Centre d'études supérieures de la Renaissance
UFR de l'université de Tours - UMR 7323 du Centre National de la Recherche Scientifique
www.cesr.cnrs.fr
www.cesr.univ-tours.fr
59, rue Néricault-Destouches - BP 12050
37020 Tours Cedex 1 - France
Tél. : +33 (0)2 47 36 77 61 - Fax : +33 (0)2 47 36 77 62

Documentation encodage xml-tei :

œuvres complètes de Voltaire, Tome XII, Poésies, Tome I.

Lafon Noémie

Master 1 : Humanités numériques

22/12/2024

Choix pour l'OCRisation : Tout d'abord, l'outil ABBYY a été utilisé pour effectuer l'OCRisation du texte initialement au format .pdf. Une fois la numérisation effectuée, les différentes erreurs potentielles dues à l'OCR ont été manuellement corrigées/révisées. Le document a ensuite été téléchargé au format .docx pour être intégré dans Oxygen. L'utilisation de Regex aurait pu me permettre de supprimer les tirets qui coupaient les mots. Toutefois, en raison du faible nombre d'erreurs récurrentes, la correction a été finalisée manuellement et directement dans Oxygen avant de passer à l'encodage du texte.

Documentation : codification du texte et encodage XML.

Le texte est divisé en plusieurs sections avec des balises et attributs spécifiques respectant les TEI Guidelines, dans le but d'avoir une visualisation claire, organisée et similaire à la source d'origine.

Choix de structure <teiHeader> : La balise **<**teiHeader**>** contient les métadonnées du projet numérisé et de la source originale. Elle inclut : la balise <fileDesc> qui comprend <titleStmt> permettant d'avoir le <titre> *Edition numériques : oeuvres complètes de Voltaire, tome XII, poésies tome I.* et <author> *Noémie Lafon* du fichier électronique. Puis la balise <publicationStmt> pour spécifier l'autorité <authority> *Cours M1 données textuelles* et la <date>22/12/2024 de la numérisation. Enfin la balise <sourceDesc> englobe les balises décrivant la source d'origine de la numérisation. Dans ce cas elle contient la balise <listBibl> qui est utilisée pour intégrer les deux références bibliographiques de la source d'origine : <bibl n="1"> et <bibl =n"2"> avec à l'intérieur de chacune le <titre> de la source, le <publisher>, <pubPlace> et <date>.

Choix de structure du <texte> : Les trois premières pages du document sont dans une balise <front> puisqu'elle permet de contenir ce qui est avant le texte principale, comme les pages de titre et les informations bibliographiques. Chaque page est encapsulée dans une balise <titlePage> et <titlePart> ce qui permet de structurer les éléments de la page avec les différentes balises comme <note>, <publisher>,<pubPlace>,<date>.

Dans la balise <text>, choix de structure du <body> : Le document comprend un avertissement, le premier poème et ses notes, le second poème et ses notes, un avertissement pour le troisième poème, le troisième poème et ses notes. Tout d'abord, chaque poème/note/avertissement débute avec son titre dans une balise <head> avec un attribut type et rend pour ajouter des détails supplémentaires sur la nature et l'apparence du titre. Tous les éléments liés au titre "poèmes" de la page 7 sont placés dans une balise c'est-à-dire du premier poème aux notes du dernier poème. À l'intérieur de cette balise , chaque nouveau poème est encapsulé dans une autre balise . Et pour chaque poème, ainsi que ses notes et/ou avertissements, sont eux-mêmes individuellement englobés dans des balises . Pour plus de clarté chaque balise a deux attributs type et n (permettant de catégoriser et de se repérer dans la hiérarchie du document).

par exemple : *poème La Bastille* *notes du poème La Bastille* . Pour plus de clarté chaque premier div à un attribut de position n.
Cela permet de garantir une structure organisée et hiérarchisée des sections.

structuration des titres : les titres sont dans la plupart des cas dans deux balises <head>, une <head type= “title”> et <head type = “subtitle”> en raison de la différence de taille d’écriture qui suggère un titre et un sous-titre, malgré la ponctuation. Si la taille de la police aurait été identique par exemple pour NOTES ET VARIANTES DE LA BASTILLE alors le titre aurait été dans une unique balise <head>.

Structuration des poèmes : Pour les poèmes, l’utilisation des balises <lg> permet de respecter leur structure et de regrouper les vers. Elle est constituée de balises <l> qui définissent chaques vers du poème. Pour chaque saut de ligne j’utilise la balise <lb/> et alinéa <l rend=”indent”> pour donner des informations sur l’aspect du document. J’ai choisi de ne pas définir les strophes des poèmes dans divers <lg></lg> car elles ne correspondaient pas au “découpage” visuel de la source originale, c’est à dire qu’après avoir fait des recherche sur ces poèmes et Voltaire les sauts de lignes ou les alinéas ne correspondent pas au début d’une nouvelle strophe. Pour ne pas faire d'interprétations erronées. J’ai simplement indiqué saut de ligne et alinéa qui étaient présents dans la source originale dans l’encodage du poème.

