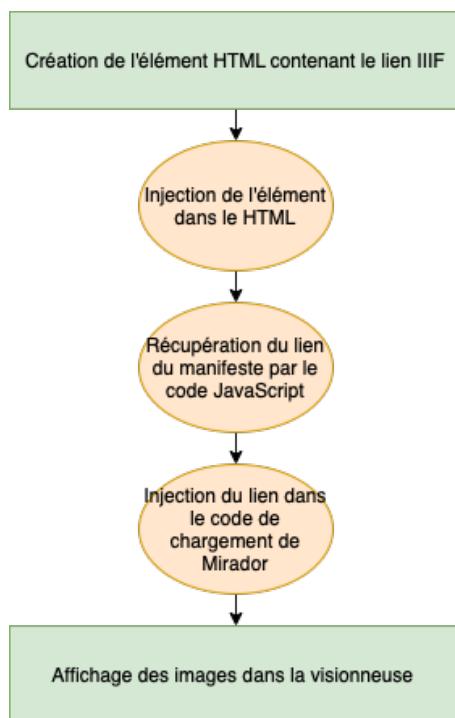


FIGURE 6.4 – Schéma des étapes pour le chargement du manifeste IIIF dans Mirador

10. *Subscribe*, URL : <https://reactivex.io/documentation/operators/subscribe.html> (visité le 13/09/2022), On définit un élément à observer dont elle attend la réussite (ou fin de chargement) pour s’exécuter. Autrement dit, elle connecte un observateur à un observé, le premier attendant de recevoir l’émission du second pour agir à son tour.

11. *Manipulating content via JS*, URL : <https://unpkg.com/@teipublisher/pb-components@1.38.4/dist/api.html#pb-view.3> (visité le 26/08/2022), La manipulation suivie est disponible à l’adresse suivante :

Finalement, le résultat du code permettant de charger la visionneuse Mirador sur TEI Publisher est le suivant :

```
<html>
    <!--head de la page HTML -->
    <body>
        <pb-view id="manif" src="document1" xpath="//teiHeader"
            view="single" subscribe="manifest">
            <pb-param name="mode" value="manifest"/>
        </pb-view>
        <pb-view id="view1" src="document1" view="single"
            before-update-event="before-javascript-update" subscribe="manifest"
            emit="manifest"/>

        <!-- div à placer au lieu d'affichage souhaité de la
        visionneuse -->
        <div id="my-mirador" class="mirador"/>
    </body>

    <script type="text/javascript">
        window.addEventListener('DOMContentLoaded', () => {
            pbEvents.subscribe('before-javascript-update', 'manifest', (ev) => {
                let manifest = manif.shadowRoot.getElementById('manifeste').attributes['href'].value

                var mirador = Mirador.viewer({
                    "id": "my-mirador",
                    language: 'fr',
                    "manifests": {
                        manifest
                    },
                });
            });
        })
    </script>
</html>
```

Maintenant que nous avons obtenu le fonctionnement technique de la visionneuse, nous pouvons présenter la structuration de la page de consultation et ses enjeux.

6.2 Organisation de la page web

6.2.1 Penser l'affichage des données

Le travail sur la construction de la page de consultation a surtout consisté en la mise en place des différents éléments, leur présentation et imbrication et non au rendu graphique. L'habillage est réalisé par le graphiste Jean-Marie Trouiller¹².

Le premier point avant de définir le positionnement précis des éléments est de penser l'organisation générale de la page. Des recommandations existent pour la mise à disposition de documents numériques¹³. Puisque nous voulons permettre la visualisation des documents, des éléments nous sont fondamentaux : le titre du document consulté, une visionneuse affichant la numérisation, une sélection de métadonnées et, pour naviguer sur le site, une barre de menu déroulant.

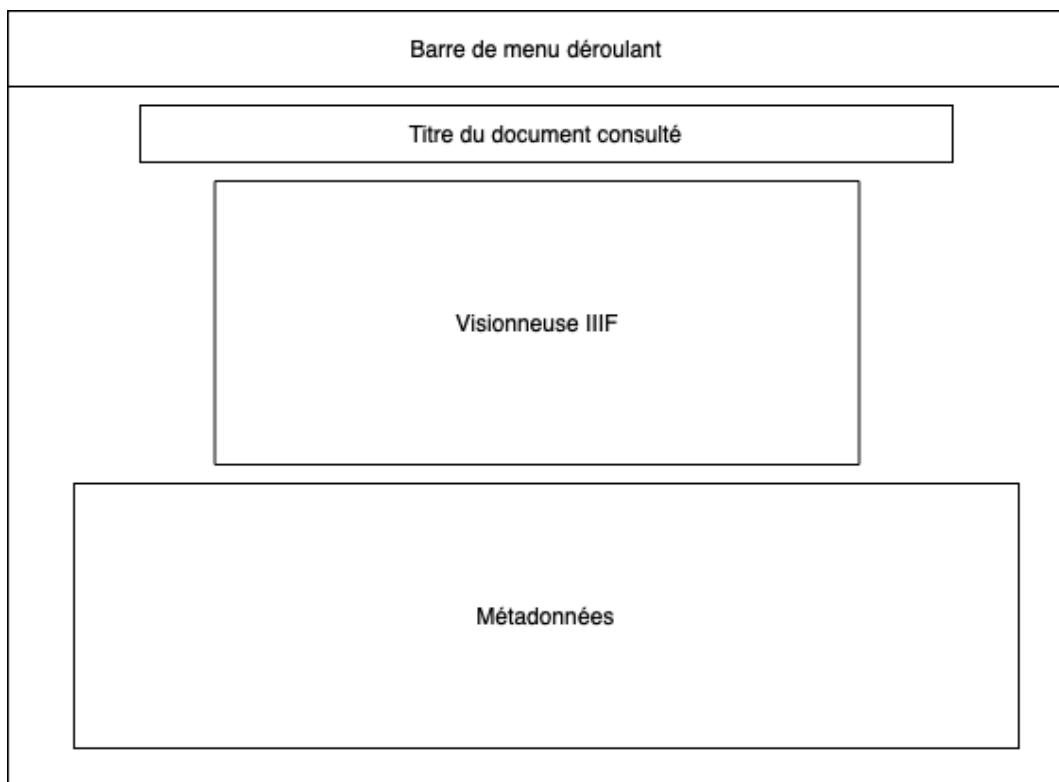


FIGURE 6.5 – Structure de la page de consultation des documents

12. <https://www.cinquantesix.com>

13. *Construire une bibliothèque numérique*, URL : <https://www.idnum.fr/methodoc/construire-une-bibliothèque-numérique/> (visité le 07/09/2022).

Le schéma ci-dessus résume les principaux « blocs », disposés de manière à optimiser l'utilisation de l'espace sur la page, pour l'occuper sans le surcharger. Il est ici question de la qualité « ergonomique » de la page¹⁴ mais aussi de la volonté de pouvoir répondre aux besoins de l'utilisateur à la recherche de contenu scientifique.

