

## Table des matières

1. Présentation de WikiDiscussion.....	2
1.1 Extraction et sélection des pages de discussion.....	2
1.2 Structuration et formatage du corpus.....	2
2. La norme TEI-P5.....	3
2.1 Qu'est ce que la norme TEI-P5.....	3
2.2 La norme TEI-P5 pour les CMC.....	4
2.2.1 Contexte.....	4
2.2.1.1 Les formes décrites par le schéma.....	4
2.2.1.2 Problématiques.....	5
2.2.2 Description du schéma.....	5
2.2.2.1 TEI Header.....	5
2.2.2.2 Principe d'anonymisation.....	6
2.2.2.3 Éléments de la structure.....	6
2.3 Mise aux normes du corpus WikiDisc.....	8
3. Observation des structures.....	10
3.1 category @type= « autre ».....	10
3.2 Category @type!= « autre ».....	11
4. Modification du corpus d'après les observations menées.....	11
4.1 Sélection du corpus à modifier.....	11
4.2 Nettoyage du contenu des balises traitées.....	11
4.3 Modification ou suppression des balises.....	12
5. Balises observées relevant des pages parallèles.....	23

# 1. Présentation de WikiDiscussion

## 1.1 Extraction et sélection des pages de discussion

Les pages de discussion du corpus WikiDiscussion ont été extraites de la sauvegarde (ou *dump*) de la Wikipédia française à la date du 12 mai 2015, ce qui correspond à 3,5 millions de pages.

57 % de ces pages étant des pages de discussion « utilisateur », ces dernières n'ont pas été conservées. Parmi les pages de discussion « article » restantes, 8 % étaient des redirections, et 68 % ont été considérées comme vides (*i.e.* contenant moins de deux mots), et ces deux cas de figures n'ont pas été retenus non plus. Le corpus WikiDiscussion contient ainsi 65 612 pages de discussion « article » contenant au minimum 2 mots (Tableau 1).

Pages de discussion « utilisateur »	1 990 927	57%
Pages de discussion « article »	1 496 553	43%
Redirections	116 432	8%
Pages de discussion vide (> 2 mots)	1 013 791	68%
Pages de discussion conservées	366 326	24%

Tableau 1: Sélection des pages de discussion du corpus WikiDiscussion

## 1.2 Structuration et formatage du corpus

Les pages restantes ont ensuite été formatées en XML et découpées selon deux unités :

- Le niveau fil de discussion (*thread*) : correspond aux divisions délimitées par les entêtes signalées par les balises `/==.*?==/` dans le wikicode.
- Le niveau message (*post*) délimité par la date et parfois la signature de l'utilisateur, mais également par un changement dans le niveau d'interaction marqué par le nombre de « : » présents au début.

Après avoir délimité les fils de discussion et les messages, les métadonnées ont été encodées dans un en-tête, selon la norme TEI-P5 (*cf.* 2. La norme TEI-P5).

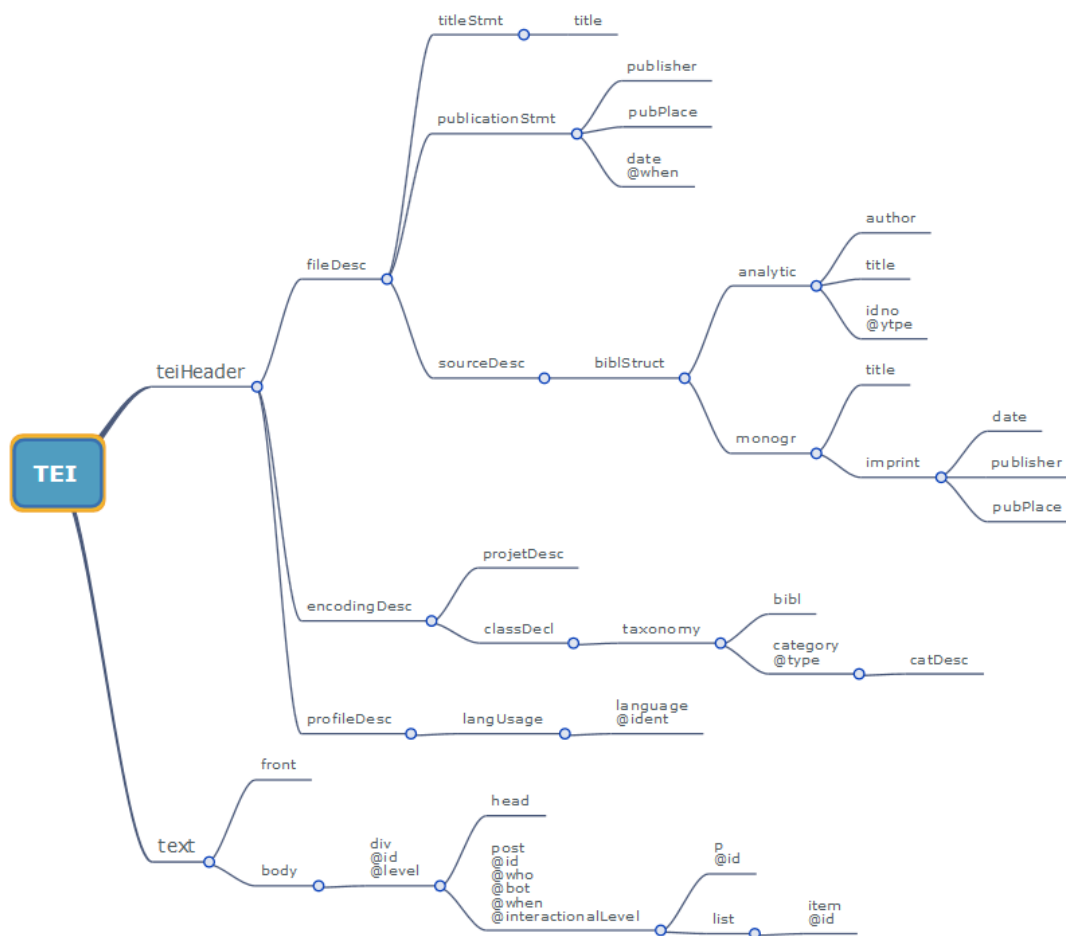


Illustration 1: Structure XML des pages de discussion

L'illustration 1 représente l'arbre XML des pages de discussion :

