# « L'édition de pièces de théâtre la plus parfaite jamais publiée » : numérisation de Lacy's Acting Edition.

Le théâtre victorien n’a pas bénéficié d’une documentation ou d’une numérisation aussi systématique que le roman victorien, ce qui reflète peut-être la perception scientifique de leur importance artistique comparée. Pourtant, il est évident que le théâtre populaire victorien et le roman victorien étaient complémentaires à bien des égards. Selon les mots d'Emily Allen, « le théâtre victorien était l'allié, l'inspiration et le concurrent du roman » [Allen] En tant qu'expression continue de la culture populaire, le théâtre du XIXe siècle a des racines profondes et de nombreuses branches. Sa lignée nous emmène du haut gothique du mélodrame romantique aux mèmes du cinéma et de la télévision moderne ; ses voix combinent spectacle sensationnel, commentaire social et réalisme domestique.

Les œuvres dramatiques du XIXe siècle ont été fréquemment réimprimées ou rééditées, tant à Londres qu'à New York, jusque tard dans le XXe siècle, ce qui témoigne d'une demande continue, malgré le manque présumé de valeur littéraire des œuvres de Douglas Jerrold, Edward Planche, Shirley Brooks. , J. Palgrave Simpson, John Madison Morton et d'autres piliers de la Dramatic Authors' Society. Mais contrairement au roman populaire, le théâtre populaire n’a produit que très peu d’auteurs ou d’œuvres que l’on qualifierait aujourd’hui de « canoniques », même si de nombreux écrivains témoignent de leur influence.

Le déclin du théâtre victorien au cours de la longue période entre les jours glorieux de Covent Garden au XVIIIe siècle et l'émergence de Shaw et d'Ibsen a souvent été attribué [par exemple Barrett 1999] au fait que les pièces n'étaient pas publiées pour être lues, mais au profit des acteurs et des managers qui envisagaient de les produire sur scène, et principalement comme moyen de conférer à l'œuvre un certain degré de protection du droit d'auteur. Le « acting edition » était une production à petit budget, purement utilitaire, conçue pour faciliter l’expérience théâtrale complète, mais en aucun cas pour la remplacer. Comme le livret ou la partition d'un opéra, il ne s'adressait pas au critique dans son cabinet, encore moins au grand public.

Par conséquent, pour qui souhaite appréhender le phénomène dans son ensemble et procéder à une sorte de lecture distanciée de ses textes, les ressources bibliographiques sont difficiles à trouver, et encore plus les recueils fiables de textes de théâtre. Il existe un petit nombre de listes de références encyclopédiques [Nicoll, Mullen], mais la documentation sur le théâtre victorien a tendance à se concentrer sur la carrière d'acteurs individuels, d'écrivains ou d'institutions. De même, lorsque des collections de sources primaires sont disponibles, c'est le résultat de collections individuelles ou de legs plutôt que d'une tentative de couverture systématique. Le canon littéraire ne fournit aucune indication sur la meilleure manière d'échantillonner la production théâtrale entre (disons) la fin des guerres napoléoniennes et la fin du 19e siècle , bien que ces années soient généreusement parsemées d'œuvres d'une vigueur, d'une originalité et d'une influence énormes – ainsi que beaucoup de scories.

Comme approximation d'un tel échantillon, nous proposons d'utiliser les quelque 1 500 titres composant *Lacy’s Acting Edition of Plays* publié par Thomas Hailes Lacy (1809-1873) entre 1848 et 1873. Lacy n'était en aucun cas la première personne a eu l'idée de réimprimer et de publier des éditions uniformes de textes dramatiques. Comme ses prédécesseurs John Dick, John Cumberland et John Duncombe, le plan commercial était d'acheter (si nécessaire) les droits d'auteur des œuvres à succès et de les réimprimer pour un marché de masse, à la fois sous forme de textes individuels et de volumes rassemblés. Le public envisagé comprenait à la fois des passionnés de théâtre et des professionnels en quête de matière à succès assuré. Par conséquent, ses « Acting Editions » fournissaient tout ce qui était nécessaire pour produire une pièce : y compris les détails des costumes, les décors, la mise en place, les gestes et tout le « business » d'accompagnement, ainsi que les listes de distribution et le texte lui-même. Il a été avancé de manière plausible [Weig 2017] qu'au cours de la longue histoire de publication de l'Acting Edition, son public cible a changé, visant davantage l'amateur de la classe moyenne qui se serait senti mal à l'aise dans un vrai théâtre, et moins l'entrepreneur théâtral professionnel. Quoi qu’il en soit, de « nouveaux » titres paraissent régulièrement jusqu’en 1873, date à laquelle Lacy vend l’ensemble de la collection à Samuel French, un éditeur américain avec qui il avait échangé des pièces contre publication au cours des deux décennies précédentes.

