Humanités numériques et science ouverte
Les collectifs

Clarisse Bardiot, Esther Dehoux, Émilien Ruiz (dir.)

La fabrique numérique des corpus en sciences humaines et sociales

Centrales pour toutes les disciplines relevant des arts, des lettres et des sciences humaines et sociales, les questions relatives à l'identification, la sélection, le classement, les modalités d'exploitation et de diffusion des matériaux nécessaires à la production de connaissances ne sont pas nées avec l'ère dite « numérique ». En histoire, pour prendre l'exemple qui nous est le plus familier, le rapport à la documentation fut ainsi d'emblée au cœur des réflexions méthodologiques qui ont accompagné la construction des savoirs historiques en discipline et l'émergence du métier d'historien. Dès 1898, dans leur *Introduction aux études historiques*, Charles-Victor Langlois et Charles Seignobos formalisent les opérations qui composent ——

Septentrion
PRESSES UNIVERSITAIRES

Phœbus e-Balzac : édition numérique exhaustive d'un monument littéraire

Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

Le contexte de recherche

1. « Avant de concevoir *La Comédie humaine* et de se battre avec l'état civil, Balzac s'est battu avec le roman de son temps. [...] La création n'est pas le prix d'une victoire du romancier sur la vie, mais sur le monde de l'écrit dont il est habité », écrivait André Malraux (1977, 155). Cette affirmation trouve sa réalisation quelques années plus tard grâce à l'avènement du numérique et du projet e-Balzac coordonné par Andrea Del Lungo[1].

2. Le site ebalzac.com, ouvert en avril 2017, propose une édition électronique de *La Comédie humaine* d'Honoré de Balzac en libre accès et dans une version inédite en ligne, ainsi que des outils d'interrogation textuelle, comme un moteur de recherche lexicale ou un comparateur de versions. L'objectif est de valoriser le patrimoine écrit français en employant des méthodes parmi les plus innovantes dans le domaine des humanités numériques. Ce site constitue une première réalisation du projet *Phœbus* (Projet d'hypertexte de l'œuvre de Balzac reposant sur l'utilisation de similarités), financé par l'ANR (Agence nationale de la recherche) pour la période 2015-2019[2].

3. La partie principale de ce projet éditorial consiste en la numérisation d'une quantité importante d'œuvres. *La Comédie humaine* est un monument littéraire composé de 95 textes, dont le projet e-Balzac ambitionne la mise en ligne, dans une version philologiquement exacte, des différents états imprimés publiés du vivant de l'auteur. Cependant, la spécificité du corpus balzacien repose sur la multiplication par l'auteur des supports de publication (livres, volumes collectifs, feuilletons) et la réutilisation de ses textes antérieurs. La numérisation de différentes versions est donc un défi de taille, tant pour l'établissement d'une chaîne de traitement efficace que pour la mise en œuvre du contrôle de la qualité des données. La formation numérique des acteurs engagés au sein du projet au fil des années est variable et nécessite souvent le recours à des outils intuitifs et faciles à prendre en

1. Voir le chapitre d'Andrea Del Lungo et Karolina Suchecka dans cet ouvrage : « Le projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles ». Pour consulter les données mobilisées dans le chapitre, voir https://hnso-corpus.nakala.fr/.

2. Ce projet est porté par les équipes CELLF (Centre d'études de la langue et de la littérature françaises) et LIP6 (Laboratoire d'informatique de Paris 6) de Sorbonne Université et par l'équipe ALITHILA (Analyses littéraires et histoire de la langue) de l'université de Lille.

main, que ce soit pour l'établissement du texte ou pour son exportation dans les formats HTML ou EPUB.

Le protocole de production des données

4. Les données conçues au sein du projet peuvent être divisées en trois sous-parties :

1. Le corpus principal constitué des différents états imprimés des œuvres de Balzac publiés du vivant de l'auteur (la dernière édition retenue étant celle dite « Furne corrigé »)
2. Le corpus comparatif qui regroupe les fichiers combinant deux versions d'un texte
3. Le corpus secondaire des auteurs contemporains ou antérieurs à Balzac

5. Pour chaque sous-ensemble, ainsi qu'au sein de ces derniers, la méthode d'acquisition des données varie en fonction de l'accessibilité des œuvres. Tandis que, pour le corpus principal, la totalité de la chaîne du traitement, à commencer par la numérisation du fac-similé, a été prise en charge au sein du projet, une partie des œuvres composant le corpus secondaire a pu être numérisée semi-automatiquement à partir des formats déjà structurés, comme le format adaptable Daisy DTBook, disponible sur Gallica pour certaines œuvres les plus connues, et le format EPUB mis à disposition par des bibliothèques numériques ou d'autres projets de recherche, comme le projet ANR *Chapitres*[3]. Quant au corpus comparatif, il a été établi automatiquement à partir des documents composant le corpus principal.

6. Les attentes concernant la qualité des données étaient aussi variables. Afin d'atteindre l'exactitude philologique de chaque texte pour le corpus principal, leurs établissements ont été très rigoureusement suivis, tandis que quelques erreurs d'océrisation du corpus secondaire ont été acceptées, dans la mesure où ce dernier est destiné à l'exploitation avec le logiciel de détection automatique des réutilisations textuelles TextPAIR (Del Lungo et Suchecka 2022) et non pas à la mise en ligne au sein de l'édition numérique. Nous décrivons ci-dessous la chaîne du traitement mise en place pour le corpus principal, dont l'édition des sources est désormais disponible sur l'entrepôt Nakala[4].

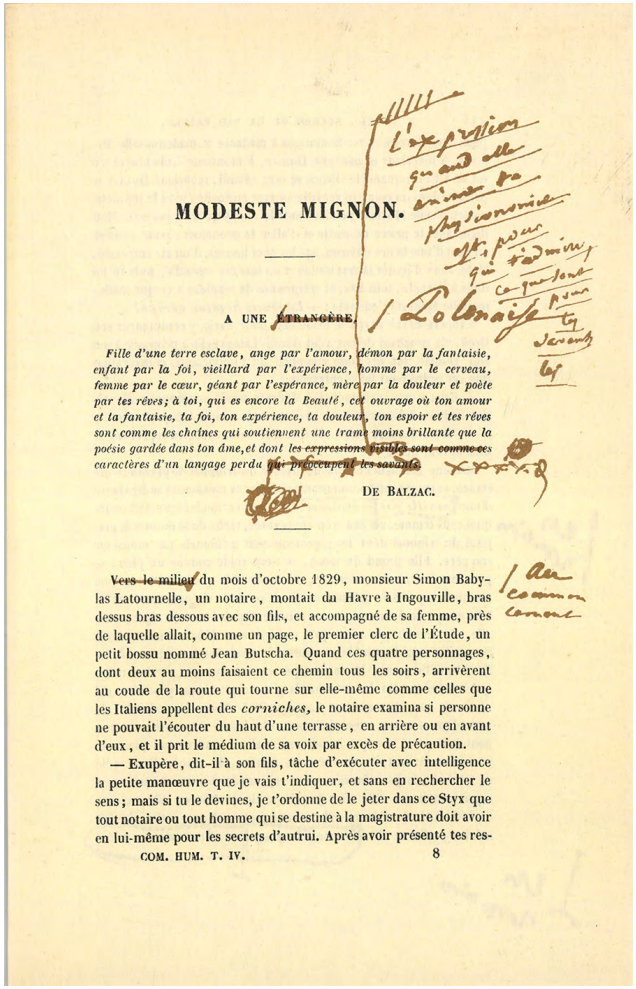
L'établissement du texte

7. La version de référence de *La Comédie humaine*, considérée comme le dernier état de l'œuvre conforme à la volonté de l'auteur, est celle appelée « Furne corrigé » (ci-après : FC). Son statut est particulier puisqu'elle intègre les corrections apportées par Balzac sur son

3. Pratiques et poétiques du chapitre du 19e au 21e siècle : génétique, rhétorique de la lecture et transmédialité – CHAPITRES, Aude Leblond (dir.), THALIM (Théorie et histoire des arts et des littératures de la modernité), Sorbonne Université, 2015-2018.