Il faut alors réfléchir aux informations exposées dans le bloc de métadonnées. Si les informations sont directement récupérées dans les fichiers XML-TEI, il ne s'agit pas pour autant de faire apparaître toutes les informations du **teiHeader** qui contiennent notamment des éléments propres à la gestion du fichier.

Il faut réfléchir en terme de diffusion scientifique mais aussi en terme d'expérience utilisateur : quelles informations un chercheur souhaite-t-il obtenir ? Ce sont d'abord les métadonnées descriptives, donnant du contexte de compréhension sur la production du document. Aussi, nous sélectionnons les informations d'ordre bibliographique : titre, auteur, éditeur, date de publication, format... Nous informons également du lieu de conservation physique de l'exemplaire affiché : institution, cote... La plateforme de consultation doit obligatoirement préciser les conditions d'utilisation des documents numérisés¹⁵, d'où la mention d'une licence (permettant l'appropriation des données hors utilisation commerciale). Enfin, nous pouvons sélectionner les métadonnées permettant de classer le document : genre et sous-genre littéraire du texte, mots-clefs renvoyant à la thématique du document, évènements ou individus cités...

Tous ces choix se paramètrent directement depuis TEI Publisher en indiquant, pour chaque métadonnée à afficher, où récupérer l'information si elle existe. Le tableau ci-dessous établit le lien entre le type de métadonnée et sa localisation dans le fichier XML-TEI.

14. Javier Barcenilla et Joseph Maurice Christian Bastien, “L’acceptabilité des nouvelles technologies : quelles relations avec l’ergonomie, l’utilisabilité et l’expérience utilisateur ?:” *Le travail humain*, Vol. 72–4 (mars 2010), p. 311-331, DOI : 10.3917/th.724.0311.

15. “Diffusion et exploitation d’un document numérique : information et mise en garde des usagers”, dans *Manuel de constitution de bibliothèques numériques*, Editions du cercle de la librairie, Paris, 2013, p. 102.

Métadonnée à afficher	Élement sélectionné dans le fichier XML-TEI (chemin XPath)
Date	//sourceDesc/bibl/date
Auteur	//sourceDesc/bibl/author/*
Imprimeur	//sourceDesc/bibl/publisher//*
Format	//sourceDesc/note
Nombre de pages	//sourceDesc/bibl/extent/measure/@quantity
Notice de la Bibliothèque Mazarine	//sourceDesc/bibl/relatedItem/@target
Licence	/
Lieu de conservation	//sourceDesc/msDesc/msIdentifier/institution & //sourceDesc/msDesc/msIdentifier/settlement
Fonds, collection, département	//sourceDesc/msDesc/msIdentifier/repository
Cote	//sourceDesc/msDesc/msIdentifier/idno
Genre	//textClass/keywords/@type="genre"
Sous-genre	//textClass/keywords/@type="subgenre"
Mots-clefs	//textClass/keywords/@type="subject"
Commentaire	//profileDesc/abstract/p

TABLE 6.1 – Localisation des métadonnées dans le fichier XML-TEI

6.2.2 Présentation du résultat

Chaque mazarinade est accessible depuis la page d'index, référençant la globalité des documents. Cette page fait alors office de catalogue pour nos mazarinades¹⁶. L'indexation TEI permet de proposer facilement un filtrage des documents d'après le contenu de certains éléments d'encodage. Par exemple, nous pouvons filtrer les documents selon leur lieu de publication, le premier élément de filtre proposé (voir l'image ci-dessous). Ce filtre correspond à l'encodage de l'élément **pubPlace** dans le fichier XML-TEI.

16. « Le catalogue en tant qu'outil de recherche associé à la réalité physique de la bibliothèque est en effet conçu pour donner une représentation de la collection, pour l'incarner. » E. Bermès et Gautier Poupeau, “1. Du catalogue de bibliothèque aux données sur le Web : un changement de paradigme du côté de l'usager”, dans *Le Web sémantique en bibliothèque*, Paris, 2013 (Bibliothèques), p. 17-28, URL : <https://www.cairn.info/le-web-semantique-en-bibliotheque--9782765414179-p-17.htm>.

FIGURE 6.6 – Index du corpus des mazarinades disponibles

Nous avons ici un exemple de l'importance de la qualité et de la précision des métadonnées qui s'ancrent complètement dans la logique des principes FAIR en facilitant la recherche des documents¹⁷.

Aussi, lorsque nous cliquons sur une mazarinade, par exemple la Moreau1556, accessible par ce lien : https://antonomaz.huma-num.fr/exist/apps/Antonomaz/corpus/Moreau1556_GBOOKS.xml, nous arrivons sur cette page :

FIGURE 6.7 – Page de consultation d'une mazarinade (visionneuse IIIF)

17. U. Méditerranée, “DoRANum-Enjeux et bénéfices...”.

La barre de menu est commune à tout le site :

- un bouton « accueil » permettant de retourner sur la page principale depuis n’importe quelle page du site.
- un onglet « corpus » donnant accès à l’index des documents disponibles.
- un onglet « ressources » proposant des accès vers des notices historiques, une bibliographie et une base de données sur les imprimeurs de mazarinades¹⁸.
- un onglet « à propos » renseignant des informations sur la réalisation du projet, son équipe, la documentation technique ou encore un formulaire de contact.
- une barre de recherche permettant de chercher des occurrences dans les textes des mazarinades.

Sous la barre de menu, le titre du document s’affiche. Avant l’affichage de la visionneuse, deux boutons ont été ajoutés :

- le premier « Métadonnées » fait glisser ce qui apparaît à l’écran sur le bloc de métadonnées, situé sous la visionneuse. Cette fonctionnalité permet à l’utilisateur de prendre facilement connaissance de l’existence d’informations liées au document. En effet, nous pouvons voir sur la capture d’écran ci-dessus que celles-ci ne sont pas visibles lors de l’affichage de la page dans le navigateur. Aussi, ce bouton sert à la fois d’indicateur et d’accès rapide. L’objectif est que l’utilisateur puisse s’approprier facilement le contenu du site¹⁹.
- le second « Accéder au texte (OCR) » indique la possibilité de consulter le texte autrement qu’à travers la numérisation, tout en précisant qu’il s’agit d’une version du texte obtenue automatiquement et donc potentiellement imparfaite.

Ensuite, si nous descendons la page internet (ou cliquons sur le bouton « Métadonnées »), nous avons accès aux informations de contexte, chargées directement depuis le fichier XML-TEI.

18. Voir au sujet de la base imprimeur le travail suivant Zoé Cappe, *Enrichissement d’une base de données et encodages en XML-TEI. L’exemple du projet Antonomaz : imprimeurs et repères*, mémoire sous la direction de Thibaut Clérice, École nationale des chartes, 2022.