Lacy a accumulé assidûment les droits des auteurs pendant de nombreuses années et a également entretenu une petite écurie d'écrivaillons qui produisaient chacun des dizaines de titres. Il semble peu probable que les éléments constitutifs de chaque volume de Lacy's Acting Edition – 15 titres pour chacun des 100 volumes étudiés – aient été sélectionnés de manière purement fortuite. Le choix des titres pour « Acting Edition » reflète probablement la perspicacité d'un éditeur avisé quant aux titres les plus susceptibles de se vendre. L’examen des caractéristiques saillantes de chacun des titres publiés dans la série au cours de sa longue histoire peut donc servir de mesure approximative de l’évolution du goût – ou peut-être des aléas de la mode – au cours de la période concernée. Des caractéristiques telles que le genre, la durée, la source, l’ancienneté et le coût, qui peuvent être déterminées relativement simplement pour chaque titre, peuvent également être utiles pour déterminer dans quelle mesure un sous-ensemble donné de titres est représentatif de l’ensemble.

Des exemplaires de titres individuels, de volumes individuels ou d'ensembles de volumes de Lacy's Acting Edition se trouvent dans de nombreuses bibliothèques du monde entier, bien que l'édition elle-même n'ait, à ma connaissance, fait l'objet d'aucune étude bibliographique approfondie. De nombreuses copies survivantes, tant de titres individuels que de volumes complets ou d'ensembles de volumes, ont été numérisées en facsimilé généralement par Google agissant au nom de la bibliothèque détentrice. Un exemple particulièrement notable est la copie du LAE provenant de l'Université du Michigan qui comporte un ensemble presque complet des 94 premiers volumes, qui ont tous été numérisés (bien qu'avec plus ou moins de succès) et sont désormais « gratuitement » accessibles à ceux qui peuvent accéder à la collection Hathi Trust. Mais une recherche sur Internet révélera rapidement l'existence de nombreux autres titres, y compris certains des six volumes restants, numérisés de la même manière par Google, à partir de copies provenant d'un certain nombre d'autres bibliothèques.

## Le **Victorian Plays Project**

La production d'une version numérisée de l'Acting Edition a été proposée pour la première fois par le professeur Richard Pearson, dont le *Victorian Plays Project* (VPP) a été financé par l'Arts and Humanities Research Council du Royaume-Uni entre 2005 et 2007. Initialement hébergé à l'Université de Worcester au Royaume-Uni, ce projet a été transféré à l'Université de Galway en Ireland en 2014. Malheureusement, le site Web du VPP n'a pas été développé depuis 2015 et le décès prématuré du professeur Pearson fin 2018 ( [https://bavs.ac.uk/ uncategorized/obituary-richard-pearso n/](https://bavs.ac.uk/uncategorized/obituary-richard-pearson/) ) jette un sérieux doute sur son développement futur. Comme c’est trop souvent le cas, la préservation d’une archive numérique s’avère dépendre autant de l’accompagnement personnel d’un individuel que de contraintes technologiques.

L'un des principaux résultats du VPP était un catalogue en ligne de l'intégralité de Lacy's Acting Edition, basé sur une copie conservée dans ce qui était alors Birmingham Central Library. Le projet actuel Digital Lacy est donc une tentative de relance du VPP, en élargissant et en corrigeant ce catalogue, et en créant un échantillon équilibré de textes transcrits.