4. Cf. https://nakala.fr/ et « § La description du jeu de données ».

exemplaire personnel de la dernière édition imprimée de son vivant (figure 1), celle du Furne (ci-après : F).



MODESTE MIGNON.

A UNE ÉTRANGÈRE.

Fille d'une terre esclave, ange par l'amour, démon par la fantaisie, enfant par la foi, vieillard par l'expérience, homme par le cerveau, femme par le cœur, géant par l'espérance, mère par la douleur et poëte par tes rêves; à toi, qui es encore la Beauté, cet ouvrage où ton amour et ta fantaisie, ta foi, ton expérience, ta douleur, ton espoir et tes rêves sont comme les chaînes qui soutiennent une trame moins brillante que la poésie gardée dans ton âme, et dont les expressions fidèles sont comme ces caractères d'un langage perdu qui préoccupent les savants.

DE BALZAC.

Vers le milieu du mois d'octobre 1829, monsieur Simon Babylas Latournelle, un notaire, montait du Havre à Ingouville, bras dessus bras dessous avec son fils, et accompagné de sa femme, près de laquelle allait, comme un page, le premier clerc de l'Étude, un petit bossu nommé Jean Butscha. Quand ces quatre personnages, dont deux au moins faisaient ce chemin tous les soirs, arrivèrent au coude de la route qui tourne sur elle-même comme celles que les Italiens appellent des *corniches*, le notaire examina si personne ne pouvait l'écouter du haut d'une terrasse, en arrière ou en avant d'eux, et il prit le médium de sa voix par excès de précaution.

— Exupère, dit-il à son fils, tâche d'exécuter avec intelligence la petite manœuvre que je vais t'indiquer, et sans en rechercher le sens; mais si tu le devines, je t'ordonne de le jeter dans ce Styx que tout notaire ou tout homme qui se destine à la magistrature doit avoir en lui-même pour les secrets d'autrui. Après avoir présenté tes res-

COM. HUM. T. IV. 8

Figure 1. Exemple d'une page de l'édition Furne

Annotation manuelle de Balzac

Source : https://www.ebalzac.com/romans/05-modeste-mignon/furne-corrige/scans/p113.jpg Licence CC-BY-NC-ND-4.0

8. Cette spécificité pose des problèmes quant à l'emploi des logiciels d'océrisation. Les annotations manuelles, situées généralement en marge et éloignées des passages qu'elles modifient, sont difficilement lisibles, même pour un œil humain. Les ratures et les signes de correction orthotypographique, quant à elles, empêchent la bonne reconnaissance de caractères imprimés. Le travail d'océrisation a donc débuté par la version non annotée de l'édition F, disponible sur Gallica tant en format image qu'en format texte[5]. Alors que l'édition numérique de la version FC reste inédite avant la création du projet e-Balzac, la F est très facilement accessible en ligne, que ce soit au sein des éditions numériques dédiées à Balzac (notamment celle réalisée par le Groupe international de recherches balzaciennes, la mairie de Paris et l'université de Chicago en 2004[6]), sur Wikisource[7], voire au sein des répertoires mettant à disposition des œuvres du domaine libre en format EPUB[8].

9. La disponibilité des œuvres en format texte permet d'accélérer la chaîne du traitement de manière importante. Cependant, elle présente un risque lié à la qualité des données mises à disposition. Pour le texte disponible sur Gallica, par exemple, le taux d'erreur d'océrisation

5. Cf. catalogue.bnf.fr/ark:/12148/cb30051006q
6. Cf. https://www.maisondebalzac.paris.fr/vocabulaire/furne/protocole.htm
7. Cf. https://fr.wikisource.org/wiki/La_Com%C3%A9die_humaine. L'édition citée est celle d'Alexandre Houssiaux, publiée de manière posthume, mais sans reprendre les corrections manuscrites de Balzac. Le texte est donc identique à celui de la version Furne (sauf erreurs d'édition).
8. Cf. par exemple https://www.ebooksgratuits.com/ebooks.php.

demeure important, alors que celui de Wikisource risque de contenir de nombreuses erreurs d'édition[9]. La comparaison de deux versions du texte disponibles en ligne (figure 2) avec le logiciel MEDITE[10] permet l'établissement d'un texte de qualité beaucoup plus satisfaisante.

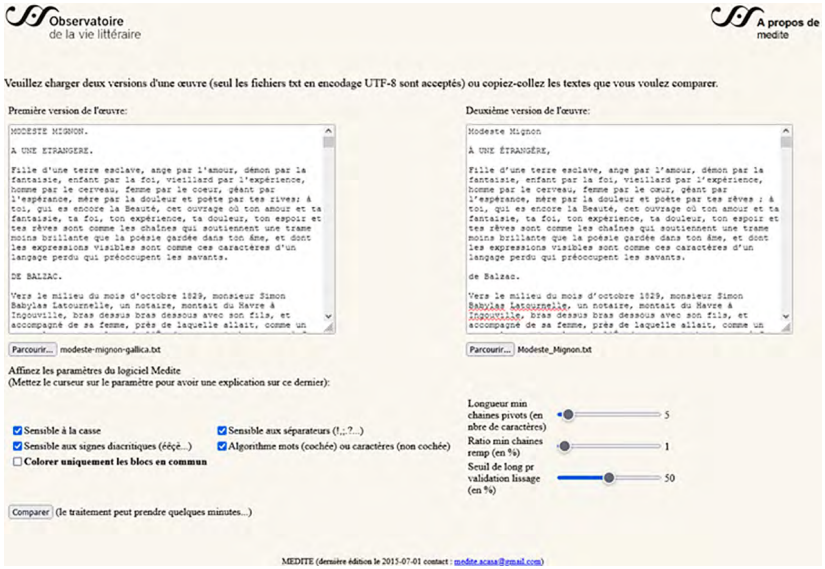


Figure 2. Interface de comparaison avec le logiciel MEDITE
Version texte de Gallica à gauche, version texte de Wikisource à droite
Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

10. MEDITE est un outil de comparaison des versions d'une œuvre qui s'appuie, entre autres, sur l'algorithme de l'alignement par fragments grâce à la détection des homologies (une méthode utilisée initialement pour l'alignement des macromolécules). Les blocs communs sont analysés et les différentes variantes sont signalées grâce à des codes de couleur. Les remplacements sont marqués en bleu, les suppressions en rouge, les insertions en vert et les déplacements en gris. Quelques modifications des paramètres initiaux sont possibles. Par défaut, le traitement est sensible à la casse, aux séparateurs ainsi qu'aux signes diacritiques. La longueur initiale des blocs communs est de cinq caractères. Pour que deux variantes soient considérées comme un remplacement, le ratio de la longueur des deux chaînes repérées doit être supérieur ou égal à 50 % (« abcd » est remplacé par « efgh », mais « a » est supprimé et « efgh » est inséré). Enfin, dans le cas de fortes densités des blocs communs et des variantes, les premiers sont insérés dans la variante si la différence de leur longueur par rapport à la longueur des variantes est supérieure ou égale à 50 %. Pour nos traitements, nous modifions uniquement le ratio des remplacements à 1 %, en considérant que les suppressions et les insertions sont des blocs qui n'ont pas leurs homologues dans l'autre texte. Les autres paramètres gardent les valeurs par défaut.

11. Dans l'exemple présenté en figure 3, nous observons une série de remplacements dus à la mauvaise reconnaissance du texte de la version Gallica (*armis/armés*, *coisi/choisit*, *invetions/inventions*, etc.). Le logiciel permet également de constater plusieurs erreurs typographiques de cette version, comme l'omission des espaces avant

9. Celles-ci sont accumulées par les reprises successives de la version Furne, à commencer par l'édition Houssiaux, jusqu'à celle des utilisateurs engagés dans la mise en ligne du texte.