Structuration des notes : Dans les différents poème il y a plusieurs balises : “<ref target="#note1b"><hi rend="sup">1</hi></ref>” la seconde balise avec son attribut permet de mettre dans ce cas 1 en exposant et la première balise avec l’attribut <ref target="#note1b"> permet de créer un lien avec un élément qui a ici pour identifiant note1b. Ainsi dans chaque page de note de chacun des trois poèmes l' utilisation d’identifiant <note xml:id="note1b"> permet de lier par renvoie les informations correspondantes entre elles (ici de l’exposant vers sa note). Les notes a se réfèrent au poème 1, b au poème 2 et c au poème 3.

Tableau récapitulatif : Les balises en gras signifient qu’elles englobent d’autres balises avant d’être fermées.

TEI éléments	Description	Attributs	Guidelines
<teiHeader>	En-tête TEI, fournit des informations descriptives et déclaratives qui constituent une page de titre électronique au début de tout texte conforme à la TEI. métadonnées	aucun	TEI element teiHeader (TEI header)
<fileDesc>	Description bibliographique du fichier électronique.	aucun	TEI element fileDesc (file description)
<titleStmt>	Mention de titre, regroupe les informations sur le titre d’une œuvre et les personnes ou institutions responsables de son contenu intellectuel.	aucun	TEI element titleStmt (title statement)
<title>	Titre principal du document	aucun	TEI element title (title)
<author>	Nom auteur de l'édition numérique	aucun	TEI element author (author)
<publicationStmt>	Regroupe des informations concernant la publication ou la diffusion d’un texte électronique	aucun	TEI element publicationStmt (publication statement)

<authority>	fournit le nom d'une personne ou d'un autre organisme responsable : la mise à disposition d'une œuvre, autre qu'un éditeur ou un distributeur.	aucun	TEI element authority (release authority)
<date>	Date	aucun	TEI element date (date)
<sourceDesc>	Source description, décrit la source à partir de laquelle un texte électronique a été produit. habituellement une description bibliographique.	aucun	TEI element sourceDesc (source description)
<listBibl>	Liste de références bibliographiques	aucun	TEI element listBibl (citation list)
<bibl>	Référence bibliographique avec ici des sous-composants.	2 sources n="1" et n="2" (repérage)	TEI element bibl (bibliographic citation)
<publisher>	Editeur de la source originale	aucun	TEI element publisher (publisher)
<pubPlace>	Nom du lieu d'une publication	aucun	TEI element pubPlace (publication place)
<text>	Texte, contient un seul texte quelconque, simple ou composite. Ici avertissement, poème, note.	aucun	TEI element text (text)
<front>	Texte préliminaire, contient tout ce qui est au début du document, avant le corps du texte : page de titre, dédicaces, préfaces, etc.	aucun	TEI element front (front matter)
<titlePage>	Contient la page de titre d'un texte qui figure dans les parties liminaires	aucun	TEI element titlePage (title page)
<titlePart>	Title part, contient une section ou division du titre d'un ouvrage telle qu'elle est indiquée sur la page de titre.	rend avec = "center" pour le rendu	TEI element titlePart (title part)
<body>	corps du texte	aucun	-TEI element body (text body)
** **	Division du texte, contient une subdivision dans le texte préliminaire, dans le corps d'un texte ou dans le	type="avertissement" type ="poem"	TEI element div (text division)

	texte postliminaire. accompagné de type et subtype pour catégorisé et n=”x” pour sa position dans le document.	type=”notes” subtype="Satiric" n="X"	
<head>	En-tête,, ici le titre d'une section. Elle est accompagnée d’attributs type pour classer les éléments et rend=”center” pour spécifier la structure visuel souhaitée.	type=”title” type=”subtitle” rend=”center”	TEI element head (heading)
<p>	Paragraphes	aucun	TEI element div (text division)
<lg>	Contient un groupe de vers dans ce cas, ça aurait pu être une strophe. Mais l’ensemble de vers appartenant au poème.	aucun	TEI element lg (line group)
<l>	Vers attribut pour la mise en page ici alinéa	rend=”indent”	https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-l.html
<pb/>	Saut de page l’attribut indique le numéro de page quand mentionné dans la source d’origine sinon rien	n=”x”	TEI element pb (saut de page)
<note>	Contient une note avec l’attribut xml:id=”notex” qui permet le renvoie direct aux notes.	xml:id=”notex”	TEI element note (note)
<hi>	Mise en évidence distingue un mot ou une expression, ici mise en exposant, en italique ou en gras.	rend=”sup” ou rend=”italic” rend=”bold”	TEI element hi (highlighted)
<ref>	Définit une référence vers un autre emplacement. Ici permet de relier les exposants aux notes qui les concernent.	target="#note4"	TEI element ref (reference)