Au printemps 2022, j'ai été financé par le projet CLS Infra pour produire une courte étude de cadrage (voir https://clsinfra.io/opportunities/tnafellowships/tna-archive/tna-fellow-lou-burnard/), visant établir l'état actuel des livrables du VPP et examiner ce qui pourrait être nécessaire pour le relancer. J'ai introduit un nouveau système de numérotation dans le catalogue, apporté de nombreuses corrections mineures et l'ai enrichi de liens vers des versions numériques existantes des textes. Certains des défis impliqués sont décrits dans un article de blog intitulé *Hunting for Lacy traces in the digital world,* daté du 16 août 2022 ; voir <https://foxglove.hypotheses.org/678>. J'ai également commencé à travailler sur la production de transcriptions conformes au TEI de certains titres.

Les titres à inclure dans le VPP original ont été sélectionnés apparemment au hasard, bien qu'il y ait nettement moins de titres choisis des volumes ultérieurs. Les titres individuels ont été numérisés et enregistrés sous forme de fichiers TIFF, soit un total de près de 16 000 pages. Ceux-ci ont ensuite été traités par Abbyy Fine Reader pour produire une transcription en texte brut, qui a ensuite été éditée pour imiter l'apparence typographique des publications originales et enregistrée au format PDF. Je crois comprendre que ce PDF a de nouveau été relu et corrigé directement à l'aide du logiciel Adobe, ce qui implique que les versions PDF disponibles en ligne (jusqu'à récemment) représentent probablement les meilleurs textes disponibles en termes d'exactitude, même si elles ne sont pas encore entièrement exemptes d'erreurs. D'un autre côté, les corrections ont introduit dans certains cas des variations dans le formatage (par exemple, des changements de taille de police ou de casse non présents dans l'original).

## Autres sources en ligne

Un autre sous-ensemble intéressant de titres Lacy se trouve en ligne dans la Hall Collection of Prompt Books détenue par l'Université de Warwick. [https://wdc.contentdm.oclc.org/digital/collection/hall] Un livret de souffleur est une bête assez différente d'un volume collecté : il contient généralement des annotations de son premier propriétaire, indiquant toutes sortes de modifications apportées au texte, destiné à faciliter la performance d'un acteur individuel. La collection Hall comprend environ 600 titres individuels, utilisés comme livrets de souffleur par Clara St. Casse, une actrice qui a voyagé avec diverses troupes itinérantes britanniques au XIXe siècle. Ceux-ci sont disponibles gratuitement à la fois sous forme d'images de page TIFF et sous forme PDF légère à partir d'Internet Archive. 116 titres sont issus du LAE.

Une numérisation sous forme d’image de page n’est bien sûr pas la même chose qu’une transcription exploitable par machine. Certains titres de LAE ont cependant été transcrits et peuvent être consultés en ligne. Le projet Gutenberg compte au total 17 titres dans différentes nuances de HTML ; le chercheur assidu trouvera également quelques cas où un site dédié à un écrivain particulier a inclus des textes de l'édition Lacy : par exemple des titres de Henry Milner, ou de W. S. Gilbert. Enfin, dans une sorte de maison de transition, environ 40 titres de LAE sont actuellement inclus dans la collection « Literature online » de Proquest. Ces titres ont été initialement préparés au format TEI pour être publiés par Chadwyck Healey dans une collection de « Nineteenth Century English Drama » sur CD, mais semblent désormais être disponibles uniquement sous forme RTF.

Les titres LAE sont donc potentiellement disponibles à partir de diverses sources et dans plusieurs formats :

(1) Images de pages des volumes provenant du bibliothèque de l’université de Michigan (1381 sur 1422 titres trouvés)

(2a) Images de pages des volumes provenant du bibliothèque de Birmingham (263 titres trouvées)

(2b) Images de pages de la version 2a corrigée par le VPP (343 titres trouvées)

(3) Images de pages provenant de la collection Hall (76 titres)

(4) Images de pages provenant de diverses autres sources (887 titres trouvés)

(5) Transcriptions dérivées des publications de Chadwyck Healey : (42 trouvées)

(6) Transcriptions en HTML distribuées par le Projet Gutenberg et (principalement) vérifiées par rapport aux images de pages d'Internet Archive (17 trouvées)

En pratique, il y a peu de chevauchements. Quelques titres LAE (82) n'ont jamais été numérisés. 957 – bien plus de la moitié – sont disponibles sous une seule forme ; J'ai trouvé 410 titres pour lesquels deux formes différentes sont disponibles, et seulement 63 pour lesquels il y en a 3 ou plus.