10. Cf. http://obvil.lip6.fr/medite/ et (Ganascia et Bourdaillet 2006; Ganascia 2011; Ganascia, Glaudes, et Lungo 2014).

les ponctuations doubles (*aigre:*) ou l'ajout de celles-ci avant les ponctuations simples (*pieu ,*). Simultanément, nous comprenons que la version Wikisource contient quelques modernisations des graphies, comme c'est le cas des *falottes* des versions F et FC remplacées par *falotes* chez Houssieux. Ainsi, l'emploi du logiciel rend le travail du contrôle moins laborieux, notamment grâce aux codes couleur qui soutiennent la vigilance du correcteur, ce qui est particulièrement utile par exemple pour la correction des mots outils à graphies proches fréquemment confondus par les logiciels d'océrisation, comme *on* et *ou*.

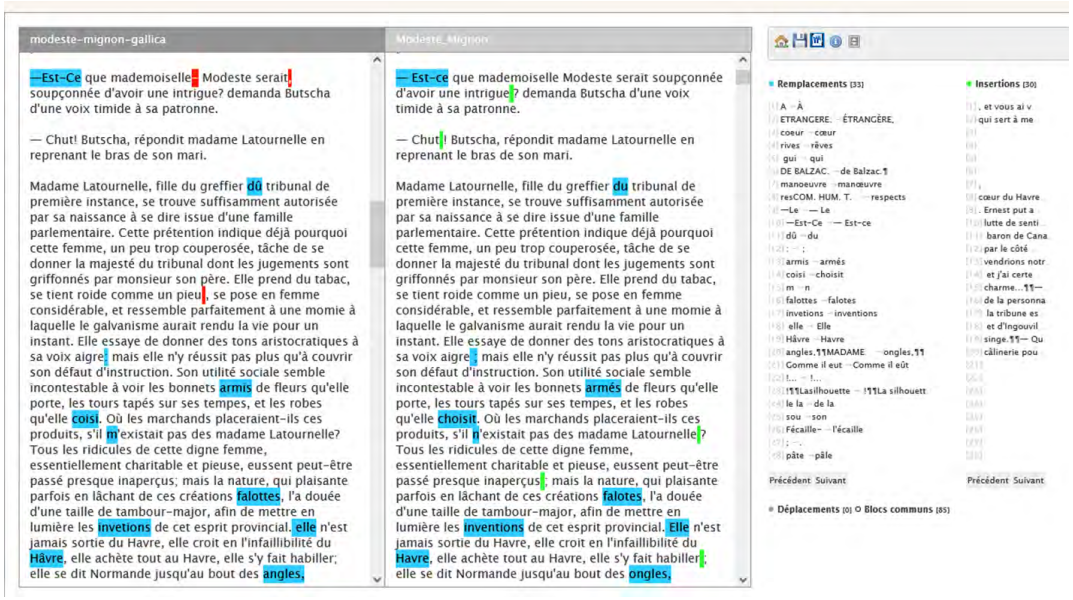


Figure 3. Extrait des résultats de MEDITE

Comparaison entre la version Gallica et celle de Wikisource de *Modeste Mignon*

Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

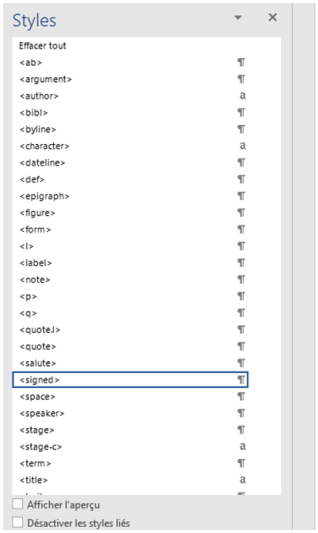
Le stylage dans un logiciel de traitement textuel

12. Après l'établissement d'une version satisfaisante du texte de l'édition F, nous suivons la chaîne de traitement Odette/Teinte (Glorieux 2015) au sein du laboratoire OBVIL (Observatoire de la vie littéraire[11]). Appuyée sur une structure restreinte issue du standard XML-TEI, Teinte, une feuille de style dédiée[12] a été proposée pour permettre de travailler avec les logiciels de traitement de texte. Même si les outils WYSIWYG[13] sont peu performants pour les structures complexes, ils ont l'avantage de faciliter l'intégration dans la chaîne éditoriale des acteurs externes – stagiaires, doctorants, vacataires ou chercheurs expérimentés –, sans qu'ils soient initiés au langage XML.

13. Optimisée pour les textes romanesques, cette méthode compose avec les fonctionnalités initiales des outils (mise en italique, style du paragraphe par défaut) et les noms de styles particuliers (mis entre chevrons) permettant l'annotation sémantique d'un bon nombre d'éléments textuels (citations, illustrations, correspondance, poèmes, etc.). Le développement est accompagné d'un guide

11. Actuellement ObTIC (Observatoire des textes, des idées et des corpus).
12. Pour celle utilisée au sein du projet e-Balzac, cf. https://sharedocs.huma-num.fr/wl/?id=RuGwKHl5KJPgDWSjXg8tJNBCOCqvIsFc.
13. *What You See Is What You Get*, acronyme qui renvoie aux éditeurs visuels de texte.

d'emploi détaillé[14]. Le nombre restreint des éléments limite l'hétérogénéité des annotations choisies par les éditeurs impliqués dans le processus, dont l'interprétation de la structure textuelle peut varier.



La maison Charles Mignon suspend ses payements. Mais les liquidateurs soussignés prennent l'engagement de payer toutes les créances passives. On peut, dès à présent, escompter aux tiers-porteurs les effets à terme. La vente des propriétés foncières couvre intégralement les comptes courants.

Cet avis est donné pour l'honneur de la maison et pour empêcher tout ébranlement du crédit sur la place du Havre.

Monsieur Charles Mignon est parti ce matin sur le Modeste pour l'Asie-Mineure, ayant laissé de pleins pouvoirs à l'effet de réaliser toutes les valeurs, même immobilières.

DUMAY (*liquidateur pour les comptes de banque*) ; LATOURNELLE, (*liquidateur pour les biens de ville et de campagne*) ; GOBENHEIM (*liquidateur pour les valeurs commerciales*).

[p. 133] Latournelle devait de sa fortune à la bonté de monsieur Mignon, qui lui prêta cent mille francs, en 1817, pour acheter la plus belle Étude du Havre. Ce pauvre homme, sans moyens pécuniaires, premier clerc depuis dix ans, atteignait alors à l'âge de quarante ans et se voyait clerc pour le reste de ses jours. Il fut le seul dans tout le Havre dont le dévouement pût se comparer à celui de Dumay ; car Gobenheim profita de la liquidation pour continuer les relations et les affaires de monsieur Mignon, ce qui lui permit d'élever sa petite maison de banque.

Figure 4. Application des styles. Exemple des éléments <q> et <signed> dans *Modeste Mignon*
Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

14. En figure 4[15], les bordures entourent le texte d'un article du *Courrier du Havre*. L'enchâssement d'autres textes dans le récit, notamment d'articles et de lettres, est très fréquent dans *La Comédie humaine*. Différenciés du corps du texte par une mise en page spécifique[16], ces éléments pourraient être considérés comme des citations, et donc à signaler, en suivant le standard XML-TEI Teinte, par <quote>. Or, du point de vue philologique, une citation est définie comme une partie de texte externe à l'œuvre, alors que, chez Balzac, il s'agit souvent d'extraits fictionnels, comme des lettres écrites par des personnages ou encore des articles fictifs tirés de journaux bien réels[17]. Nous choisissons donc de les signaler avec l'élément <q>, qui, dans le standard TEI : « contient un fragment qui est marqué (visiblement) comme étant d'une manière ou d'une autre différent du texte environnant, pour diverses raisons telles que, par exemple, un discours direct ou une pensée, des termes techniques ou du jargon, une mise à distance par rapport à l'auteur, des citations empruntées et des passages qui sont mentionnés, mais non employés[18]. »

15. Pour répondre aux besoins spécifiques du projet e-Balzac, le schéma Teinte a modifié cette définition en précisant que « le contenu de <q> ne peut pas être attribué à une origine extérieure au texte (origine fictionnelle ou non

14. Cf. https://obvil.github.io/Teinte/teinte.html

15. Pour consulter le document complet en format DOCX, cf. https://sharedocs.huma-num.fr/wl/?id=BwvWJnbdnROWwND1xGBP42YCzoCJnHxK.