19. J. Barcenilla et J. M. C. Bastien, “L’acceptabilité des nouvelles technologies...”, « L’appropriation renvoie à la façon dont l’individu investit personnellement l’objet ou le système et dans quelle mesure celui-ci est en adéquation avec ses valeurs personnelles et culturelles, lui donnant envie d’agir sur ou avec celui-ci, et pas seulement de subir son usage. »

Métadonnées

Date	1649	Genre	Discours adressé
Date (Carrier)	automne 1649	Genre	Harangue
Auteur (faux locuteur)	Servien Abel	Mot(s)-clé(s)	
Imprimeur	Sans Nom		
Lieu de publication	Sans Lieu		
Format	in-4		
Nombre de pages	15		
Notice Mazarine	https://mazarinades.bibliotheque-mazarine.fr/ark:/61562/bm50368		
Licence	CC-BY		
Lieu de conservation	Bibliothèque municipale (Lyon)		
Cote	SJ IF 247/188, 125		
<ul style="list-style-type: none">Commentaire de la Bibliothèque Mazarine : Relatif au traité de Münster de mai 1648. Autre émission (BM01474) avec le même titre : Paris, Claude Boudeville, 1649, 15-[1 bl.] p., in-4. Adresse restituée d'après l'autre émission. D'après Hubert Carrier, texte paru à l'automne 1649 (Bibliothèque Mazarine, Ms. 4682-3, f. 36v).Commentaire Antonomaz : La page de titre ainsi que la dernière page sont manquantes dans la numérisation.			

FIGURE 6.8 – Page de consultation d'une mazarinade (métadonnées)

Des liens permettent d'accéder directement aux notices de la Bibliothèque Mazarine, au contenu de la licence sous laquelle sont placées les informations diffusées ou encore à des indications sur les différentes datations. Ici, « Carrier » signifie que la date annoncée est une proposition d'Hubert Carrier. Le lien doit envoyer vers une page détaillant les différentes datations des documents (page à venir).

Cette présentation est commune à toutes les mazarinades encodées par le projet à partir de l'instant où le fichier XML-TEI est déposé dans les fichiers de l'application.

6.2.3 Fonctionnalités futures

Nous avons pu livrer une première proposition de page de consultation pour nos mazarinades opérationnelle mais nous pouvons aussi proposer quelques idées d'ajouts pour compléter le travail.

Depuis le début des années 2000, la Science ouverte est un « un mouvement politique et social visant à la libéralisation de l'accès aux données²⁰ ». Au-delà de rendre accessibles les résultats de la recherche, elle vise également à « pérenniser et dans la mesure du possible partager les données, brutes ou pré-traitées, collectées au cours des projets de recherche²¹ ». Dans cette optique, nous pourrions penser à proposer le téléchargement du fichier XML-TEI depuis la page de consultation. Les principes FAIR encouragent à « ouvrir ses données autant que possible²² ». Précisons cependant que l'ensemble des

20. Bernard Jacquemin, Joachim Schöpfel et Renaud Fabre, “Libre accès et données de recherche. De l'utopie à l'idéal réaliste”, *Études de communication*–52 (juin 2019), p. 11-26, DOI : 10.4000/edc.8468.

21. *Ibid.*

22. Inist - CNRS, “DoRANum-Enjeux et bénéfices : Cycle de vie des données, un outil pour améliorer

documents encodés est déjà disponible et ouvert à tous sur la plateforme Github du projet²³ mais un lien pourrait être effectué depuis la page du site. De même, nous pourrions penser à rendre les métadonnées exportables dans un format CSV ou JSON.

Nous pouvons également penser à rendre le document facilement citable en proposant une citation bibliographique du document consulté qui ne serait qu'à copier depuis sa machine.

Une section de suggestion de documents similaires en fin de page pourrait être mise en place à partir des mots-clefs thématiques indiqué dans les fichiers XML.

Enfin, il serait intéressant de trouver un moyen pour conditionner la page affichée lors de l'ouverture de la visionneuse IIIF à la première page de texte numérisé. En effet, la plupart des documents ont des pages « parasites » pour l'accès direct au texte (reliure, page blanche...).

la gestion, la mise en qualité et l'ouverture des données” (, 2021), Publisher : DoRANum, doi : 10.13143/GDBG-CF63.

23. <https://github.com/Antonomaz/Corpus>

Conclusion

Le projet Antonomaz se situe au cœur des enjeux d'intéropérabilité et d'accessibilité de la donnée en sciences humaines, enjeux qui dynamisent le monde des humanités numériques. La chaîne de traitement démontre à la fois l'intérêt de pouvoir réutiliser des données déjà produites et l'importance de permettre la réutilisation des données issues du projet.

Aussi, nous soulignons ici l'importance de la propreté des données, les rendant exploitables, réutilisables sous d'autres formats et surtout diffusables. Si la préparation de la donnée est une phase de traitement nécessaire à sa visualisation, penser sa plateforme de diffusion l'est également.

Une attention particulière a été portée à proposer au visiteur une visualisation agréable de la numérisation avec la visionneuse IIIF Mirador. La structuration de la page doit porter de manière évidente à la connaissance du visiteur la présence de métadonnées et l'accès au texte océrisé. Toutes les données produites au cours du projet sont pensées pour s'aligner sur les standards, facilitant leur réutilisation. Elles sont aussi librement utilisables à des fins de recherche²⁴. De même, toute la chaîne de traitement est reproductive pour les futures récupérations de mazarinades. À partir d'une numérisation de mazarinade, le projet Antonomaz propose un jeu de métadonnées complet, une récupération automatique du texte, la mise à disposition du document par les technologies IIIF (image) et TEI Publisher (texte et métadonnées).

En vue de leur exposition, les données ont donc subi une « préparation » dans laquelle s'articule le traitement automatisé d'un ensemble massif et le traitement manuel, individualisant le traitement apporté à certaines étapes. L'objectif de l'utilisation du numérique sur le projet Antonomaz n'est pas tant celui d'une structure cherchant l'innovation et le développement de nouveaux outils généralisables aux humanités numériques (bien que tous les scripts soient mis en ligne sur Github et donc réutilisables ou adaptables²⁵), mais plutôt de se saisir d'outils déjà existants pour monter un projet de recherche et trouver des solutions techniques permettant de répondre à des besoins spécifiques aux données traitées. Aussi, notre script de nettoyage des PDFs Google Books devait faire avancer le projet et la vérification manuelle a permis d'accompagner complètement l'objectif.

Ce travail de mise en forme des données en vue de leur diffusion a permis d'introduire des réflexions sur la valorisation numérique. Notre mission de stage se rapproche finalement de la création d'une page de consultation assez typique de celle d'une bibli-

24. Placées sous une licence *Creative Commons*, les données sont, sauf mentions contraires, librement réutilisables hors utilisation commerciale.

25. Les scripts du projet sont notamment disponibles au lien suivant : <https://github.com/Antonomaz/tools>

thèque numérique mettant en ligne une collection. Le point intéressant est que ce résultat a été permis par le biais de tout un travail sur le texte qui vient compléter la mise en ligne de la numérisation et de ses métadonnées.

Il faut là souligner le besoin constant de concilier l'objectif scientifique et l'objectif technique dans l'avancement d'un projet de ce type : avant d'approfondir le traitement du corpus avec des outils de TAL (traitement automatique de la langue) ou la nouvelle production d'un HTR, les forces de l'équipe se sont concentrées sur la création d'une plateforme de mise à disposition des documents que les futurs traitements viendront faire évoluer.