## Vers **des** transcriptions TEI

J'ai expérimenté un certain nombre de flux de travail dans la poursuite de mon objectif chimérique de produire une version du LAE conforme à la TEI. Aucune d’entre elles n’a été une réussite à 100 %, dans la mesure où toutes ont nécessité un certain degré de travail manuel, variable en termes de temps nécessaire.

Des approches sophistiquées d'apprentissage automatique pour créer des versions TEI-XML plus ou moins automatiquement à partir de sources PDF ont été utilisées, par exemple dans les projets Europarl ou RDA. Mais ces projets traitent pour la plupart de très grandes quantités – des milliers – de documents formatés de manière assez cohérente. Presque tous les PDF que j'ai consultés dans le VPP ont un format légèrement différent. Il y a peu de cohérence dans la manière dont les préfixes des locuteurs ou les didascalies sont formatés. Les variations dans la taille et le format des caractères sont imprévisibles. Des collègues possédant une expertise dans ce domaine ont confirmé mes soupçons selon lesquels il serait peu pratique d'adopter cette approche tant qu'un nombre raisonnablement important (par centaines) de textes de jeu correctement étiquetés ne seraient pas disponibles pour servir d'ensemble d'entraînement.

En l'absence d'un ensemble d'entraînement, j'ai expérimenté certains outils disponibles gratuitement prétendant convertir un PDF en HTML ou XML. De tels outils tentent de représenter la manière dont les caractères d'un fichier PDF reconnu doivent être affichés sur la page, en termes de lignes, de blocs, de pages, etc. Il existe deux normes XML ouvertes largement utilisées pour coder ces informations : PAGE et ALTO ; un projet intéressant appelé SegmOnto développe des outils génériques pour construire un pipeline de conversion basé sur ALTO, bien qu'il n'existe pas encore de produit stable. J'ai expérimenté un outil beaucoup plus ancien, appelé *pdftohtml,* dérivé de la vénérable bibliothèque *xpdf* , initialement développée pour produire une sortie imprimée à partir d'un fichier PDF indépendamment du logiciel Adobe. Cela peut être utilisé pour générer des fichiers XML conformes à un schéma de type ALTO qui peuvent ensuite être traités par une série de scripts XSLT (par exemple, ceux fournis par Dario Kampkaskar dans sa bibliothèque pdftotei) pour générer quelque chose de plus traitable et de type TEI.

J'ai également réalisé quelques expériences avec Abby FineReader, probablement le leader actuel du marché des logiciels OCR commerciaux. La service de recherche française Huma-Num donne accès à Abbyy et aussi à Tesseract, son principal concurrent open source, ce qui m'a permis de faire une comparaison entre les deux.

Le diable se cache, comme toujours, dans les détails. Parmi les systèmes que j'ai expérimentés, tous ont échoué à un ou plusieurs des obstacles suivants :

* l'espace blanc entre les mots individuels doit être préservé
* l'ordre des sous-chaînes dans le document doit être conservé
* les distinctions de police (italique ou romaine) doivent être préservées
* les unités structurelles adjacentes (par exemple une série d'éléments <hi> avec les mêmes valeurs pour @rend) doivent être fusionnées correctement (presque impossible lorsque, par exemple, une mise en scène en italique contient des mots non en italique)

Aucune des approches que j'ai envisagées n'a été en mesure de fournir une solution complète pour un PDF VPP typique. La solution « la moins pire » que j'ai trouvée était un pipeline dans lequel Abbyy était utilisée pour générer une représentation docx afin de conserver une grande partie du formatage d'origine, qui était ensuite affiné par une série de feuilles de style XSLT, en commençant par la feuille “docxtotei” fournie par la TEI, et se terminant par une vérification manuelle de chaque transcription par rapport aux images de la page.

La reconnaissance de la structure globale (division en actes et scènes) a été partiellement réussie pour la plupart des textes que j'ai examinés, car les feuilles de style TEI distinguaient pour la plupart correctement les lignes de titre du corps du texte. De même, il s'est avéré possible d'identifier automatiquement des discours individuels par correspondance de modèles (par exemple en recherchant des blocs de texte commençant par un mot en italique suivi d'un point) ; pour identifier les indications scéniques (blocs individuels commençant par un caractère [ ou sous-chaînes données entre parenthèses), etc. Cependant, étant donné que ces modèles typographiques variaient considérablement d’un texte à l’autre, et même à l’intérieur d’un texte, un certain degré d’intervention manuelle reste nécessaire.