- La balise *teiHeader* est l'entête dans lequel sont encodées les métadonnées relatives à la page de discussion.
- La balise *text* correspond au contenu de la page de discussion : *div* représente un fil de discussion, composé d'un titre (*head*) et d'un ou plusieurs messages (*post*). Les messages peuvent contenir des paragraphes (*p*) ou bien des listes (*list*) d'objets (*item*).

### 1.3 Évaluation de l'extraction ?

## 2. La norme TEI-P5

### 2.1 Qu'est ce que la norme TEI-P5

La Text Encoding Initiative est un projet visant à établir des normes de structuration et d'encodage de textes électroniques afin de favoriser leur traitement et leur diffusion.

La représentation structurée de documents s'effectue au format XML, selon des schémas de balises et d'annotation strictes établis dans le cadre de ce projet. Ainsi, pour être valide, autrement dit reconnu conforme à la norme TEI-P5, un document doit respecter plusieurs critères.

1. Tout d'abord la structure XML du document doit être conforme aux recommandations à ce sujet du W3C.
2. Ensuite, le document doit respecter les schémas décrits dans la *guideline* du projet.

Un schéma conceptuel représente les différents éléments qu'il peut contenir, ainsi que leur hiérarchie. Ces éléments sont définis par des critères très strictes, et il n'est en aucun possible de changer la disposition des éléments ou bien leur type de contenu sans entraîner une annulation de la validité du document.

Exemples :

- L'élément <l> doit absolument contenir une ligne de vers, et non pas une ligne de prose ou un titre.
  - L'élément <p>, qui contient un paragraphe, est forcément contenu dans l'élément à la hiérarchie supérieure <body>, qui représente le corps de texte. Il est impossible, selon un schéma valide, que <p> contienne <body>.
3. Les schémas sont modulables et adaptables en fonction du genre et du contenu de chaque document, mais certains composants sont néanmoins obligatoires. Ainsi, un document est forcément composé d'un élément TeiHeader qui contient des informations sur le titre, la publication ou encore les sources du texte, suivi d'un ou plusieurs éléments issus de la classe d'éléments *model.ressourceLike*
  4. Un document TEI conforme est toujours accompagné d'un fichier .ODD qui décrit la structure du document XML en renseignant la sémantique des éléments définis. Ainsi, le fichier .ODD spécifie clairement les modules, éléments, classes utilisés dans le schéma, et sert également à valider la structure du document XML.

La norme TEI-P5 permet ainsi de décrire et de s'adapter à de nombreux genres textuels, allant de la poésie au théâtre en passant par le discours oral, en combinant différents éléments selon les schémas dédiés. Mais tous les genres textuels ne sont pas encore concernés, notamment des genres récents tels que

les communications médiées par ordinateur. C'est pourquoi des chercheurs travaillant sur ce type de corpus ont proposé une nouvelle version du schéma.

## 2.2 La norme TEI-P5 pour les CMC

Une équipe de chercheurs, Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, et Angelika Storrer, dans le cadre de recherches sur un corpus constitué de pages de discussion de la Wikipédia allemande (DÉRiK), ont décidé d'établir un schéma conforme à la norme TEI afin de décrire des documents représentant des communications médiées par ordinateur (désormais CMO). Ce schéma est présenté dans l'article *A TEI Schema for the Representation of Computer-mediated Communication*<sup>1</sup>.

### 2.2.1 Contexte

L'outil informatique a fait émerger de nombreux genres nouveaux de communications interpersonnelles qui ont, dans un premier temps, été rattachés à des conversations orales ou des échanges épistolaires. Or ces interactions nouvelles qui sont régies par l'adaptation des utilisateurs aux limites et aux capacités des nouvelles technologies, comportent des traits linguistiques et structurels qui peuvent pas être définis selon les traditionnelles descriptions de l'oral et de l'écrit.

Nos communications quotidiennes sont de plus en plus effectuées par des outils électroniques, et pourtant ces interactions sont encore mal documentées, notamment à cause de l'absence de modèles universels pour les décrire. Il est ainsi important d'établir des outils et des modèles permettant de représenter les CMO et leurs particularités linguistiques, afin de mieux les comprendre, de pouvoir les observer de manière empirique et à grande échelle, et surtout diffuser et partager les résultats.

C'est dans cette optique que l'équipe de Michael Beißwenger a cherché à établir un schéma XML conforme au cadre défini par TEI. Les CMO générant de nombreux types d'interaction différents, le schéma présenté dans l'article n'a pas pour ambition de pouvoir tous les décrire, mais a pour objectif de faire émerger une dynamique autour de ces recherches pour compléter et approfondir ce projet.

#### 2.2.1.1 Les formes décrites par le schéma

Le schéma se concentre sur les formes de communication suivant les protocoles TCP/IP pour l'échange de données, dont le principal moyen d'encodage de l'interaction est écrit. Ces formes de communication sont également dialogiques, c'est à dire qu'elles établissent un dialogue entre différents utilisateurs, dans lequel chaque utilisateur peut basculer du rôle de récepteur/lecteur à celui de producteur/auteur.