Annexes

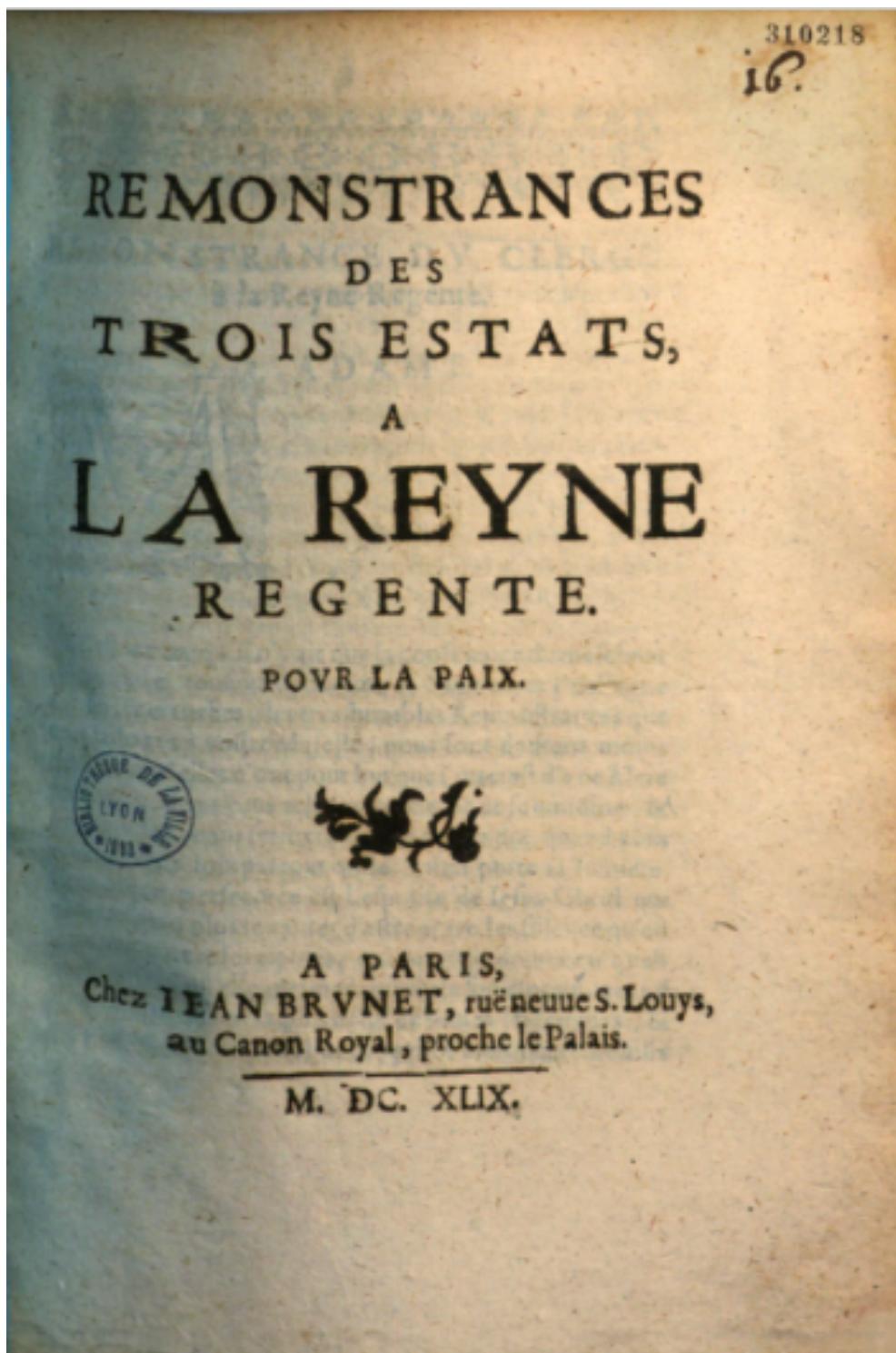
Annexe A

Exemple d'encodage XML-TEI

Pour proposer un exemple d'encodage représentatif, nous utiliserons la mazarinade *Remontrance des trois états à la reine régente pour la paix*¹ identifiée comme Moreau3312.

1. <https://books.google.fr/books?id=5E5mM-9tK0YC>

A.1 Numérisation de la page de titre



A.2 Encodage de la page de titre en XML-TEI

```
<text>
  <body>
    <p>
      <pb n="1"/>
      <lb/>REMONSTRANCES

      <lb/>DES

      <lb/>TROIS ESTATS,

      <lb/>A

      <lb/>LA REYNE

      <lb/>REGENTE.

      <lb/>POVR LA PAIX.

      <figure type="decoration"/>

      <lb/>Chez JEAN BRVNET, ruë neuue S. Louys,
      <lb/>au Canon Royal, proche le Palais.

      <lb/>M. DC. XLIX.
```

A.3 Encodage du *teiHeader*

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title source="Moreau" type="main">Remontrance des trois états à la reine régente pour la paix.</title>
      <title type="sub">Édition Antonomaz</title>
    <respStmt>
      <resp>Chaine de traitement</resp>
      <persName ref="orcid:0000-0001-9518-1040" role="Project_manager">Abiven Karine</persName>
      <persName ref="orcid:0000-0002-4795-2362" role="Project_manager">Lejeune Gaël</persName>
      <persName ref="orcid:0000-0002-0007-1664" role="OCRisation">Tanguy Jean-Baptiste</persName>
      <persName ref="orcid:0000-0003-0850-8266" role="engineer">Bartz Alexandre</persName>
      <persName ref="orcid:0000-0001-5815-9506" role="intern">Faure Margaux</persName>
    </respStmt>
  </titleStmt>
  <publicationStmt>
    <publisher>Projet ANTONOMAZ / Sorbonne Université</publisher>
    <ref target="https://github.com/Antonomaz"/>
    <date type="file_creation" when="2021-11-02">2 novembre 2021</date>
    <availability n="cc-by" status="restricted">
      <licence target="https://creativecommons.org/licenses/by/4.0"/>
    </availability>
  </publicationStmt>
  <sourceDesc>
    <bibl>
      <ref target="https://books.google.fr/books?id=SE5mM-9tK0YC"/>
      <author source="Mazarine" ref="isni:0000000112972841">
        <persName>
          <forename>Cyrano</forename>
          <surname>de Bergerac</surname>
        </persName>
      </author>
      <title source="Moreau">Remontrance des trois états à la reine régente pour la
        paix.</title>
      <pubPlace ref="geonames:2988507" source="Moreau">Paris</pubPlace>
      <publisher ref="isni:000000001831368" xml:id="Il_BrunetJean.xml">
        <persName>
          <forename>Jean</forename>
          <surname>Brunet</surname>
        </persName>
      </publisher>
      <date source="Moreau" when="1649">1649</date>
      <date source="Carrier" notBefore="1649-03-05" notAfter="1649-03-11">entre le 5 et le 11 mars 1649</date>
    <extent>
      <measure quantity="24" unit="page"/>
    </extent>
    <note type="format">in-4</note>
    <relatedItem source="Mazarine" type="identifier" subtype="BM01653" cert="high"
      target="https://mazarinades.bibliotheque-mazarine.fr/ark:/61562/bm50544"/>
  </bibl>
  <msDesc>
    <msIdentifier>
      <settlement ref="geonames:2996944">Lyon</settlement>
      <institution>Bibliothèque municipale</institution>
      <repository/>
      <idno type="cote">Rés 310220</idno>
    </msIdentifier>
    <history>
      <provenance>
        <stamp>True</stamp>
      </provenance>
    </history>
  </msDesc>
  </sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc>
  <p>Cette édition a été réalisée dans le cadre du projet ANTONOMAZ. Son objectif principal
  est de fournir un texte destiné à l'exploration avec des outils électroniques. De ce fait,
  ce n'est ni une édition philologique, ni une édition pédagogique ou de redécouverte d'un
  auteur oublié.</p>
  <p>Les textes encodés dans le cadre du projet ANTONOMAZ sont issus de numérisations de

```