## Le système de balisage Lacy

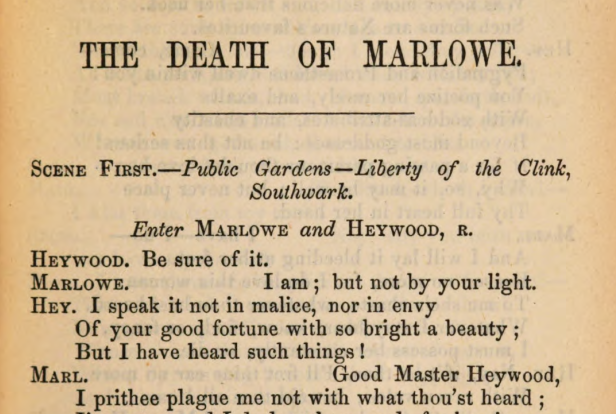
Comme pour tout autre projet TEI, le schéma de balisage ciblé des documents Lacy TEI est défini par un ODD, qui spécifie tous les éléments TEI utilisés, leurs attributs, types de données, etc. La première ébauche de ce ODD a été produite automatiquement à partir d'un ensemble préliminaire de six pièces balisées à la main, puis adaptées au fur et à mesure que d'autres pièces étaient testées. Cela reste un travail en cours.

Le Lacy ODD exige que toutes les pièces suivent la même structure de base et n'enregistre que les caractéristiques textuelles susceptibles d'être d'une utilité générale dans l'analyse. Il ne tente pas de prendre en charge un balisage philologique détaillé, par exemple des variations textuelles, des effets de formatage, une analyse métrique ou linguistique, etc. Il propose cependant plusieurs balises bien adaptés à l’encodage des métadonnées détaillées sur les détails de l'interprétation, le genre, la taille, etc. pour chaque titre. Tous les détails de la spécification sont fournis dans le manuel Lacy, qui est maintenu en ligne avec les documents TEI (voir http://lb42.github.io/Lacy/).

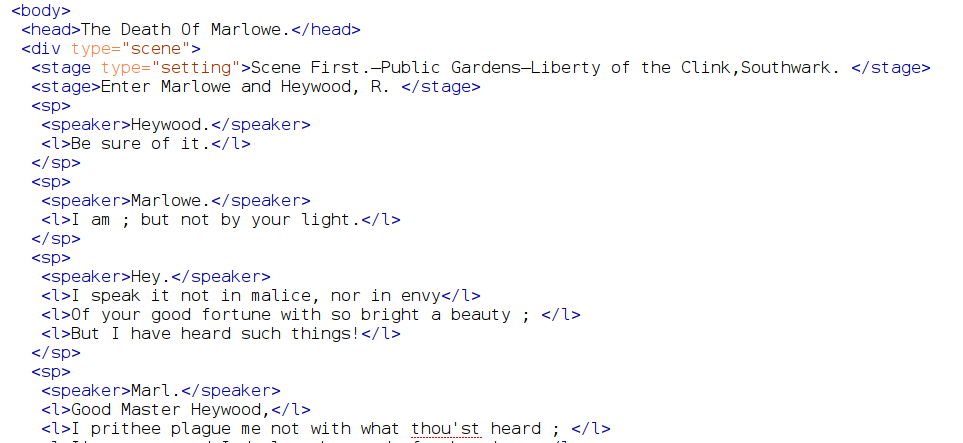
Chaque titre (c'est-à-dire chaque œuvre réimprimée dans le LAE) est considéré comme un seul élément <TEI>, contenant un en-tête TEI et un élément <text>. Chaque élément de texte comporte un début, un corps et un verso facultatif.

