Medict37020d (Littré-Robin 1873), programme de balisage au 1er novembre 2023

# État

Le Littré-Robin 1873 est un dictionnaire spécialisé de termes de médecine, fortement structuré, avec notamment des sous-entrées, ainsi que des renvois, tant aux entrées principales, qu’aux sous-entrées. La source XML/TEI à date a été structurée pour alimenter une base de termes médicaux, réunissant 50 ouvrages, s’étendant sur 200 volumes, avec un niveau de structuration disparate, et généralement beaucoup moins élevé. Il en résulte que le fichier Medict37020d assure actuellement que les renvois retrouvés pointent vers les entrées principales, mais pas les sous-entrées. Par ailleurs, il n’a jusqu’ici jamais été envisagé de publier ce dictionnaire en ligne, le texte OCR n’étant pas corrigé.

# Objectif

L’objectif de ce programme de balisage est tout d’abord de restaurer des informations typographiques, afin de ne rien perdre de la typographie initiale (Ex : les renvois <ref> sont majoritairement en petites capitales, toutefois, certains mots en italiques peuvent être interprétés comme des renvois, mais la distinction typographique est pour l’instant perdue). L’objectif n’est plus seulement l’extraction d’informations terminologiques, mais de ne pas empêcher d’établir une édition électronique fac-similé (si des forces sont trouvée pour corriger le texte OCR).

Il en résulte désormais la doctrine suivante de balisage : les balises de fonctions lexicographiques sont des interprétations, la source est conservée sous forme de texte, des sauts de ligne, et d’attributs typographiques (normal est le cas par défaut, ne sont précisés que petites capitales et italiques).

Le deuxième objectif est d’augmenter le réseau des vedettes et des renvois, en assurant la structuration des sous vedettes.

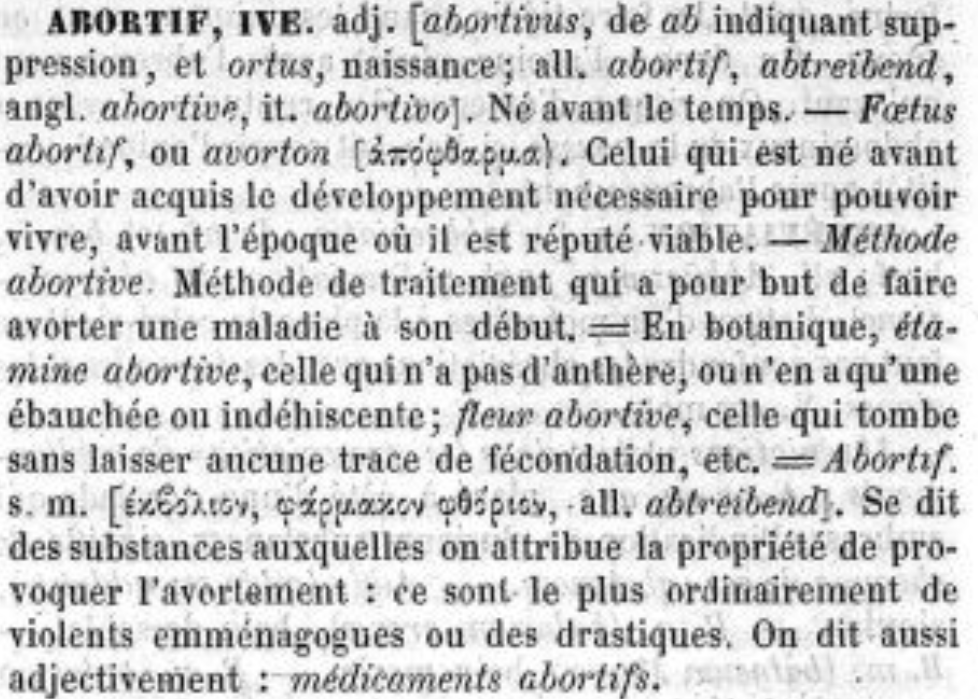
# Tâches

Description des tâches souhaitables

**Rétablir l’ordre des traductions (grec après latin si présent, ou en premier).**

XSLT [FG]

Structurer automatiquement des sous articles



État actuel du balisage