*Exemple* : fils de discussions d'un forum, chat, page de discussion Wikipédia, conversation sur les réseaux sociaux, etc.

---

<sup>1</sup> Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer, « A TEI Schema for the Representation of Computer-mediated Communication », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 15 October 2012, connection on 04 February 2017. URL : <http://jtei.revues.org/476> ; DOI : 10.4000/jtei.476

Ce schéma ne décrit donc pas les CMO qui ne transitent pas par internet, sont sous forme de monologue ou bien dont le moyen de communication est audiovisuel.

*Exemple* : Les SMS, les pages web statiques, les logiciels de visioconférence.

Afin de décrire au mieux ces interactions, le schéma respecte certains critères :

- S'adapter aux particularités structurelles des CMO qui sont principalement composées d'enchaînement d'interventions des différents interlocuteurs. Le but n'est donc pas de représenter les paragraphes ou les énoncés, mais bien les interventions.
- Fournir la possibilité d'annoter des éléments typiques de la langue utilisée sur internet (appelés *interaction signs*).
- Permettre l'anonymisation simple et réversible des données relatives aux utilisateurs.
- Fournir toutes les informations et métadonnées nécessaires pour exploiter les documents.

### **2.2.1.2 Problématiques**

La constitution du schéma résulte de la personnalisation de schémas et modules déjà établis par la TEI, mais dont la structure et les éléments n'étaient pas adaptés aux CMO. Il a ainsi fallu prendre en compte certaines problématiques afin d'adapter ce nouveau schéma :

- Comment représenter les contributions écrites des utilisateurs en tant qu'éléments principaux des dialogues.
- Comment définir les types spécifiques de groupement de plusieurs contributions en unités plus grandes (*Ex* : fil de discussion).
- Comment différencier :
  - La structure interne de chaque contribution : contrôlée par chaque utilisateur
  - La structure du dialogue : résultat de toutes les participations

## **2.2.2 Description du schéma**

Cette section se contente d'explicitier certains éléments considérés comme principaux et importants du schéma afin de mieux comprendre leur fonctionnement et leur emploi. Un exemple de représentation concrète de la structure du schéma est fourni dans la partie suivante (2.3)

### **2.2.2.1 TEI Header**

Cet élément n'est pas une spécificité du schéma CMC, mais un élément indispensable à n'importe quel document XML conforme à la norme TEI. Le header est ensemble d'informations relatives au document, et peut être constitué de cinq catégories principales.

1. **fileDesc** File description

Informations bibliographiques, précisions sur les sources.

(Exemple : titre, éléments relatifs à la publication, nom de la source, etc)

2. **encodingDesc** encoding description

Informations relatives au processus de numérisation et/ou de récupération du document.

(Exemple : catégories reliées au document lors de sa récupération)

3. **profileDesc** profile description

Informations contextuelles du document.

(Exemple : langue)

4. **xenoData**

Cet élément permet d'ajouter facilement des métadonnées non représentées dans le schéma TEI.

5. **revisionDesc** revision history

Permet d'ajouter des informations concernant d'éventuelles modifications effectuées sur le document (Exemple : date, auteur des modifications)

### 2.2.2.2 Principe d'anonymisation

Afin de faciliter la communication et la diffusion des recherches, le schéma permet d'anonymiser simplement et de manière réversible les données. Il répond ainsi à deux critères :

- Pour un observateur lambda des données : pouvoir associer chaque utilisateur à un ensemble de messages postés, sans que cet utilisateur ne soit identifiable en tant qu'individu du monde réel.
- Pour un observateur privilégié : avoir accès à plus d'informations et métadonnées concernant chaque utilisateur.

Cette stratégie d'anonymisation a été mise en place au moyen de plusieurs étapes.

Pour commencer, toutes les données personnelles disponibles peuvent être récupérées et affectées dans l'élément <person> (person profile). Cet élément fournit un attribut @xml:id auquel est associée une valeur unique dans le document XML. Autrement dit, <person> attribue à chaque utilisateur un identifiant unique. Tous les éléments <person> sont stockés dans l'élément <header>, plus particulièrement dans <sourceDesc> puis <listPerson>, et facilement dissociables de l'élément <body>, ce qui permet de les dissimuler aux lecteurs non autorisés.


Ainsi, chaque élément <posting> est relié à un profil par l'attribut @who qui porte non pas le vrai nom/pseudo de l'utilisateur, mais la valeur unique définie par l'attribut @xml:id de l'élément <person>.

### 2.2.2.3 Éléments de la structure

Le Tableau 1 décrit les éléments de la macrostructure du schéma, tandis que le Tableau 2 revient sur des éléments de la microstructure.

Élément	Description	Classe	Exemple
<posting>	La plus grande unité de structure pouvant être assignée à un auteur.  Pivot entre les plus hauts niveaux de la macrostructure et le contenu de la microstructure	<u>Élément</u> : model.divLike  <u>Attributs</u> : Att.ascribed att.datable att.global att.typed	<b>@synch</b> : heure du post <b>@who</b> : référence au profil de l'auteur <b>@revisedWhen</b> , <b>@revisedBy</b> : quand et par qui le post a été modifié <b>@indentLevel</b> : marque le niveau d'indentation (par un entier)
<div>	Description de la manière dont les messages sont organisés (ce n'est pas l'action d'un seul utilisateur, mais l'enchaînement des interventions et la gestion du server)	<u>Élément</u> : model.divLike  <u>Attributs</u> : deux attributs possibles	<b>@type</b> <b>=logfile</b> : organise les séquences de posts de manière chronologique <b>=thread</b> : structure les posts selon deux dimensions (dessus/dessous : relation avant/après, droite/gauche : indentation pour appuyer le lien d'un post avec le précédent)