```
<p>Les textes encodés dans le cadre du projet ANTONOMAZ sont issus de numérisations de  
bibliothèques numériques publiques et de Google livres.</p>  
<p>L'édition présentée ici est issue d'un processus d'OCRisation réalisé avec Kraken.</p>  
</projectDesc>  
<editorialDecl>  
<p/>  
</editorialDecl>  
</encodingDesc>  
<profileDesc>  
<langUsage>  
<language ident="fra">Document en français</language>  
</langUsage>  
<abstract>  
<p source="Mazarine">Édition composite : 3 états du cahier C (C1/C2/C3) : pas de marque  
décorative sur la p. de titre (C1) / marque décorative (cul-de-lampe) sur la p. de titre  
(C2, C3) ; p. 12 chiffrée 4 (C2) / p. 12 chiffrée 12 (C1, C3) ; 2 états du cahier F  
(F1/F2) : feuillets F signé B / feuillets F signé F ; p. 21-24 chiffrées 5-8 / p. 21-24  
chiffrées 21-24 Signé "D. B.", identifié comme étant Cyrano de Bergerac d'après Jacques  
Prévot (Cyrano de Bergerac Œuvres complètes, p. 342-349). Chaque partie a sa page de  
titre particulière Paru probablement entre le 5 et le 11 mars 1649, d'après Hubert Carrier  
(Bibliothèque Mazarine, 8° 40523-21/5 [Res] f. 148v) </p>  
<p source="Moreau">3312. Remontrance des trois états à la reine régnante l1 r %, pour la  
paix. Paris, Jean Runet, 1649, 24 pages. Signé D.B. La pièce se divise en trois  
parties avec sous-titre; une pour chaque état, le clergé, la noblesse et  
le peuple. Elle est surtout remarquable de style ; mais elle ne contient ni un fait  
ni une anecdote. 193(51) Voir le Conseiller fidèle.</p>  
</abstract>  
<textClass>  
<keywords>  
<term type="form">prose</term>  
<term type="genre">Rhétorique délibérative</term>  
<term type="subgenre">Remontrances</term>  
<term type="subgenre"/>  
<term type="subject">Conférence de Rueil</term>  
<term type="subject">France, États généraux</term>  
<term type="handwritten note" subtype="no"/>  
<term type="table_of_content" subtype="no"/>  
<term type="illustration" subtype="no"/>  
</keywords>  
</textClass>  
</profileDesc>  
<revisionDesc>  
<change status="corrected" when="2022-07-18" who="MF"/>  
</revisionDesc>  
</teiHeader>
```


Annexe B

Exemple des diverses qualités des PDFs GoogleBooks

B.1 Exemple 1

Au téléchargement, le PDF de la mazarinade *Le manifeste véritable des intentions de M. le Prince*¹, contenait 31 vues pour un imprimé de 8 pages. Les deux dernières étaient largement repétées, créant ce PDF d'une trentaine de pages. De plus, la qualité générale du document numérique est particulièrement mauvaise, une bande de chaque page est systématiquement répétée.

Pour des soucis de confort, nous ne joignons pas le PDF entier mais seulement quatre pages. L'ensemble est téléchargeable depuis le lien en note de bas de page. Nous présentons ici un extrait de la version obtenue une fois le script passé, réduit à 11 pages.

1. *Le manifeste véritable des intentions de M. le Prince, qui ne tendent qu'au rétablissement de l'autorité souveraine et du repos des peuples, présenté à nos seigneurs du Parlement, 1651.* Source : <https://books.google.fr/books?id=AIxAAAAAcAAJ&hl=fr>

tesse Ro yale, & m'en suis rédu le Maistre, pour empêcher que les sectateurs de Mr le Cardinal ne s'en saisissent eux-mesmes, pour luy donner vn poste si auantageux. I'ay en suite voulu informer les principaux habitans & officiers de cette ville d'Angers, des motifs & des raisons de cette conduite; Comme ce n'estoit que pour apuyer l'execution des Déclarations du Roy, desdits Arrests du Parlement, & executer les ordres de Vostre Altesse Royale. Tout ce qu'il y a eu de gens bien affectionnés au bien de l'Estat & au repos des peuples l'ont vnamiment approuuée. Il n'y a eu, Monseigneur, que le sieur Boylefue Lieutenant general, & quelques autres de sa cabale qui l'ont voulu descrirer & la rendre suspecte aux peuples de cette ville. Et comme ils ont veu qu'ils n'y pouuoient pas réussir, ils ont enuoyé vers Mr le Cardinal Mazarin, ont pris des liaisons estroites avec luy pour le rendre Maistre de cette ville, & par consequent de toute cette prouince: & pour y paruenir ont obtenu par son moyen des Lettres du petit cachet du Roy, pour m'arrester & se faire de ma personne par toutes sortes de voyes & de moyens. Ayant été heureusement aduerty de ce dessein, i'ay creu qu'il estoit de ma conduite d'en preuenir l'execution. Pour cet effet, ayant appris que ledit Boylefue Lieutenant general auoit sans mon ordre assemblé le Presidial au Palais le vingt-septiesme du passé, & mandé à tous les Corps & Compagnies de la ville d'en faire de mesme, pour y faire prendre quelque resolution non seulement