Le corps d'un texte consiste en une séquence de mises en scène, de discours et de groupes de discours (par exemple des numéros musicaux), représentés respectivement par les éléments TEI <stage>, <sp> ou <spGrp>. Un élément <sp> contiendra généralement une indication initiale du locuteur (<speaker>) suivie d'une série de paragraphes (<p>) ou de lignes de vers (<l>). Ces éléments contiennent tous uniquement du texte courant, bien qu'un petit nombre d'autres éléments de niveau texte soient également autorisés, par exemple des sauts de page, des notes, des mots ou des phrases imprimés en italique et des corrections de transcription. Les actes et les scènes, s'ils sont indiqués dans le texte source, sont marqués à l'aide de l'élément TEI <div>.

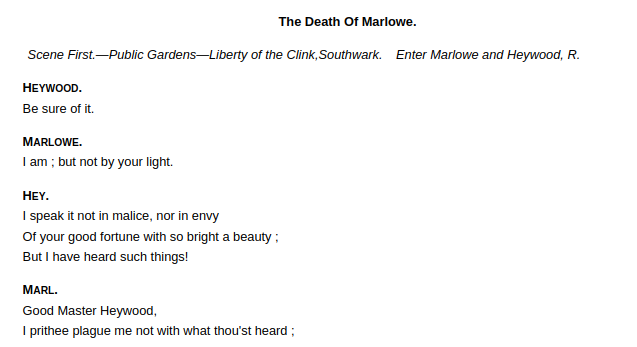
Voici par exemple le début de *The Death of Marlowe* (L1325) de R .H. Horne :



Et voici la version codée TEI du même passage :



Notez que la variation de police, en particulier l'utilisation de petites majuscules et d'italique, a été entièrement abandonnée, tout comme l'indentation supplémentaire utilisée pour indiquer la fragmentation de la ligne de vers. Le but de cet encodage n’est pas de représenter toutes les caractéristiques visuellement saillantes, mais simplement celles qui sont susceptibles d’être fonctionnellement significatives. Un programme tel que CETEICEAN peut profiter de ce balisage pour produire quelque chose de visuellement satisfaisant :



Les pages préliminaires d’une édition Lacy peuvent contenir plusieurs composants distincts :

* une page de titre
* notes de performances
* une liste de distribution, répertoriant d’habitude les rôles et les acteurs
* informations sur les costumes
* une liste de scènes ou de spectacles
* autre passages préliminaires ou dédicatoire

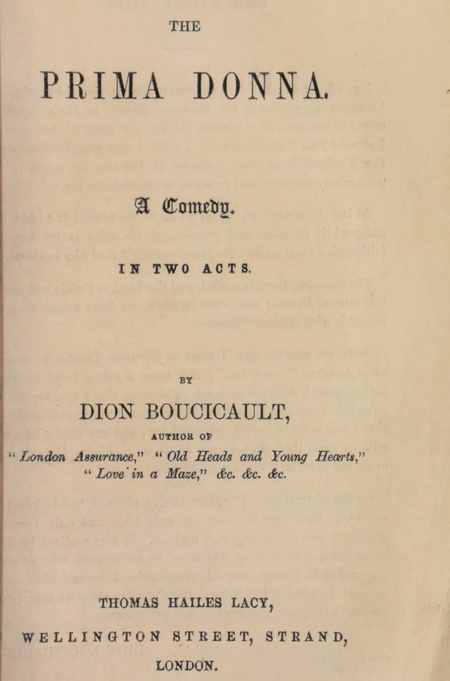
Pour chacun d’eux, nous utilisons les éléments TEI standards et leurs sous-composants, comme indiqué ci-dessous. Encore une fois, nous n'essayons pas de reproduire entièrement l'apparence de l'original, même si les pages de titre et les listes de programmes en particulier manifestent souvent une exubérance typographique considérable.

### Titre de page

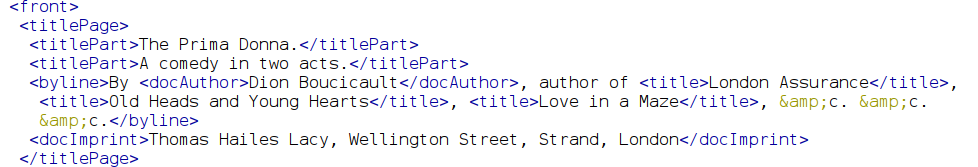
L'élément <titlePage> contient une transcription du texte imprimé sur la page de titre d'une pièce de théâtre. Celui-ci contiendra généralement