Tableau 1: Éléments de la macrostructure

Balise	Description	Exemple	Nature	Attributs Exemples
<emoticon>	-Unité iconique générée grâce au clavier -Utilisée pour représenter des expressions faciales et marquer des émotions, de l'ironie	:-) :-( :o/  Contient caractères non Unicode et glyphes	<u>Élément</u> : gLike  <u>Attributs</u> : att.global att.typed	<b>@style</b> : origine (Western, Japanes, Korean, Other) <b>@systemicFunction</b> : sens de l'expression <b>@contextFunction</b> : interprétation en contexte de l'expression (emotionMarker:positive negative neutral, unspec, responsive, ironyMarker, illocutionMarker, virtualEvent) <b>@topology</b> : position de l'unité par rapport au texte (front_position, back_position, intermediate_position, standalone.)
<interactionWord>	Unité linguistique symbolique : construction morphologique basée sur un mot/ une phrase décrivant, symbolisant une expression, une action corporelle, une gestuelle	*Sing* *sort* L'unité linguistique illustre l'action virtuelle du locuteur	<u>Élément</u> : model.global.spoken  <u>Attributs</u> : att.global att.typed	<b>@formType</b> : propriétés morphologiques (simple, complex, and abbreviated) <b>@systemicFunction</b> <b>@contextFunction</b> <b>@topology</b> <b>@semioticSource</b> (mimic, gesture, bodilyReaction, action, sentiment, sound, process, emotion)
<interactionTemples>	Unités non générées par le clavier : l'utilisateur active un modèle qui introduit un élément graphique ou textuel prédéfini		<u>Élément</u> : model.global  <u>Attributs</u> : att.global	<b>@type</b> (iconic, verbal, iconicverbal) <b>@motion</b> (static, animated) <b>@systemicFunction</b> <b>@contextFunction</b>



<addressingTerms>	Lorsqu'un utilisateur s'adresse à un ou plusieurs utilisateurs ciblés.	@toto to:jeanjacques	Élément : model.nameLike  > le contenu est restreint à deux éléments : voir cases suivantes  Attributs : att.global	
<addressMarker>	Unité marquant le fait que l'utilisateur s'adresse à une cible	@ to :	Élément : model.labelLike  Attributs : att.global	
<addressee>	Utilisateurs ciblés	toto	Élément : model.nameLike.agent  Attributs : att.global	<b>@who</b> : adressé sous la forme de son id <b>@scope</b> : une ou plusieurs personnes (group, all, individual, nospec) <b>@formType</b> (persNameFull, persNameAbbreviation, persNameNickname)
<autoSignature>	Position de la signature de l'utilisateur > le contenu de la signature est récupéré dans la balise <person>		Élément : model.pPart.edit	<b>@target</b> : référence à l'élément <signatureContent> qui contient la signature et qui se trouve dans <person>
<seg>	Unités similaires à celles utilisées dans échanges épistolaires		Élément model.segLike	<b>@type</b> =postscript =opener =closer

Tableau 2: Éléments de la microstructure

## 2.3 Mise aux normes du corpus WikiDisc

Le corpus WikiDisc n'est pas encore complètement conforme aux normes proposées par le schéma de TEI-P5 pour les communications médiées par ordinateur. Voici les différents éléments à modifier afin de rendre la structure XML des pages de discussion du corpus conforme.

- Anonymisation des utilisateurs

Pour l'instant, le pseudo des utilisateurs ayant posté des messages au sein des fils de discussion est précisé directement dans l'élément *post* sous l'attribut *@who*. Il faudrait ainsi attribuer un identifiant numéral à

chaque utilisateur, et regrouper les informations reliant le pseudo de l'utilisateur à son identifiant unique dans l'élément `<person>`, situé en `<teiHeader><sourceDesc><listPerson>`.

- Regrouper les informations temporelles

Les informations concernant la date de chaque message posté sont actuellement précisées directement dans l'élément `post` sous l'attribut `@when`. Le schéma préconise de regrouper ces informations dans un élément `<timeline>`, situé dans `<text><front>`. Cet élément contiendra ainsi plusieurs éléments *when*, au sein desquels la date sera précisée, et un identifiant unique leur sera attribué, identifiant qui sera précisé dans les éléments *post*.

- Changer de dénomination

A l'heure actuelle, les éléments de la structure XML qui correspondent aux messages postés par les utilisateurs sont appelés `<post>` tandis que dans le schéma, ces éléments sont appelés `<posting>`.

- Utiliser des éléments de la microstructure

Il existe en effet des éléments pour caractériser des séquences particulières au sein même du contenu des messages dans le schéma, mais ils ne sont pas employés dans le corpus `wikiDisc`. Ces éléments peuvent être particulièrement intéressants, notamment celui pointant la signature des utilisateurs, `<autoSignature>`, qui renvoie à la signature présente dans l'élément `<person>` correspondant à l'utilisateur, et permettant ainsi d'anonymiser les messages.

- Préciser le type des `div`

Il est en effet préconisé dans le schéma de préciser le type des `div`. Dans le cas du corpus `WikiDisc`, les `div` sont de type *thread*, car elles correspondent à un fil de discussion. Par ailleurs, le numéro et le niveau d'indentation des `div` sont précisés dans `wikiDisc`, alors que ce n'est pas obligatoire dans le schéma.