stre, pour empes-
rdinal ne s'en sai-
r vn poste si auan-
es principaux ha-
gers, des motifs
comme ce n'estoit
Declarations du
& executer les
out ce qu'il y a eu
de l'Estat & au
t approuuée : Il
Boylesue Lieu-
te sa cabale qui
esté aux peuples
qu'ils n'y pou-
Mr le Cardinal
es avec luy pour
consequenc de
nir ont obtenu
achet du Roy,
nne par toutes
esté heureuse-
u qu'il estoit de
Pour cet ef-
fet enant genet
residial au Pas-
ndé à tous les
ire de mesme,
on non seule-
ment

ce, si le Prince de Conty, poussé par les mesmes sen-
timens que son frere, n'eust iugé qu'il estoit à propos
d'arrêter le cours de ce mal en sa haissance, & que la
diligence fait d'ordinaire la meilleure partie des con-
quêtes.

Son Al. ne fut pas arriver à Agen, qu'ayant dissipé
& par sa presence & par l'adresse de sa conduite, quel-
ques petits nuages qui troubloient le serain de cette
Ville, il donna tous les ordres pour le siege de Cau-
decoste, & fit auancer ses troupes jusques au devant
de la Ville, où d'abord les soldats, allumez par vne
genereuse impatience de faire voir en leur premier
essay ce qu'ils deuoient faire esperer pour l'avenir,
s'attacherent si vigoureusement à vne demy-Lune,
qu'elle fut presque aussi-tost gaignée qu'attaquée, la-
quelle ils abandonnerent foudain, soit qu'elle ne
peut pas estre gardée, ou par le desplaisir qu'ils auoient
de se voir postez en vn trauail qui auoit mis dans le
peril par des blesseures mortelles, quoy que glorieu-
ses, la vie des sieurs de Barbeziere, Gensiac & Monde,
sous la conduite desquels ils auoient fait cet exploit.

Cependant le P. de Conty, qui employe utilement
tous ses momens, ayant reconnu dans vn autre en-
droit le foible de cette Place, y fit dresser vne batte-
rie dès le lendemain, & se logea ce jour mesme sur le
bord du fossé, lequel il eut passé en cette même nuit
pour attacher le minceur à la muraille, si le sieur de
Bourgoigne Mareschal de bataille, n'eust remarqué
qu'il n'estoit pas soustenu par le Regiment de Gon-
drin, comme il en auoit deceulé ordre, ce qui ébola

es mesmeſ ſen-
eſtoit à propos
nce, & que la
partie des con-

u'ayant diſſipé
conduite, quel-
ſerain de cette
ſiege de Cau-
ques au deuant
lumez par vne
i leur premier
pour l'auenir,
e demy-Lune,
u'attaquée, la-
oit qu'elle ne
r qu'ils auoient
it mis dans le
que glorieux
ſac & Monde,
it cét exploit.
oye utilement
vnu autre en-
ſer vne batte-
meſme ſur le
meſme huis
ſi leſeur de
t remarqué
entide Gonq
e qui donra
gle

B.2 Exemple 2

La numérisation de la mazarinade suivante témoigne également des répétitions de pages². La main ayant numérisé n'a pas été supprimée avant la mise en ligne du document. L'extrait suivant en témoigne.

2. Voir <https://books.google.fr/books?id=fL1vR-ja5xgC&hl=fr>

Et de ne plus parler enfin
De frondeur ny de Mazarin,
Qui sont des de discorde,
Sous peine, non pas de la corde
Du coutelas, ny de la hant
Mais de mal souper chez Renaud.

F I N.

Et de ne plus parler enfin
De frondeur ny de Mazarin,
Qui sont des de discorde,
Sous peine, non pas de la corde
Du coutelas, ny de la hant,
Mais de mal souper chez Renard.

F I N.

Et de ne plus parler enfin
De frondeur ny de Mazarin,
Qui sont des de discorde,
Sous peine, non pas de la corde.
Du coutelas, ny de la hant,
Mais de mal souper chez Renard.

F I N.

Annexe C

Organisation des fichiers dans l'application Antonomaz

La capture d'écran ci-dessous présente l'organisation interne de l'application.

- Le répertoire *data* contient les fichiers XML-TEI dans lesquels les informations des mazarinades sont puisées.
- Les fichiers .xql (aussi contenus dans *modules*) sont des fichiers de transformation permettant le bon fonctionnement de l'application.
- Le dossier *resources* contient des fichiers utiles à la mise en page du site, c'est-à-dire les feuilles de style CSS, les images (logos du projet, des partenaires...) ou encore les schémas de validation ODD des fichiers XML-TEI et de l'application.
- Le dossier *templates* contient les pages HTML.

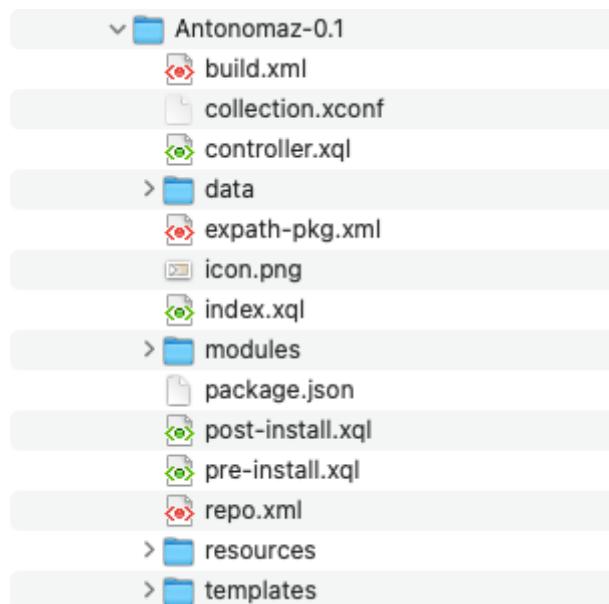


FIGURE C.1 – Structure de l'application

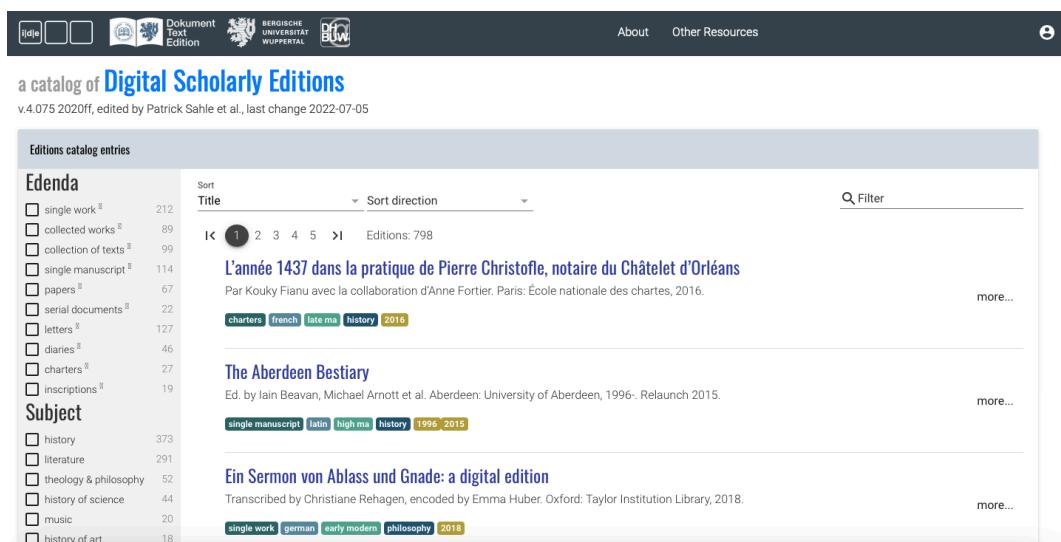
Annexe D

Exemples de site TEI Publisher

D.1 La structure commune des sites

Les sites disposent d'une organisation proche : barre latérale de menu avec lien vers l'accueil, la collection, une logique de visionnage avec alignement page de texte et numérisation... un site TEI Publisher se repère. L'identité du site se joue donc au niveau de l'habillage et de la personnalisation des éléments.