16. Dans l'exemple à la figure 4, le texte de l'article est entouré des guillemets et séparé du corps du texte par deux lignes horizontales, cf. https://gallica.bnf.fr/ark:/12148/bpt6k6116517k/f145.highres.

17. *Le Courrier du Havre* a été publié quotidiennement de 1839 à 1906, cf. https://data.bnf.fr/fr/32751243/courrier_du_havre/.

18. Cf. https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-q.html

identifiable). Exemples : récit enchâssé, lettre insérée, article de presse, acte notarié, poème, traité, axiomes[19]… » La structure interne de l'élément textuel codé par <q> est également balisée à l'aide de styles dédiés. En figure 4, par exemple, les signatures des trois liquidateurs sont marquées avec le style du paragraphe <signed> et leurs noms sont mis en petites capitales.

16. Le texte corrigé et stylé de l'édition F est ensuite repris pour l'établissement de la version FC en intégrant manuellement les annotations de Balzac. Cette méthode, certes chronophage et minutieuse, garantit l'exactitude philologique de la transcription et permet de corriger d'éventuelles erreurs des autres éditions scientifiques appuyées sur ce texte de référence et disponibles notamment sur le marché du livre imprimé. L'éditeur travaille sur un texte auquel les styles ont déjà été appliqués, ce qui lui permet de focaliser toute son attention sur les annotations de Balzac. Cela implique également que l'expert de la main de l'auteur ne doit pas simultanément maîtriser les techniques de l'édition numérique. Enfin, une relecture finale, appuyée par la comparaison avec MEDITE est effectuée avant la transformation vers le format XML[20].

19. Cf. https://obvil.github.io/Teinte/teinte.html#el_q

20. Pour l'établissement des versions antérieures, l'éditeur peut s'appuyer sur le texte d'une édition déjà établie et, par exemple, la comparer avec l'océrisation de la version en cours de préparation.

La transformation automatique en format XML-TEI Teinte avec Odette

17. Le texte stylé dans un traitement de texte et enregistré en format ODT est ensuite transformé en format XML-TEI grâce au logiciel Odette[21] (figure 5). Une fois encore, la mise en place d'un outil intuitif et disponible en ligne augmente l'autonomie des éditeurs sans demander des compétences informatiques particulières.

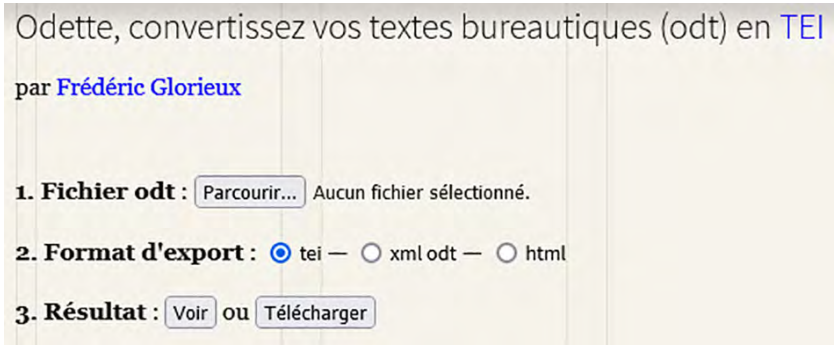


Figure 5. Interface Web de l'outil Odette

Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

18. Grâce à Odette, trois étapes simples (chargement du fichier stylé, choix du format d'export, puis téléchargement du résultat en format choisi) suffisent pour transformer rapidement un document ODT vers le format XML-TEI Teinte. Pour l'extrait textuel présenté en

21. Cf. http://obvil.lip6.fr/Odette/

figure 4, par exemple, nous obtenons la structuration XML suivante :

```
<q type="press">
  <p>La maison Charles Mignon suspend ses payements. [...]</p>
  <p>Cet avis est donné pour l'honneur de la maison [...].
  </p>
  <p>Monsieur Charles Mignon est parti ce matin [...].
  </p>
  <signed>
    <hi rend="sc">Dumay</hi> (<hi rend="i">liquidateur pour les
    comptes de banque</hi>) ;
    <hi rend="sc">Latournelle</hi>, (<hi rend="i">liquidateur pour
    les biens de ville et de campagne</hi>) ;
     <hi rend="sc">Gobenheim</hi>, (<hi rend="i">liquidateur
     pour les valeurs commerciales</hi>).
  </signed>
</q>
<p>
  <pb n="133" xml:id="p133"/>
 Latournelle devait de sa fortune à la bonté de monsieur Mignon,
 [...].
</p>
```

19. Pour les utilisateurs plus expérimentés, le code source du logiciel est disponible en ligne[22]. Le logiciel est conçu principalement à l'aide des feuilles de style XSL et scripts PHP. Il est possible, en l'utilisant localement, d'appliquer la transformation massivement à l'aide d'une ligne de commande exécutée à partir du dossier avec les ressources Odette :

20. user@user:~/odette$ php Odt2tei.php "*.odt" XML/?

21. La variable "*.odt" indique que tous les fichiers en format ODT présents dans le dossier doivent être traités. En revanche, il n'est pas permis de les regrouper dans un sous-dossier : la variable "sous-dossier/*.odt" ne sera pas reconnue. Les résultats peuvent être sauvegardés dans un sous-dossier spécifique (XML/ dans notre exemple).

22. La qualité de la transformation est très satisfaisante (à condition, bien sûr, que le stylage soit fait correctement et en respect du schéma Teinte). Sporadiquement, des balises auto-fermantes <anchor/> peuvent apparaître dans le XML. Il est également conseillé de revoir les éléments de l'emphase avec la requête xPath //hi[@rend='i'] et //emph, car la mise en italiques des longs fragments textuels produit parfois l'accumulation inutile des balises. Le <teiHeader> doit être modifié et complété en fonction des usages de chaque projet.

La transformation vers des formats éditoriaux (HTML et EPUB)

23. Il est également possible, dans la continuité de la chaîne du traitement Odette/Teinte, de générer des fichiers HTML et EPUB à partir du document XML ainsi obtenu. La procédure devient, à ce point, plus complexe, notamment si l'on souhaite utiliser le développement sur un

22. Cf. https://github.com/oeuvres/odette

serveur Web[23]. Toutefois, pour une utilisation basique et locale, les compétences de base en XSLT sont suffisantes.