### 3. Observation des structures

L'observation de certaines pages de discussion a permis de mettre en évidence des problèmes dans la structure, et plus particulièrement au sein du contenu parfois confus des balises *category*, situées dans le *header*. Afin de mieux comprendre ce phénomène, nous avons observé séparément les éléments *category* dont l'attribut *type* était « autre » et les éléments *category* restantes.

#### 3.1 category @type= « autre »

Dans un premier temps, les balises *category* avec un attribut *type*= « autre », ont été identifiées comme les plus déviantes. Ces éléments possèdent des éléments enfants, *catDesc*, qui contiennent les séquences potentiellement problématiques à observer. Cette observation a fait émerger trois cas de figure :

- **Balises significatives** : certains contenus sont porteurs d'informations à garder, ces derniers ont été identifiés et regroupés dans un tableau contenant le nom de l'information, le mot clef attribué à cette information ainsi qu'une explication, bien souvent un extrait de la page Wikipédia.  
Traitement : le mot clef est intégré dans la structure XML de la page de discussion, sous forme d'attribut de l'élément *category*, afin d'y accéder plus facilement.  
Exemple : traduction, projet Tintin, page soupçonnée de faire de la publicité, etc.
- **Balises bruit** : certains contenus relèvent surtout de la mise en page ou de la structure de la page de discussion originale, ou alors portent des informations redondantes déjà identifiées par les balises significatives.  
Traitement : ces balises sont ignorées et effacées de la structure XML, mais sont malgré tout relevées et expliquées dans un tableau.  
Exemple : balise indiquant le menu déroulant, balise de redirection vers une autre page WP, etc.
- **Balises problème d'extraction** : deux cas de figures majeurs ont été repérés comme dus à un problème lors de l'extraction.
  - Premièrement, certains messages, devant initialement se situer dans la partie *text* de la structure, contenant les messages de la page de discussion, se retrouvent dans *category*. Étant les plus compliqués à repérer automatiquement car n'ayant pas de signe distinctif récurrent, les contenus considérés comme des messages sont identifiés par élimination. En d'autres termes, tous les contenus n'ayant pas été catégorisés dans l'une des autres classes de balises décrites ici, sont considérés comme des messages.  
Traitement : les messages sont déplacés dans la bonne partie de la structure XML de la page de discussion : en tant qu'élément *post*, dans l'élément *div* id=0 (<text><body><div><post>)
  - Deuxièmement, certaines informations telles que le projet rattaché à l'article ou le niveau d'avancement, sont initialement contenues dans les balises *category* du type adapté (en l'occurrence discipline ou avancement), se retrouvent au mauvais endroit, c'est à dire dans un élément *category* de type *autre*.  
Traitement : remettre ces informations dans les balises adéquates.  
Exemple : avancement, contenu normalement dans *category type*= « avancement »

## 3.2 Category @type!= « autre »

Le contenu des *catDesc* des balises *category* ayant un attribut *type* différent de *autre* a été récupéré dans une table de fréquence. Afin d'identifier dès le début d'éventuels contenus déviants, définis de manière arbitraire par une longueur supérieure à 50 caractères, le script produisant la table de fréquence permet également de sortir dans un document annexe et d'ignorer dans la table de fréquence les contenus dépassant le seuil.

Les résultats n'ont sorti qu'un seul contenu déviant, pouvant donc être corrigé manuellement.

Par ailleurs la table de fréquence a permis d'examiner une à une toutes les catégories, et de contrôler si le *type* associé était cohérent. Aucune modification n'a semblé nécessaire suite à cette observation.

## 4. Modification du corpus d'après les observations menées

Le contrôle et l'identification des erreurs à corriger dans les pages XML ont permis de mettre en évidence les manipulations à effectuer afin d'ajuster le corpus. Ces modifications ont été apportées en plusieurs étapes.

### 4.1 Sélection du corpus à modifier

Le corpus est constitué de pages de discussion principales et de pages de discussion parallèles. Ces dernières sont des pages de discussion au sein desquelles les thèmes abordés sont très ciblés (suppression, article de qualité, droits d'auteur, etc). Les modifications décrites par la suite ne sont effectuées que sur les pages de discussion principales, ainsi que sur les pages de discussion dites archivées, qui sont des pages de discussion principales mises de côté.

### 4.2 Nettoyage du contenu des balises traitées

Avant de modifier ou supprimer les balises, un premier nettoyage consistant à supprimer certaines séquences relevant de problèmes d'encodage ou d'extraction a été effectué au sein des balises, car elle étaient porteuse de bruit. Ainsi, toutes les séquences du Tableau 1 ont été préalablement supprimées du contenu des balises traitées.

Balise
{{NAMESPACE}}
{{BASEPAGENAME}}
&amp;lt;
&amp;gt;
centerbig
centerbig
/big/center
&amp;quot ;
center

br
small
nowiki
big
}}
div

Tableau 1 : Séquences bruit supprimées lors du nettoyage

## 4.3 Modification ou suppression des balises

Les balises relevées comme étant problématiques ont donc été traitées différemment selon leur type : les balises dites « bruit » ont été supprimées des documents, et les balises significatives ont été conservées et modifiées en fonction de leur nature. Le Tableau 2 contient un échantillon de contenu considéré comme « bruit ».