1. au moins un et généralement plusieurs fragments de titre, chacun dans un <titlePart>
2. une signature indiquant l'auteur et ses autres œuvres
3. une empreinte, comprenant généralement l'adresse de Lacy

Par exemple:



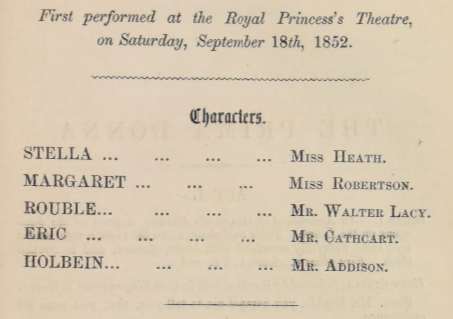
Cela serait encodé comme suit :



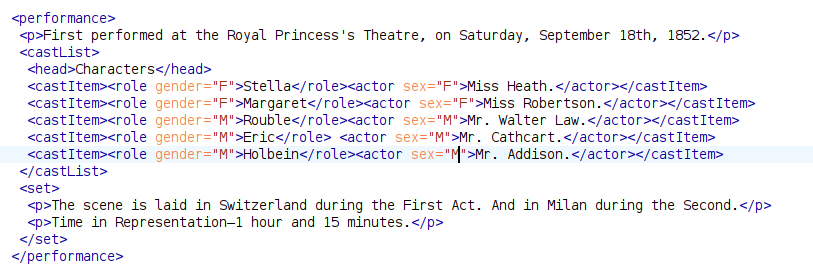
Le formatage de l'original est en grande partie abandonné : le balisage ne distingue que le titre principal de son sous-titre, et indique également les limites des autres titres mentionnés. La casse est normalisée et les sauts de ligne ignorés, tout comme lors de la transcription du texte de la pièce.

### Avis de performance et liste des acteurs

La plupart des pièces de théâtre contiennent une déclaration formelle indiquant la date et le lieu de leur première représentation, car cela constitue une affirmation du droit d'auteur. Il est souvent associé à un *dramatis personae* , indiquant les personnages de la pièce et généralement le nom d'un acteur.

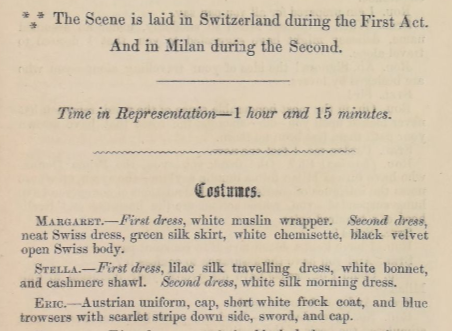


La liste des rôles est potentiellement importante pour analyser, par exemple, les différentes contributions apportées par les rôles masculins et féminins, et pour identifier la fréquence à laquelle une actrice assume un rôle masculin. Nous incluons ces informations dans la version balisée :

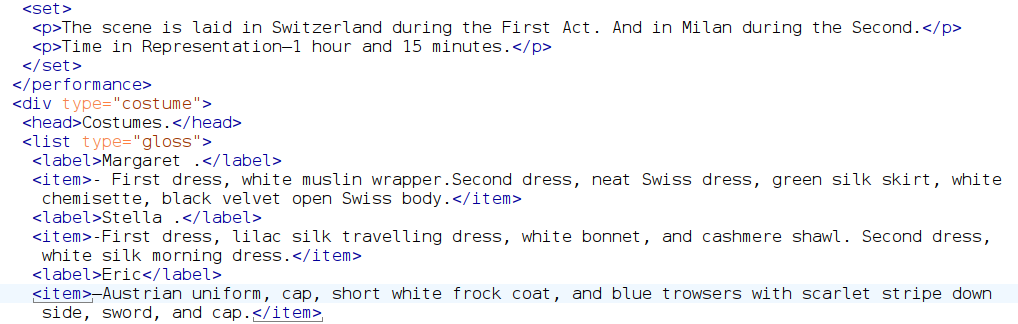


### Décor et costumes

Des informations sur le décor, la durée de la présentation et des notes détaillées sur le costume se trouvent également couramment dans les pages préliminaires.