D.2 Exemples de pages de sites TEI Publisher



The screenshot shows the homepage of the Digital Scholarly Editions catalog. At the top, there is a dark header bar with icons for home, search, and user profile, followed by the text "Dokument BERGISCHE UNIVERSITÄT WÜPFERLT DTW". Below the header, the title "a catalog of Digital Scholarly Editions" is displayed, along with the note "v.4.075 2020ff, edited by Patrick Sahle et al., last change 2022-07-05". The main content area has a light gray background. It features a sidebar on the left with sections for "Edenda" (listing categories like "single work", "collected works", etc.) and "Subject" (listing categories like "history", "literature", etc.). The main content area includes a search bar, a sort dropdown, and a filter button. It lists several entries, each with a title, author, publication details, and a "more..." link. For example, the first entry is "L'année 1437 dans la pratique de Pierre Christofle, notaire du Châtelet d'Orléans" (Par Kouky Fianu avec la collaboration d'Anne Fortier. Paris: École nationale des chartes, 2016). Other entries include "The Aberdeen Bestiary" (Ed. by Iain Beavan, Michael Arnott et al. Aberdeen: University of Aberdeen, 1996-. Relaunch 2015) and "Ein Sermon von Ablass und Gnade: a digital edition" (Transcribed by Christiane Rehagen, encoded by Emma Huber. Oxford: Taylor Institution Library, 2018).

FIGURE D.1 – Page d'index du projet Digital Scholarly Editions

Source : <https://www.digitale-edition.de/exist/apps/editions-browser/\protect\TU\textdollarapp/index.html>

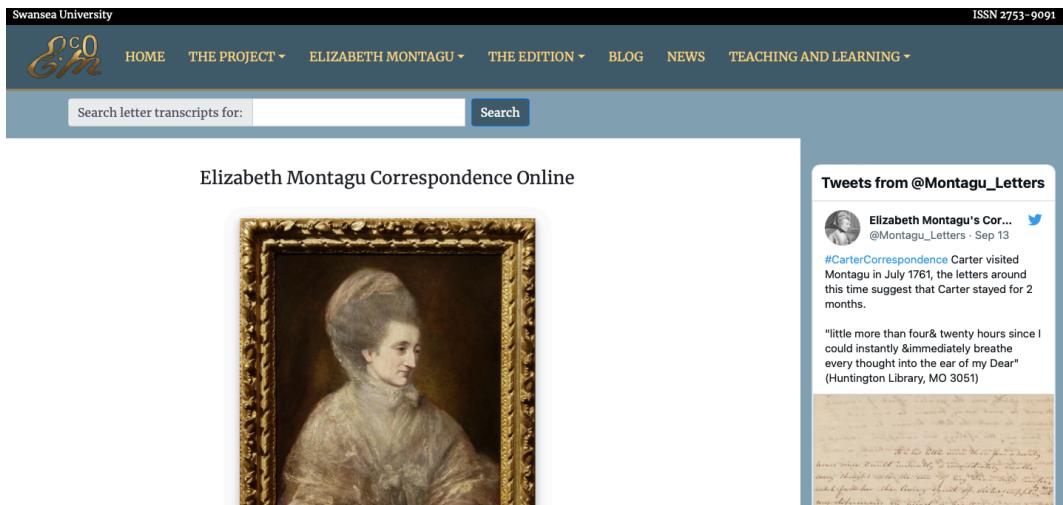


FIGURE D.2 – Page d'accueil du projet EMCO

Source : <https://emco.swansea.ac.uk/home/>

D.3 Utilisation de Mirador par le projet Démêler le cordel

Le projet propose un lien externe vers la visionneuse Mirador par un bouton indiquant la possibilité de comparer le document avec d'autres sources.

Exemple tiré du lien suivant : https://desenrollandoelcordel.unige.ch/Pliegos/Moreno_319.xml

FIGURE D.3 – Intégration de Mirador sur la page de consultation

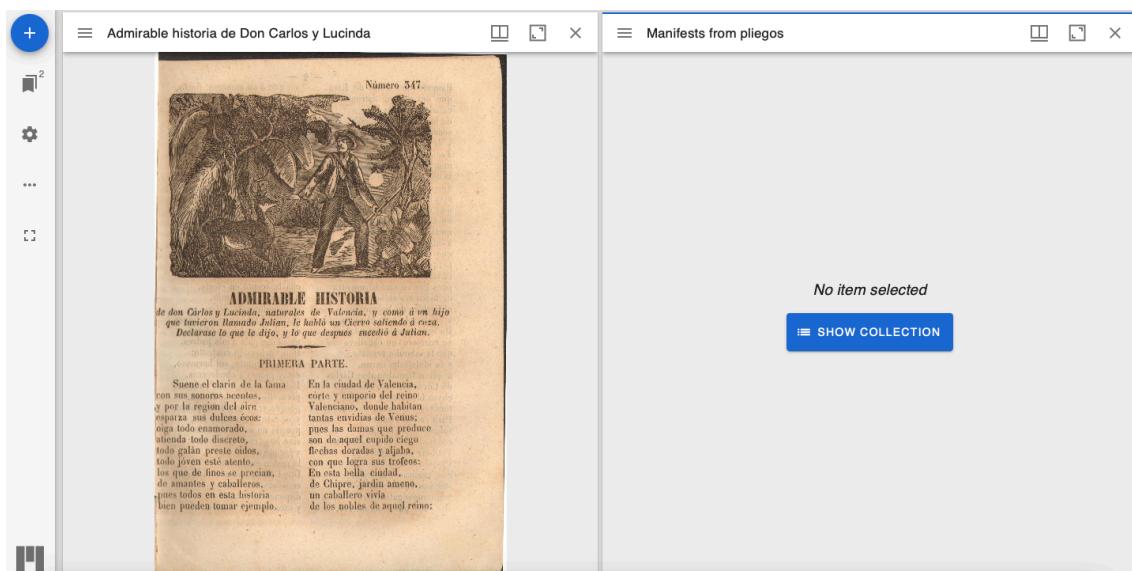


FIGURE D.4 – Affichage de Mirador après le clic

Annexe E

Livrables

Les annexes sont accessibles à l'adresse suivante : <https://github.com/margauxfre/MemoireTNAH-2022/tree/main/Livrables>

Le dossier *Nettoyage-PDFs* contient le script réalisé pour le nettoyage des PDFs ainsi que deux exemples de PDFs Google avant et après traitement. Le Moreau495 représente un exemple complètement satisfaisant du nettoyage tandis que le Moreau692 comporte encore une page dupliquée.

Le dossier *ManifestIIIF* contient le script de création des manifestes IIIF ainsi que deux exemples de fichier.

Le dossier *ODD* centralise le travail effectué sur la révision de l'ODD des fichiers XML-TEI. La documentation est accessible au format HTML. Deux scripts ont été réalisés pour ajouter ou modifier automatiquement des éléments sur tous les fichiers.

Enfin, le dossier *TEI-Publisher* contient l'ODD réalisé pour la structuration du site (état arrêté à la fin du stage), ainsi que le travail réalisé sur la page HTML de consultation des documents (*view.html*). Comme les deux fichiers sont adaptés à la logique de TEI Publisher, ils sont difficilement lisibles en dehors de l'application. Pour y accéder, nous conseillons leur ouverture en texte brut.

Toutes les données produites par le projet sont accessibles par le lien suivant : <https://github.com/Antonomaz>

Table des figures

1.1	Extrait de la <i>Bibliographie des Mazarinades</i>	16
1.2	Extraits de la <i>Supplément à la Bibliographie des Mazarinades</i> , p. 15	17
1.3	<i>Le salut de la France dans les armes de la ville de Paris</i> , Paris, 1649.	20
2.1	Extrait de la numérisation et de l'océrisation du Moreau3707	30
2.2	Extrait de la numérisation et de l'océrisation du Moreau3780	30
3.1	Schéma du traitement opéré sur l'image	40
3.2	Sélection des zones sur une page de document	42
4.1	Image IIIF affichée dans son entièreté	47
4.2	Image IIIF zoomée sur un point d'intérêt	48
4.3	Structuration du manifeste IIIF	49
4.4	Schéma du chemin ARK dans l'URI	53
4.5	Extrait d'un fichier au format JSON	54
5.1	Notes additionnelles dans un fichier XML-TEI	70
5.2	Attribution d'un prédicat dans TEI Publisher	71
5.3	Définition d'un <i>template</i> dans TEI Publisher	71
5.4	Définition du mode d'affichage dans TEI Publisher	72
5.5	Intégration de l' <i>abstract</i> pour l'affichage des métadonnées	73
5.6	Résultat HTML des éléments sélectionnés	73
6.1	Interface au chargement de la visionneuse IIIF	78