Contenu de la balise identifié	Raisons de la suppression
Instructions LANN	Redondant avec Archive LANN
*Commentaire	Commentaire suite à un vote
fait	Redondant avec autre balise
{{Boîte déroulante {{Palette {{liste de vérification	Structure de la page
{{Début cadre {{saut 170px}} {{DEFAULTSORT	Structure de la page
{{Modèle :	Structure de la page
{{ligne grise	Structure de la page
{{Titre en italique	Structure de la page
{{Aller au sommaire	Redirection vers section de la page
{{Sommaire à droite	Redirection vers section de la page
{{Sommaire	Redirection vers section de la page
{{Voir aussi Wikipédia	Redirection vers section WP
état  demandeur etc	Précisions après demande de relecture
{{#switch	Après page supprimée
{{ns :	Après page supprimée
#default	Après page supprimée
{{Auteurs crédités	
{{En-tête de page de discussion	Structure de la page
{{sources	Indique sources
trad cop scis fus	Précisions après wikicrédit
{{!}} [[Image	Image
{{ok}}	Puce de check liste état « OK »

<span>{{clr}}</span>	
Discussions archivées [[../Archive	Redirection vers les autres discussions archivées de la page
'''Cette page est un &lt;u>{{rouge espace de travail}}''' '''Cette page est un u{{rouge espace de travail}}/u''' [[Fichier:Tecleando.gif]]	Balises espace de travail, indiquée sur page bac à sable
* &lt;strong>Depuis mai 2007 : Page principale de discussion&lt;/strong>	Indique que c'est la page principale de discussion
[[Catégorie:Wikipédia:Archives Pàs supprimé {{SUBPAGENAME}}]]	Accompagne certaines pages thème « suppression »
<span>{{étonné}}</span>	Smiley
[[Catégorie:Wikipédia:Archives Pàs supprimé {{ARTICLESPACE}} ]]	Suit toujours balise
[[../Archive ou [[Discussion:	Lien vers les autres discussions archivées
(actu) (diff) (+/-) (diff) # (diff)	Liste auteurs dont le nom a été effacé suite à destruction de l'historique/fusion
# s(	Liste de choses à faire
<span>{{nobr</span>	Liste règles WP dans tableau
<span>{{!}}}</span> <span>{{\!}}</span> padding:1em;padding-top:0\,5em;&quot; {{#ifeq: {{encart important {{titre mis en forme {{devoir}} {{traité}} {{done}} {{DR début {{Article principal	Balises éléments de structures peu récurrents
<span>{{Suivi de traduction</span>	Suivi de traduction / traduction en cours
<span>{{Suivi</span>	Suivi → ne porte pas d'information pertinente
***** _____	
<span>{{Tableau GroupeComm {{ébauche intercomm</span>	Modèle de tableau encadré pour article donné en exemple
'''[{{SERVER}}{{localurl:{{PAGENAME}}  action=edit&amp;section=new}} Pour commencer une nouvelle discussion cliquez ici]'''	Direction vers ouverture de nouvelle page
<span>{{Terres Contées}}</span>	Précision que le bandeau terres contées a été retiré
<span>{{Notif</span>	Balise indiquant image flèche verte
<span>{{Méta bandeau d'avertissement {{Avertissement}}</span>	Affichage d'un bandeau d'avertissement → divers thèmes abordés dans ce bandeau Affichage triangle rouge

{{Infobox	Marque début d'une infobox → pas de contenu à retenir
{{Wikipédia:Pages à supprimer/préchargement2	Encart mal chargé
{{Crédit d'auteurs	Cet article est partiellement ou en totalité issu de l'article intitulé « Ensemble Théâtral Mobile » (voir la liste des auteurs).
centerCeci est une autorisation globale pour l'ensemble des articles créés	Suit confirmation OTRS
Commence par	Balise mal découpée → éléments qui sont censés se suivrent
{{drapeau	Balise avec image de drapeau
!-- Concernant l'article original – !--NE RIEN MODIFIER SUR LES DEUX etc	Commentaire → pas d'info à récupérer  Problème : ne fonctionne pas avec ^ → sans : toute balise contenant com supprimée
{{supprimer	Balise image
{{BS3bis {{BS5bis	Balise structure schéma
{{Utilisateur GIMPS}}	Balise « userbox » gimps
{  {{#ifeq:{{{}}} yes class=messagebox-talk class=messagebox-standard-talk}}	Balise box
'''Bienvenue sur Wikipédia	Balise bienvenue sur WP
{{#ifeq: Discussion    }}	
{{avancement	Barre pourcentage d'avancement





- OTRS  
teiHeader/profileDesc/textDesc/derivation @type = 'traditional' texte = texte  
Le contenu ou une partie du contenu est issu d'une source extérieure

```
#1 Traduction
if re.search(r"^ *({ ?(T|t)radu?(it|ction)( |_)(de)?.*?|({ ?(T|t)raduit)", txt):

    #Compteur
    if "Traduction" not in freq :
        freq["Traduction"] = 1
    else :
        freq["Traduction"] = freq["Traduction"] + 1

    #Supprimer l'element actuel
    cat.remove(catDesc)
    section.remove(cat)

    #Deplacer l'information dans l'element adequat
    trad = root.find('teiHeader/profileDesc')
    trad.append(etree.Element('textDesc'))
    trad2 = root.find('teiHeader/profileDesc/textDesc')
    trad2.append(etree.Element('derivation'))
    deriv = root.find('teiHeader/profileDesc/textDesc/derivation')

    deriv.set('type', 'translation')
    deriv.text = txt

    m1 = "Trad_"

    sos.write("\tx")
```

Illustration 2: Extrait du script - partie traitant les balises "Traduction"