24. À partir du dossier des sources, le document XML traité doit être transformé à l'aide de la feuille de style « tei2html.xsl ». Pour ce faire, il est possible d'utiliser une ligne de commande, en exécutant, à partir du dossier Teinte :

25. ```
user@user:~/teinte$ xsltproc tei2html.xsl texte.xml > ../HTML/texte.html
```

26. Pour que le fichier HTML ainsi généré puisse être mis en forme (à l'aide d'une feuille de style CSS « tei2html.css »), il doit être enregistré dans un dossier frère de celui contenant les ressources Teinte. La visualisation proposée est basique, mais offre tout de même quelques fonctionnalités facilitant la lecture dans l'environnement numérique (figure 6). Par exemple, le clic sur le numéro de page affiché sur la marge de gauche permet d'afficher le fac-similé correspondant[24], la génération de la table de matières accélère la navigation dans le texte, et la mise en forme des éléments balisés facilite la révision et le contrôle qualité du document.

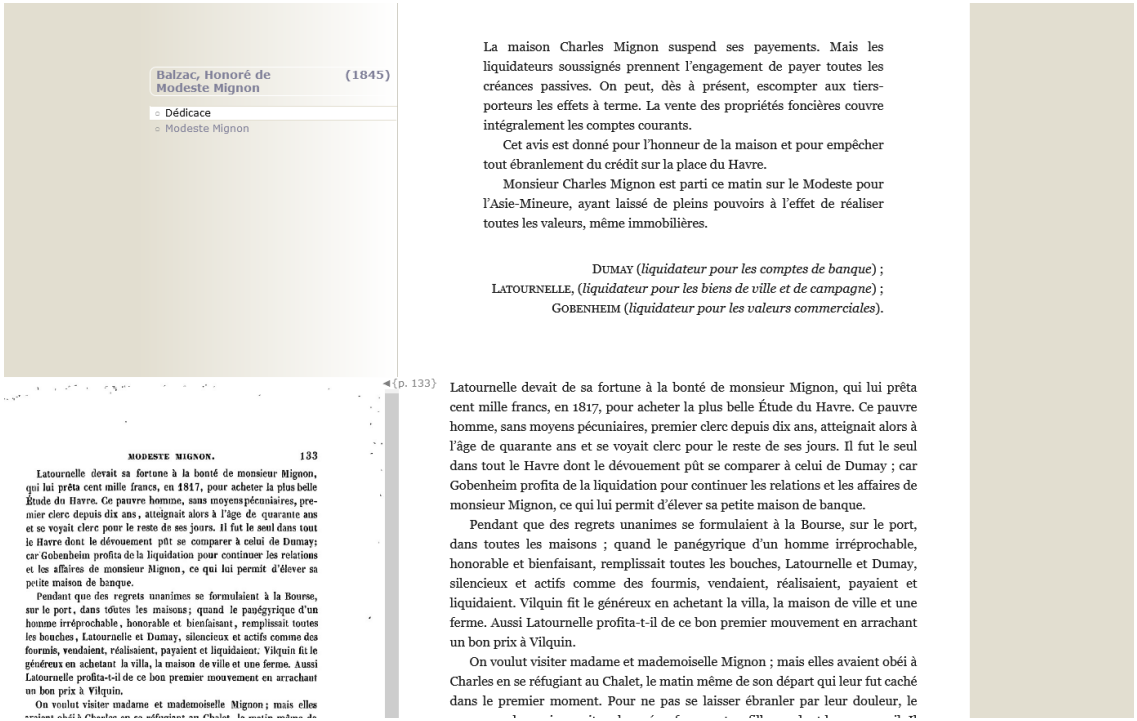



Figure 6. Visualisation HTML avec Teinte. Exemple d'affichage pour *Modeste Mignon*
Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

23. Dans ce cas, nous renvoyons les lecteurs vers la documentation détaillée disponible sur https://github.com/oeuvres/teinte. Pour l'utilisation avec Omeka, cf. https://github.com/oeuvres/Bookmeka.

24. À condition que le lien vers l'image soit renseigné en tant que valeur de l'attribut @facs de l'élément <pb>. Dans l'exemple présenté à la figure 6, la page 133 est déclarée ainsi dans le XML final : <pb n="133" xml:id="p133" facs="https://gallica.bnf.fr/ark:/12148/bpt6k6116517k/f146.highres"/>. Nous ajoutons les attributs @facs automatiquement après la génération du XML, cf. https://nakala.fr/10.34847/nkl.d1066aod.
```

La description du jeu de données

27. Le jeu de données décrit dans cet article a été déposé dans l'entrepôt de données Nakala suivant une modélisation des métadonnées préparée dans le cadre du dépôt. Avant d'aller plus loin dans la description du jeu de données, précisons qu'il s'agit ici d'informations sous format numérique, sorties de leur contexte de recherche, qu'il est nécessaire de traiter, structurer et présenter d'une certaine manière.

Le choix de l'entrepôt

28. Nakala est un service de l'IR* Huma-Num[25] permettant aux chercheurs, enseignants-chercheurs ou équipes de recherche de partager, publier et valoriser tous types de données numériques documentées (fichiers texte, sons, images, vidéos, objets 3D, etc.) dans un entrepôt sécurisé afin de les publier en accord avec les principes du *FAIR data*[26] et des valeurs de la science ouverte. Choisir Nakala comme entrepôt de données permet de s'inscrire dans la logique du Web de données ouvertes (*Linked Open Data*), qui rend possible la connexion à d'autres entrepôts existants, et prolonge le travail sur l'intertextualité autour de l'œuvre de Balzac.

29. Une fois les données publiées sur Nakala, il est possible de les récupérer, moissonner[27], et exposer ailleurs grâce au protocole documentaire OAI-PMH[28], au modèle RDF[29] ou à une API REST[30]. Par ailleurs, Nakala fait partie d'un dispositif cohérent de services mis en place par Huma-Num suivant la chaîne de traitement des données. L'accès, le signalement, la conservation et l'archivage à long terme des données de la recherche en SHS sont facilités par les outils développés par Huma-Num.

30. Nakala est destiné au stockage de données stabilisées (décrites et complètes) et est utilisé par les porteurs de projets une fois le processus de collecte des données terminé. Afin de réaliser une collecte optimale, les projets sont encouragés à :

- mettre en sécurité les données sur un outil de stockage externe (disques durs...)
- ordonner et organiser les données

25. L'Infrastructure de Recherche (IR*) des humanités numériques (anciennement TGIR, Très Grande Infrastructure de Recherche), cf. https://www.huma-num.fr/.

26. Il s'agit des données qui respectent les principes de l'ouverture des données appelés *FAIR* (*Findability*, *Accessibility*, *Interoperability* et *Reusability*).

27. Cela veut dire que les données peuvent être récoltées et incluses dans une base de données regroupant les documents aux références bibliographiques similaires. Seulement les données encodées avec les mêmes procédés techniques peuvent être moissonnées.

28. *Open Archives Initiative – Protocol for Metadata Harvesting*, protocole pour la collecte des métadonnées de l'initiative pour les archives ouvertes, est un dispositif permettant l'échange des métadonnées entre plusieurs institutions qui donnent accès aux documents numériques, cf. http://www.openarchives.org/OAI/openarchives-protocol.html. Il est utilisé, par exemple, par la Bibliothèque nationale de France, cf. https://www.bnf.fr/fr/les-entrepots-oai-de-la-bnf.

29. *Resource Description Framework* est un modèle descriptif des données Web et des métadonnées attachées.

30. Interface de Programmation d'Application (*Representational State Transfer*) est un style d'architecture logicielle.

- consigner toutes les informations disponibles sur les données
- mettre en place un plan de nommage harmonisé[31] des fichiers

31. Nakala accepte tous les types de formats de fichiers pour le dépôt et permet d'en visualiser certains[32]. Les données sont respectées dans leur intégrité et ne subissent pas de modification une fois déposées. De plus, Huma-Num effectue une copie sécurisée des données et des métadonnées au sein de son infrastructure. Celles-ci sont identifiées grâce à l'attribution d'identifiants pérennes[33]. Il est ainsi possible de citer chaque donnée de manière précise et exacte grâce au DOI[34]. Par exemple, l'édition électronique du texte de *Modeste Mignon* parue chez Furne (1845) est connue sous l'identifiant 10.34847/nkl.defelc0z. Il est également possible de citer la référence complète de la donnée (Balzac 2021). Même si cette manière de référencer est encore à retravailler, notamment en ce qui concerne la reprographie des textes imprimés, l'identifiant alphanumérique assure que la citabilité de chaque donnée et chaque collection soit pérenne et aussi exacte que possible. Afin de suivre toutes les recommandations *FAIR*, le projet e-Balzac s'est également assuré que les ressources déposées ne contiennent pas des informations personnelles, dont la diffusion est encadrée par le Règlement général sur la protection des données (RGPD). Par conséquent, il n'est pas nécessaire de mettre en place des procédures de pseudonymisation[35] ou de *mash-up* (application composite) des données. Elles doivent tout de même être mises sous une licence. Celle reprise pour la publication des données se conforme au choix initial effectué lors de la création du site ebalzac.com : il s'agit de la licence Creative Commons Attribution Non Commercial No Derivatives 4.0 International[36] (CC-BY-NC-ND-4.0).