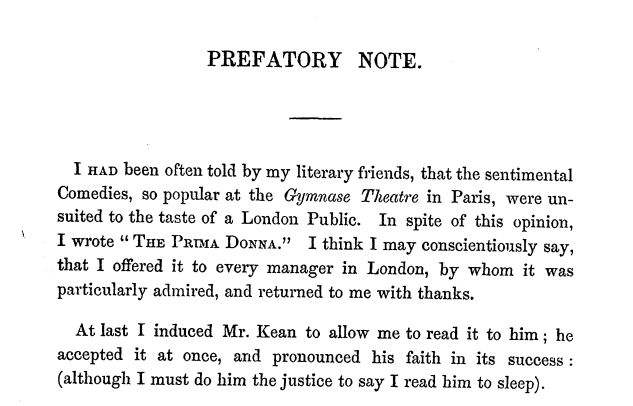


Ceux-ci sont balisés comme suit :

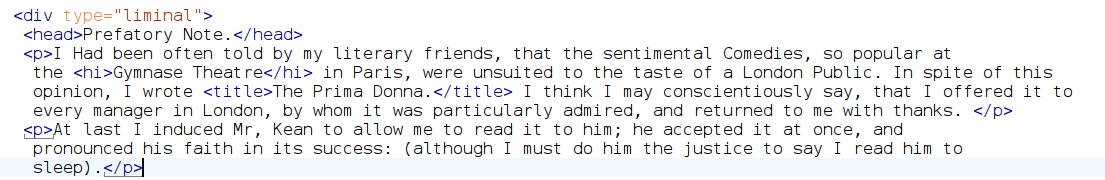


### Autres parties préliminaires

Les pièces incluent parfois d'autres types de sujets préliminaires, comme dans ce cas



Ce type de mention peut être balisé à l'aide des éléments TEI habituels <div>, <head>, <p>, <list>, <signed>, <epigraph>, <quote>, etc. Dans ce cas, le premier paragraphe contient une phrase en surbrillance et un titre, pour lesquels nous utilisons respectivement <hi> et <title>.



D'autres exemples sont fournis dans le manuel Lacy référencé ci-dessus.

## Prochaines étapes

Au moment d'écrire ces lignes, plus de 50 titres ont été convertis au format TEI, et le travail se poursuit, quoique lentement. En parallèle, je travaille sur des feuilles de style et d'autres outils pour analyser le corpus émergent. Mais la première priorité sera de lancer un nouveau site Internet « Digital Lacy » présentant le catalogue révisé, enrichi de liens vers d'autres ressources utiles.

## **R**éférences

1. “Thomas Hailes Lacy” [DNB entry] <https://en.m.wikisource.org/wiki/Dictionary_of_National_Biography,_1885-1900/Lacy,_Thomas_Hailes>) ; see also <https://doi.org/10.1093/ref:odnb/15862> (paywall)
2. *Truly Yours: One Hundred and Fifty Years of Play Publishing and Service to the Theatre.* Samuel French, 1980.
3. Cadwallader , Jen & Larry Mazzeno. "Teaching the 'Forgotten' Genre: Victorian Drama." Teaching Victorian Literature in the 21st Century: A Guide to Pedagogy. Palgrave Macmillian, 2017. 167-79.
4. Mattacks, Kate. "‘A Woyage O’Diskivery’: Thomas J. Williams' The Peep-Show Man (1868) and the Victorian Performative Text." *Nineteenth Century Theatre and Film* 36.2 (2009): 53-62.
5. Mattacks, Kate. "Sensationalism on Trial: Courtroom Drama and the Image of Respectability in His Darling Sin." *New Perspectives on Mary Elizabeth Braddon*. Brill, 2012. 211-230.
6. Mattacks, Kate. "Reading Theatre Writing: TH Lacy and the Sensation Drama." *Media and Print Culture Consumption in Nineteenth-Century Britain*. Palgrave Macmillan, London, 2016. 183-197.
7. Mattacks, Kate. "Regulatory bodies: Dramatic creativity, control and the commodity of Lady Audley's Secret." *19: Interdisciplinary Studies in the Long Nineteenth Century* 8 (2009).
8. Mullin, Donald. *Victorian Plays: a record of significant productions on the London stage, 1837-1901.* Greenwood Pr. 1987.
9. Nicoll, Allardyce, *A History of English Drama 1660-1900.* Cambridge University Press.1959