Numéro Ligne dans le script	Catégorie	Mot clef attribué	Texte dans wikipédia	Nombre de pages de discussion concernées
#1 86	Traduction	Traduction		9498
#2 111	Contenu issu d'une source extérieure sous licence creative commons	OTRS	<p>OTRS icon      Le contenu de cet article est issu en tout ou partie de <a href="http://webwave.fr/Blog/?p=220">http://webwave.fr/Blog/?p=220</a>. La permission de distribuer ce travail sous la licence Creative Commons paternité partage à l'identique 3.0 (CC BY-SA 3.0) a été reçue sur OTRS, via le ticket no 2010031710048876.</p> <p>Ce modèle est utilisé par des volontaires autorisés sur le système de suivi de tickets (OTRS) de la Wikimedia Foundation. Il ne doit être apposé qu'après la réception d'une autorisation claire de permission, via permissions-fr@wikimedia.org. Pour demander une permission, n'utilisez pas ce modèle mais consultez Aide:Republication. En cas de doute sur une autorisation, vous pouvez demander sur le bulletin des agents OTRS de vérifier la validité de celle-ci.</p>	669

Tableau 3: Synthèse 1 des balises significatives

- Soit la balise précise une information sur l'article même, et n'a pas à voir avec ses source. Dans ce cas, l'élément *catDesc* reste tel quel, mais un mot clef vient préciser la présence de l'information dans l'élément parent *category*, afin d'en améliorer la visibilité.

```
#42 Projet Tintin
elif re.search(r"^ *{{ALaDecouverteDeTintin", txt) :
    #Compteur
    if "Tintin" not in freq :
        freq["Tintin"] = 1
    else :
        freq["Tintin"] = freq["Tintin"] + 1

    #Rajout de l'attribut info
    cat.set('info', 'Tintin')
```

*Illustration 3: Extrait du script - Traitement des balises "Tintin"*

Le Tableau 4 synthétise toutes ces balises, en précisant leur numéro et leur mot clef attribué, mais également la ligne à laquelle elles sont traitées dans le script (Illustration 3), un petit texte explicatif issu de la wikipédia, ainsi que le nombre de pages de discussions qui contiennent ce type de balise.

Num	Catégorie	Mot clef attribué	Texte dans la wikipédia	Nombre de pages de discussion concernées
#3 160	To do	Todo		8001
#4 174	Ancien AdQ	Ancien AdQ		259
#5 188	Archives	Archives		642
#6 202	Page à supprimer	PàS	Suppression de la page débattue	642
#10 254	Intention de proposer article au label BA	Intention BA		254
#11 267	Historique détruit	Historique détruit		703
#12 280	Page conservée	Page conservée		3746
#14 306	Avertissement hors-sujet Discussion hors-sujet Hors-sujet	HS	<p>Cette page de discussion sert uniquement à aider à la coordination de l'amélioration de l'article, et non à engager une discussion à propos de l'intérêt ou de la pertinence du sujet. N'utilisez pas cette page comme un forum de discussion.</p> <p>Bonjour, Vous êtes sur la page de discussion concernant l'article Nâdiya de l'encyclopédie Wikipédia. Cette page est réservée aux discussions visant à l'élaboration et l'amélioration de l'article correspondant. La Fondation Wikimedia n'est en aucune façon liée aux personnes physiques ou morales citées dans ces pages. Aussi, merci de ne pas déposer de message personnel ou de demande d'information complémentaire sur cette page et référez-vous si possible au site officiel de la personne ou de la société en question. Pour toute question non spécifiquement liée à l'élaboration des articles encyclopédiques, vous êtes cordialement invité à poser vos questions à l'Oracle où d'autres contributeurs vous répondront dans la mesure de leurs possibilités</p>	207
#17 345	Demande d'atelier de relecture	Demande relecture		30
#18 358	Invitation au projet éphéméride	Ephemeride		354
#20 384	Article copié par un site / livre	Copie	Le contenu, ou une partie du contenu, de cet article de Wikipédia a été repris par le livre ou la revue : Lille Métropole 2013 écrit par Dominique Auzias et Jean-Paul Labourdette sans mention de la licence CC-BY-SA 3.0 (ou le cas échéant de la GFDL) ni des auteurs. L'article de Wikipédia n'enfreint donc pas le droit d'auteur sur ce point.	211
#21 397	Avertissement Homonymie	Homonymie		150
#22 410	Avertissement ISBN	ISBN		34

#23 423	Page à vérifier	PàV		452
27	Projet (divers)	Projet		615
#26 463	EPDD	EPDD		57
#27 476	Article projet Valdensia	Valdensia	Projet Valdensia Cet article a été traité par le projet Valdensia. Dans le cas où cet article a été créé par un robot, sa source principale est l'article correspondant sur la base de données de la Bibliothèque cantonale et universitaire de Lausanne (BCU).	79
#28 489	Article mentionné par la presse	Presse		31
#29 502	Article conservé	Article conservé	Sursis accordé après débat de suppression OU Cette page a été anciennement proposée à la suppression : il a été décidé de la conserver suite à ce vote. Si vous souhaitez la proposer à nouveau à la suppression, veuillez consulter la partie « Comment proposer à nouveau une page pour la suppression ? » de Wikipédia:Pages à supprimer.	13
#30 516	Pub	Pub	Le ton de cet article ou de cette section est trop promotionnel ou publicitaire	2
#31 530	Wikicrédit	Wikicrédit	Après copie/trad/scission etc : indique crédits → précisions souvent mal placées : split dans balises suivantes (à récupérer ? Pour l'instant ignorées)	103
35	Contenu reporté	Contenu reporté	Indique que du contenu de l'article a été reporté dans un autre article	14
#32 543	Article indexé Rivalités dans le football	Rivalité football	L'article Classico mouloudéen fait partie d'un ensemble d'articles traitant des rivalités dans le football géré par le groupe de travail sur les rivalités du Projet:Football. Si vous êtes intéressés n'hésitez pas à améliorer les articles, voire à nous rejoindre.	9
38	Portail	Portail	Indique portail thématique → possibilité de récupérer le thème du portail	13
#34 569	Relecture	Relecture	Un émissaire a contacté Boungawa, ce dernier étant un spécialiste du sujet. Après relecture de l'article, cela a abouti à cette version.  OU  Cet article a été relu par l'Atelier de lecture.	16
#35 582	Article à illustrer	Illustrer	Cet article a besoin d'être illustré (en discuter) (indiquez la date de pose grâce au paramètre date).  Pour améliorer cet article, des vidéos sous licence libre ou du domaine public sont les bienvenus. Si vous êtes l'auteur d'un média que vous souhaitez partager, importez-le. Si vous n'êtes pas l'auteur, vous pouvez néanmoins faire une demande de libération d'image à son auteur.	3
#36 595	Film Disney Standard	Disney	Cet article a été standardisé selon le modèle : Film Disney (rubriques, mise en page, etc.)	56