32. Outre la facilitation du respect des principes *FAIR*, d'autres fonctionnalités de Nakala visent à simplifier la gestion collective du dépôt. Tous les utilisateurs pourvus des droits d'édition peuvent déposer dans les collections et modifier les métadonnées avant la publication. Cette gestion fine des droits d'accès aux dépôts, collections et fichiers est un atout de taille pour répartir le travail entre deux personnes n'appartenant pas à la même institution et ne pouvant se voir régulièrement. Les droits sur les données sont partagés entre le déposant[37] (ROLE_OWNER) et les administrateurs ajoutés[38] (ROLE-ADMIN). La vue et l'édition de la donnée dépendent ainsi des rôles attribués par la suite, qui peuvent être modifiés par le déposant. Il est aussi possible d'ajouter des listes d'utilisateurs autorisés à avoir certains rôles.

31. Cela implique que les fichiers sont toujours nommés de la même manière, par exemple : Type_date_numéro.
32. Par exemple l'utilisation de la visionneuse d'images OpenSeadragon (cf. https://openseadragon.github.io/).
33. Depuis 2020, il s'agit du DOI. Auparavant, c'est Handle qui était plus largement utilisé.
34. Le *Digital Object Identifier* permet d'identifier les ressources de manière unique.
35. Moins forte que l'anonymisation, qui empêche totalement et de manière irréversible la possibilité d'identifier une personne, la pseudonymisation permet d'empêcher l'identification d'une personne en fonction des données présentes.
36. Cf. https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode
37. Ici : Victoria Le Fourner.
38. Ici : Karolina Suchecka.

Les métadonnées

33. La modélisation des métadonnées a été effectuée à partir des informations présentes sur chaque page d'une œuvre sur le site ebalzac.com, les informations présentes dans le <teiHeader> des fichiers XML ainsi qu'à partir des métadonnées obligatoires pour chaque dépôt dans Nakala. Chaque donnée déposée peut contenir plusieurs fichiers et est décrite par un certain nombre de métadonnées. La clarté dans l'exposition des métadonnées est fondamentale pour une réutilisation et une compréhension par autrui. Aussi, notre travail insiste sur la précision de la description de chaque donnée. Cinq métadonnées sont rendues obligatoires pour le dépôt : le type de la donnée, le titre, l'auteur, la date de création et la licence (tableau 1). Il est admis qu'un auteur soit anonyme ou qu'une date soit inconnue. Le choix du type de dépôt est restreint par les valeurs présentes au sein d'une liste déroulante. Le type « Édition de sources » convient à la majorité des données du projet. Les choix de mettre Balzac en créateur principal et de renseigner la date de l'édition originale de la source ont été basés sur celles du projet *Nénufar*, sur lequel nous nous sommes appuyés[39]. Nous avons également suivi les recommandations du consortium CAHIER pour les éditions numériques (Galleron et al. 2018) afin de savoir s'il est permis de garder la structuration des informations contenues dans le <teiHeader> ou s'il est préconisé de respecter les usages du standard Dublin Core. La multiplication des mots-clés nous a semblé importante pour un meilleur référencement dans le moteur de recherche de Nakala et pour un moissonnage extérieur.

34. Le choix de mettre dans une donnée tous les types de fichiers générés à partir du XML initial (plutôt que de créer une collection par type de fichier) est motivé par la volonté de faciliter la recherche et l'extraction ciblée des informations par un utilisateur externe. Par ailleurs, le type de fichiers présents dans les données varie en fonction de l'édition de chaque source. Un fichier image (en format PNG) est attaché uniquement aux documents de l'édition FC. Autrement, 171 fichiers HTML sont publiés, 92 EPUB, 179 XML et 93 PNG. Le fichier attendu est le fichier XML, les autres n'étant pas produits pour toutes les éditions. Le volume total des données publiées est de 193,6 Mo. Elles ont été déposées en août 2021 et publiées en septembre 2021. Toutes les données sont disponibles dans l'entrepôt Nakala du projet sous l'identifiant 10.34847/nkl.89fa9i03[40]. Enfin, un tableau qui spécifie, pour chacune des 95 œuvres, son appartenance aux collections et les identifiants DOI des différentes versions éditoriales est disponible dans la collection « e-Balzac »[41].

39. Cf. http://nenufar.huma-num.fr/presentation/

40. Cf. https://nakala.fr/u/collections/10.34847/nkl.89fa9i03

41. Cf. https://nakala.fr/10.34847/nkl.b2afnpoa

42. Dans ce cas spécifique, nous ne renseignons pas de date pour l'édition FC, qui n'a pas été publiée du vivant de Balzac, pour la différencier de celle du F (cf. « § L'établissement du texte »).

Tableau 1. Liste des métadonnées remplies pour chaque dépôt

L'astérisque (*) indique le caractère obligatoire de la métadonnée lors du dépôt.

Nom de la métadonnée	Type	Explication	Exemple (*Modeste Mignon*)
Type de dépôt* nakala:title	Liste	Type de la donnée déposée	Édition de sources
Titre* nakala:type	dcterms:box	Titre de la version déposée	*Modeste Mignon*, dans *La Comédie humaine*, *Études de mœurs*, *Scènes de la vie privée*. Furne corrigé
Auteurs* nakala:creator		Auteur de la source	Honoré de Balzac
Date de création* nakala:created		Date de création de la source. Celle-ci peut-être inconnue.	Inconnue[41]
Licence* nakala:licence	Liste	Licence attribuée sur le dépôt, en lien avec le type de donnée	Creative Commons Attribution Non Commercial No Derivatives 4.0 International
Description dcterms:description	dcterms:Box	Description de l'édition électronique reprise des indications du <teiHeader> du fichier	Cette édition électronique de *La Comédie humaine* constitue la première édition en ligne de l'œuvre de Balzac dans la version dite du « Furne corrigé », qui intègre les corrections manuscrites apportées par l'auteur sur son exemplaire personnel de la première édition de *La Comédie humaine* parue chez Furne de 1842 à 1847.
Mots-clés dcterms:subject	dcterms:Box	Liste de mots-clés destinés au référencement sur Nakala	Édition en ligne ; Édition électronique ; Balzac ; Modeste Mignon ; Comédie humaine ; e-Balzac ; …
Langues	dcterms:language	Langue pour les fichiers de données	Français
dcterms:publisher	dcterms:Box	Éditeur	e-Balzac

Nom de la métadonnée	Type	Explication	Exemple (*Modeste Mignon*)
dcterms:contributor	dcterms:Box	Éditeur(s) scientifique(s)	Directeurs du projet : Andrea Del Lungo, Jean-Gabriel Ganascia et Pierre Glaudes Éditeur : Maxime Perret Correction OCR : Dimitri Julien Établissement du texte et stylage TEI : Maxime Perret Édition XML-TEI : Amélie Canu Informatique éditoriale : Frédéric Glorieux Traitement des images : Claire Carpentier Dépôt Nakala : Victoria Le Fourner et Karolina Suchecka
dcterms:relation	dcterms:URI	Lien vers un document en relation	https://nakala.fr/10.34847/nkl.defelc0z
dcterms:abstract	dcterms:Box	Résumé	Au Havre, Modeste Mignon et sa mère attendent patiemment le retour de Charles Mignon de La Bastie, parti tenter de rétablir sa fortune après une faillite fracassante. En l'absence de son père, Modeste tombe sous le charme de la poésie de Melchior de Canalis et trouve le moyen d'initier une correspondance avec son grand homme. Elle ne se doute pas que c'est le secrétaire du poète, Ernest de La Brière, qui a emprunté le nom de Canalis et qui est le véritable interlocuteur de cet échange épistolaire.
dcterms:available	dcterms:URI	Lien vers la ressource disponible sur le site e-Balzac	https://www.ebalzac.com/edition/05-modeste-mignon/furne-corrige
dcterms:issued	dcterms:Box	Date de publication du fichier électronique	2017