6.2	Zoom de la visionneuse IIIF sur le bas droit du document	78
6.3	Modèle défini pour reconstruire les liens des manifestes IIIF hébergés par la Bibliothèque Mazarine	81
6.4	Schéma des étapes pour le chargement du manifeste IIIF dans Mirador .	82
6.5	Structure de la page de consultation des documents	84
6.6	Index du corpus des mazarinades disponibles	87
6.7	Page de consultation d'une mazarinade (visionneuse IIIF)	87

6.8	Page de consultation d'une mazarinade (métadonnées)	89
C.1	Structure de l'application	115
D.1	Page d'index du projet Digital Scholarly Editions	117
D.2	Page d'accueil du projet EMCO	118
D.3	Intégration de Mirador sur la page de consultation	118
D.4	Affichage de Mirador après le clic	119

Table des matières

Résumé	iii
Remerciements	v
Liste des sigles et abréviations	vii
Bibliographie	ix
Introduction	3
I Le projet Antonomaz et les mazarinades : présentation du corpus et de sa chaîne de traitement	7
1 Le corpus des mazarinades : l'expression des frondeurs et des anti-frondeurs	9
1.1 Contexte historique : La Fronde	9
1.1.1 De la fronde à la Fronde	9
1.1.2 La figure de Mazarin	10
1.1.3 Déroulé des évènements	11
1.2 Qu'est-ce qu'une mazarinade ?	13
1.2.1 Origine et signification du mot	13
1.2.2 L'hétérogénéité des textes	14
1.2.3 Bibliographes et constitution du corpus	15
1.2.4 Bibliographies et identification numérique des mazarinades	18
1.3 La production et circulation en masse d'imprimés	19
1.4 La conservation des documents	19
1.4.1 Aspects matériels	19
1.4.2 Lieux de conservation des mazarinades	22

2 Préparer et nettoyer les (méta)données par le traitement de masse	25
2.1 Repérer et collecter les documents numérisés	25
2.1.1 S'adapter à la matière	25
2.1.2 Procédure mise en place	26
2.1.3 Bilan	27
2.1.4 Nommer les fichiers	27
2.2 La TEI et le premier objectif scientifique du projet : analyser automatiquement le texte	28
2.2.1 L'océrisation des textes	28
2.2.2 Où stocker le résultat ?	29
2.3 Assurer la qualité des métadonnées	31
2.3.1 Générer semi-automatiquement les fichiers	31
2.3.2 Comment s'assurer de la qualité de l'encodage ?	32
2.3.3 Pourquoi contraindre ?	33
II Réalisation technique autour du traitement de l'image : l'opportunité IIIF	35
3 Le nettoyage des PDFs des numérisations Google BOOKS	37
3.1 Pourquoi nettoyer ?	37
3.2 Comment nettoyer ?	38
3.2.1 La librairie python imagededup	38
3.2.2 Logique du script	39
3.3 Premiers résultats sur le traitement de numérisations d'imprimés anciens	41
3.3.1 Adapter le script aux résultats : le découpage en zones	41
3.3.2 Trouver le bon degré de similarité	42
3.4 Limites de la procédure	43
4 Le standard IIIF pour l'hébergement des images	45
4.1 Standards et intéropérabilité de l'hébergement image	45
4.1.1 Qu'est-ce que le IIIF ?	45
4.1.2 Présentation de la chaîne technique de IIIF	48
4.1.3 La communauté IIIF	50
4.2 Production des manifestes à partir des métadonnées des XML-TEI	51
4.2.1 Où héberger les images ?	51
4.2.2 Ce que doit contenir le manifeste	54
4.2.3 Créer un script python écrivant les manifestes	56

4.3	La valorisation et la réutilisation des données et métadonnées IIIF	61
4.3.1	Regrouper numériquement le corpus	62
4.3.2	Exploiter le corpus	63
III	La construction d'une page de consultation pour chaque mazarinade	65
5	Mettre en ligne le corpus avec TEI Publisher	67
5.1	TEI Publisher : un outil pour l'édition numérique	67
5.1.1	Qu'est-ce que TEI Publisher ?	67
5.1.2	Une communauté grandissante	69
5.1.3	La personnalisation de l'affichage des éléments XML-TEI	70
5.2	Choisir son outil de mise en ligne	73
5.2.1	Les options possibles	74
5.2.2	Antonomaz et le choix de TEI Publisher	75
6	Développement applicatif : la construction de la page HTML dédiée à l'affichage du document	77
6.1	L'implémentation de la visionneuse IIIF Mirador sur TEI Publisher	77
6.1.1	Qu'est-ce que Mirador ?	77
6.1.2	Pourquoi Mirador ?	79
6.1.3	Mise en place sur TEI Publisher	80
6.2	Organisation de la page web	84
6.2.1	Penser l'affichage des données	84
6.2.2	Présentation du résultat	86
6.2.3	Fonctionnalités futures	89
Conclusion		93
Annexes		97
A	Exemple d'encodage XML-TEI	97
A.1	Numérisation de la page de titre	98
A.2	Encodage de la page de titre en XML-TEI	99
A.3	Encodage du <i>teiHeader</i>	100

B Exemple des diverses qualités des PDFs GoogleBooks	103
B.1 Exemple 1	103
B.2 Exemple 2	108
C Organisation des fichiers dans l'application Antonomaz	115
D Exemples de site TEI Publisher	117
D.1 La structure commune des sites	117
D.2 Exemples de pages de sites TEI Publisher	117
D.3 Utilisation de Mirador par le projet Démêler le cordel	118
E Livrables	121