L'organisation en collections

35. En choisissant Nakala comme entrepôt de données, il faut composer avec la spécificité de celui-ci : les collections n'ont pas de niveaux hiérarchiques. En revanche, il est possible de mettre les données dans autant de collections que nécessaire (une collection publique n'acceptant que des données publiées et non déposées[43]). Sans hiérarchisation, il est a priori difficile de rendre compte de la composition conceptuelle de *La Comédie humaine*. En effet, l'œuvre se subdivise en *Études* puis en *Scènes* (et parfois même en diptyques). La subtilité du projet réside également en la présence de plusieurs éditions d'un même texte, qui doivent donc être reliées de manière compréhensible. Cette hiérarchie, bien visible sur le site du projet, est pensée comme une arborescence, alors que Nakala demande de la représenter avec un graphe. Pour ce faire, nous choisissons d'explorer les champs Dublin Core (préfixés par dcterms), qui permettent de reconstituer le lien entre les collections, notamment grâce à dcterms:hasPart, dcterms:isPartOf ou encore dcterms:relation. L'utilisateur peut alors retrouver l'appartenance d'une collection à une autre ou voir le lien entre les textes[44].

36. Ainsi, il est nécessaire de mettre chaque donnée dans différentes collections afin de la restituer clairement dans *La Comédie humaine*. Par exemple, la donnée *Modeste Mignon* (10.34847/nkl.1cffmy50) appartient aux collections « e-Balzac » (10.34847/nkl.89fa9103), « Édition Furne Corrigé » (10.34847/nkl.19e543l9), « Comédie humaine » (10.34847/nkl.bb45t7np), « Études de mœurs » (10.34847/nkl.7d1b0m32) et « Scènes de la vie privée » (10.34847/nkl.da7c094z). La donnée est mise en relation avec une donnée similaire, *Modeste Mignon* de l'édition Furne (10.34847/nkl.defelc0z). Les métadonnées entre elles sont sensiblement les mêmes. Toutefois, ces données ne recouvrent pas la même réalité matérielle. Par conséquent, une description philologique et littéraire est ajoutée à chaque collection afin d'accompagner l'utilisateur au mieux dans l'exploration de la structure complexe de *La Comédie humaine* et dans les différentes éditions des œuvres. Les données sont ainsi disséminées dans 28 collections (tableau 2). La collection « e-Balzac » regroupe directement toutes les ressources produites dans le cadre du projet. Elle est destinée à être reliée avec la collection *Humanités numériques et science ouverte*.

La navigation dans l'entrepôt avec Nakala_Press

37. Une fois l'ensemble des collections publiées, il est possible de mettre en place un site Nakala_Press[45]. Ce site

43. Une donnée déposée est une donnée accessible uniquement au déposant et aux personnes ayant un accès à cette donnée. Une donnée publiée est une donnée visible par n'importe quel utilisateur de Nakala. Si la donnée est visible, cela signifie que ses métadonnées sont visibles, mais le contenu peut être masqué et mis en embargo. Les utilisateurs sont toutefois au courant de son existence.

44. Ces liens sont désormais visibles directement grâce à l'interface Nakala. Ils n'étaient pas directement apparents au moment du dépôt en août 2021.

45. Si le remplaçant de Nakalona n'a pour le moment que peu d'options de stylage et de valorisation des données, il n'est plus lié à Omeka, ce qui permet d'éviter le problème de pérennité lié à la mise à jour et à l'obsolescence des technologies.

Tableau 2. Liste des collections, leurs identifiants DOI et leurs liens pérennes

Nom de la collection	Id Nakala	Lien pérenne vers la collection
e-Balzac	10.34847/nkl.89fa9i03	https://nakala.fr/u/collections/10.34847/nkl.89fa9i03
Comédie humaine	10.34847/nkl.bb45t7np	https://nakala.fr/u/collections/10.34847/nkl.bb45t7np
[Édition] Furne	10.34847/nkl.d3dazxns	https://nakala.fr/u/collections/10.34847/nkl.d3dazxns
[Édition] Furne Corrigé	10.34847/nkl.19e543l9	https://nakala.fr/u/collections/10.34847/nkl.19e543l9
[Édition] Mame	10.34847/nkl.ff47434q	https://nakala.fr/u/collections/10.34847/nkl.ff47434q
[Édition] Béchet	10.34847/nkl.814e3gt6	https://nakala.fr/u/collections/10.34847/nkl.814e3gt6
[Édition de] La Presse	10.34847/nkl.ea325v1w	https://nakala.fr/u/collections/10.34847/nkl.ea325v1w
[Édition du] Constitutionnel	10.34847/nkl.f9182h84	https://nakala.fr/u/collections/10.34847/nkl.f9182h84
[Édition du] Pétion	10.34847/nkl.06f00638	https://nakala.fr/u/collections/10.34847/nkl.06f00638
[Édition du] Siècle	10.34847/nkl.f4f5e445	https://nakala.fr/u/collections/10.34847/nkl.f4f5e445
[Édition de l'] Union	10.34847/nkl.b5d6k760	https://nakala.fr/u/collections/10.34847/nkl.b5d6k760
[Édition du] Canel	10.34847/nkl.7dcc76uu	https://nakala.fr/u/collections/10.34847/nkl.7dcc76uu
[Édition] Chlendowski	10.34847/nkl.8babp549	https://nakala.fr/u/collections/10.34847/nkl.8babp549
[Édition de l'] Europe Littéraire	10.34847/nkl.e5579r98	https://nakala.fr/u/collections/10.34847/nkl.e5579r98
[Édition] Charpentier	10.34847/nkl.4bbczmc6	https://nakala.fr/u/collections/10.34847/nkl.4bbczmc6
Études de mœurs	10.34847/nkl.7d1b0m32	https://nakala.fr/u/collections/10.34847/nkl.7d1b0m32
Études philosophiques	10.34847/nkl.4b5e8zw2	https://nakala.fr/u/collections/10.34847/nkl.4b5e8zw2
Études analytiques	10.34847/nkl.3ff5mulk	https://nakala.fr/u/collections/10.34847/nkl.3ff5mulk
Scènes de la vie privée	10.34847/nkl.da7c094z	https://nakala.fr/u/collections/10.34847/nkl.da7c094z
Scènes de la vie de province	10.34847/nkl.2fadvhq8	https://nakala.fr/u/collections/10.34847/nkl.2fadvhq8
Scènes de la vie parisienne	10.34847/nkl.6c2ef1n9	https://nakala.fr/u/collections/10.34847/nkl.6c2ef1n9
Scènes de la vie politique	10.34847/nkl.fc77m86e	https://nakala.fr/u/collections/10.34847/nkl.fc77m86e
Scènes de la vie militaire	10.34847/nkl.cca809ji	https://nakala.fr/u/collections/10.34847/nkl.cca809ji
Scènes de la vie de campagne	10.34847/nkl.fcf53ybd	https://nakala.fr/u/collections/10.34847/nkl.fcf53ybd
Les Célibataires	10.34847/nkl.adfbt5u7	https://nakala.fr/u/collections/10.34847/nkl.adfbt5u7
Les Parisiens en province	10.34847/nkl.357cr836	https://nakala.fr/u/collections/10.34847/nkl.357cr836
Les Rivalités	10.34847/nkl.76a6ll98	https://nakala.fr/u/collections/10.34847/nkl.76a6ll98
Les Parents pauvres	10.34847/nkl.f31403ds	https://nakala.fr/u/collections/10.34847/nkl.f31403ds

n'a pas pour vocation de remplacer le site original du projet, mais de faciliter la lecture et la navigation au sein des données partagées. Dans la même veine que les expositions du *Research Data Journal for the Humanities and Social Sciences* (Brill), le but de cette présentation des données est de rendre la recherche des ressources dans l'entrepôt moins fastidieuse pour les personnes non initiées aux services et outils d'Huma-Num. Le site où les données peuvent être consultées, e-balzac.nakala.fr, a été construit en relation avec le projet *Humanités numériques et science ouverte* et a été connecté au site principal de celui-ci[46], permettant ainsi de restituer le chapitre du présent ouvrage dans son contexte. Le site principal oriente l'utilisateur vers les données des différents projets. Elles sont liées, mais ne se mélangent pas. La modélisation des données reste également propre à chaque collection et adaptée à chaque projet. Elle est renseignée à l'aide du standard Dublin Core.

38. Différents types de contenu peuvent être créés à partir de la page d'accueil (lien vers une ressource externe, liste de données, métadonnées, visualisation du contenu…). Les données contenues dans la collection et les métadonnées sont mises en forme, ce qui facilite leur lecture. En cliquant sur « Chronologie des publications », il est également possible de filtrer les données par année de création. Ainsi, un utilisateur extérieur peut facilement récupérer uniquement un échantillon chronologique du corpus qui l'intéresse. De même, il est possible de voir les différentes collections auxquelles sont liées les données. Une donnée pouvant être contenue dans plusieurs collections, il est plus aisé pour un utilisateur extérieur, par exemple de naviguer directement dans la collection « Études de mœurs » à partir de la collection « e-Balzac » depuis laquelle le site Nakala_Press est généré. Enfin, une recherche par type de licence est également possible.

39. Des pages annexes ont été ajoutées pour structurer davantage le site. Il s'agit notamment des liens vers le site du projet et vers les outils exploités dans le cadre du projet. Un moteur de recherche interne à cette collection est également disponible. Ainsi, il s'agit non seulement de rendre les données disponibles, mais aussi visibles, et surtout lisibles par des personnes extérieures au projet, voire peu familières des outils et méthodes des humanités numériques. La consultation de l'entrepôt n'est peut-être pas encore un réflexe, mais la navigation sur un site Web est une pratique généralement démocratisée.

La réutilisation des données

40. Le projet e-Balzac s'appuyant sur la production littéraire du XIX^e^ siècle, toutes les œuvres mises à disposition dans le cadre de l'édition numérique sont dans le domaine public et ne sont pas restreintes par les droits d'auteur. Cela limite de manière importante d'éventuelles questions éthiques soulevées par la collecte des données. Plus encore, alors que des éditions numériques des sources ont déjà suscité des controverses liées au travail édito-

46. Cf. https://hnso.nakala.fr/

rial, critique ou scientifique[47], le FC a été établi pour la première fois à la fin du XIXe siècle[48]. Par ailleurs, le projet e-Balzac a pris en charge l'établissement du texte à partir de l'exemplaire personnel de Balzac. La version de référence de la Pléiade a été consultée dans certains cas épineux (mais l'interprétation des corrections manuscrites a parfois été divergente).

41. Dans le cadre du projet, les données sont employées de nombreuses manières. Grâce au partenariat avec le projet ARTFL de l'Université de Chicago, il est possible d'interroger le corpus à l'aide d'un moteur de recherche lexicale. L'adaptation du logiciel MEDITE et la création des fichiers XML regroupant les variantes des deux versions différentes d'un texte ont permis de proposer une comparaison informatique de différents états du texte qui facilite une étude génétique de l'œuvre de Balzac. Enfin, grâce à la réutilisation des données mises à disposition par d'autres projets ou des archives numériques, il a été possible de débuter le dernier axe du projet, orienté vers une édition hypertextuelle et visant à recomposer une bibliothèque virtuelle qui comprend l'ensemble des textes littéraires et non littéraires dont Balzac a pu s'inspirer dans la création de *La Comédie humaine*. Mais les possibilités sont loin d'être épuisées. Alors que nous nous intéressons aux inspirations balzaciennes, il paraît stimulant d'investiguer l'influence de l'auteur sur ses successeurs, et comparer *La Comédie humaine* à un corpus postérieur. Le corpus comparatif, composé des fichiers XML qui regroupent deux versions d'un texte et précisent le type des variantes[49], pourrait être exploré massivement, par exemple en extrayant les variantes seules pour déterminer ce qui caractérise les modifications les plus

47. Cf. par exemple (Rageot 2014).

48. Il s'agit de l'édition Michel Lévy (puis Calmann-Lévy) parue de 1869 à 1876. Elle a été ensuite complétée par Conar en 1912-1940, puis perfectionnée enfin par la deuxième édition de la Pléiade (1976-1981), devenue depuis la version de référence. Cf. (Del Lungo 2017).

49. Pour ce faire, nous employons l'élément <choice> dont @ana précise le type de modification (remplacement, suppression, insertion, déplacement). <choice> peut avoir deux fils, <orig>, qui code le contenu du texte publiée antérieurement, et <reg>, qui contient la variante du texte postérieur. Voici, par exemple, la manière dont sont codées les modifications du dédicace de *Modeste Mignon* entre l'édition F et celle du FC (figure 1) :

```
<div type="dedication">
    <pb n="113" xml:id="p113"/>
    <salute>À UNE
            <choice ana="remplacement">
                <orig>ÉTRANGÈRE</orig>
                <reg>POLONAISE</reg>
            </choice>
    </salute>
    <p rend="noindent">Fille d'une terre esclave, ange par l'amour, [...]
            <choice ana="remplacement">
                <orig>les expressions visibles</orig>
                <reg>l'expression quand elle anime ta [...]</reg>
            </choice>
    sont
            <choice ana="remplacement">
                <orig>comme ces</orig>
                <reg>pour les savants les</reg>
            </choice>
    caractères d'un langage perdu
            <choice ana="remplacement">
                <orig> qui préoccupent les savants.</orig>
                <reg>. </reg>
            </choice>
    </p>
            <signed><hi rend="sc">De Balzac.</hi></signed>
</div>
```

fréquentes. Le cas de Balzac est réplicable pour d'autres auteurs, aussi bien Zola que Hugo par exemple.

42. Par ailleurs, différents niveaux de réplicabilité peuvent être mis en avant à travers ce *data paper*, que ce soit au niveau du jeu de données, que des pratiques et méthodes décrites. À la multiplication des utilisations, nous ajoutons la multiplication des formats : la transformation du XML-TEI dans les formats HTML et EPUB permet la diffusion des textes au public plus large et l'utilisation de l'édition au-delà du milieu universitaire (dans le cadre scolaire, voire plus une lecture de plaisir). L'emploi du standard TEI garantit l'interchangeabilité et l'ouverture des données produites dans le cadre du projet et l'emploi de la syntaxe Teinte, plus restreinte, mais adapté au projet et développé avec lui, permet de proposer des textes dotés d'une structuration sémantique. L'établissement rigoureux du format XML a été priorisé, puisque celui-ci est considéré comme un format pivot, à partir duquel nous procédons aux transformations automatiques vers d'autres formats éditoriaux. L'automatisation de la chaîne du traitement garantit par ailleurs la conformité des versions des fichiers et facilite la correction plus rigoureuse d'éventuelles coquilles signalées par les utilisateurs.

43. Comme le constatent Marcello Vitali-Rosati et Michael Sinatra (2014, 60) : « [I]l ne s'agit pas seulement de choisir, de légitimer, de mettre en forme et de diffuser un contenu, mais il s'agit aussi de réfléchir à l'ensemble des techniques que l'on va utiliser ou créer pour le faire, ainsi qu'aux contextes de circulation produits par l'espace numérique. Si les humanités numériques s'occupent de produire des outils et de réfléchir à leur impact sur la production et la circulation du savoir, alors l'éditorialisation devient l'objet central de leur travail. »

44. Le concept de la science ouverte invite à donner accès non seulement aux résultats, mais aussi aux chaînes de traitement qui ont permis leur production. Nous sommes persuadés que le dépôt des données collectées et produites au sein de chaque projet, accompagné d'une documentation facilitant leur réutilisation, permet de remettre l'éditorialisation au centre de la réflexion sur les objets et outils numériques.