#37 608	Discussion biographie	Biographie	Cette page est destinée à accueillir les discussions à propos de l'article Chris Benoit de Wikipédia. Elle sert à formuler des remarques, commentaires ou suggestions quant au fond ou à la forme de l'article. Elle ne sert ni de forum, ni à entrer en contact avec la personne ou l'institution à laquelle cet article est consacré.	12
#38 621	En travaux → Pour l'instant rencontré uniquement sur pages BâS : à conserver ou redondant ?	Travaux	Cette page est en cours de réécriture ou de restructuration importante (indiquez la date de pose grâce au paramètre date).  Un utilisateur prévoit de modifier cet article pendant quelques jours. Vous êtes invité(e) à en discuter en page de discussion et à participer à son amélioration de préférence en concertation pour des modifications de fond.	1
#39 634	Guerre d'édition	R3R	Bonjour. Vous semblez participer à une guerre d'édition. La recommandation appelée règle des trois révocations dit qu'un contributeur ne peut effectuer trois révocations ou davantage sur tout ou partie d'un article sur une durée de 24 heures consécutives. Un compte utilisateur qui révoque les contributions d'autres wikipédiens sur un même article à plusieurs reprises peut se voir bloqué pour éviter une guerre d'édition. Merci de cesser immédiatement vos révocations et de rechercher préférentiellement un consensus avec les autres wikipédiens en engageant avec eux un dialogue constructif . Merci.	2
#40 647	Wikipédia takes Montréal	WTM	Wikipedia Takes Montreal L'article « Jardin botanique de Montréal » a été sélectionné par les membres du projet:Québec pour figurer sur la liste des cibles à photographier le dimanche, 28 août 2011, lors de la chasse photographique Wikipedia Takes Montreal.	12
#41 660	Wikiconcours	Wikiconcours	L'article Drone a été choisi par l'équipe n°620 pour participer au Wikiconcours septembre 2010 qui s'est déroulé du 15 septembre au 14 novembre 2010.  Le Wikiconcours est l'événement rédactionnel collaboratif par excellence de Wikipédia. D'abord divertissant et enrichissant, il est ouvert à tous sans exception, même si vous n'avez jamais édité sur Wikipédia.	64
#42 673	Projet Tintin	Tintin	Les wikipédiens ont formé le projet Tintin afin d'organiser l'écriture autour des articles liés aux Aventures de Tintin. Tous ceux qui veulent s'intégrer au projet sont les bienvenus Sourire. Vous pouvez consulter la page de discussion du projet, y poser vos questions et répondre à celles des autres, ou bien contacter directement l'un des participants au projet	23
#43 686	Concours automne français	Automne	L'article « Femen » a été créé ou amélioré pendant le concours « Automne français ». La taille et la qualité de cet article en date du 10 décembre 2011 a déterminé le score final de l'utilisateur qui l'a créé ou amélioré.	10
#45 699	Du bistro	Bistro	Note : le texte ci-dessous a été déplacé depuis le Bistro de Wikipédia.	25
#46 712		ASS	???	64
#47 725	Blanchiment de courtoisie	Blanchiment	Ceci est l'archive d'une page blanchie en vertu du blanchiment de courtoisie.	6

#48 738	Bandeau article audio	Audio	La version audio de cet article a été enregistrée à partir de cette version de l'article dans le cadre du Projet Articles audio, un projet visant à proposer des versions lues à voix haute des articles de Wikipédia. Si vous souhaitez participer à ce projet, vous pouvez consulter la page principale afin de le découvrir.	22
#49 751	Confusion	Confusion	Cette discussion traite de l'article Environnement naturel. Pour les autres sens du mot Environnement, voir l'article Environnement (homonymie).	1
#50 764	Champions Foot	Champions Foot	L'article Paris Saint-Germain Football Club fait partie d'un ensemble d'articles concerné par la création d'un thème de qualité sur les clubs de football champions de France. Si vous souhaitez participer à ce projet, nous vous invitons à venir suivre l'évolution de ce dernier sur la page du projet.	19
#51 777	Admissibilité	Admissibilité	L'admissibilité de cet article est à vérifier (mars 2015). Motif : l'utilisateur qui a apposé ce bandeau n'a pas précisé le motif de la pose. Vous êtes invité à le compléter pour expliciter son admissibilité, en y apportant des sources secondaires de qualité, ainsi qu'à discuter de son admissibilité. Si rien n'est fait, cet article sera proposé à la suppression un an au plus tard après la mise en place de ce bandeau.	4
#52 790	À sourcer	Source	Ce modèle est-il pertinent ? Cliquez pour en voir d'autres. Cet article ne cite pas suffisamment ses sources (indiquez la date de pose grâce au paramètre date).  Si vous disposez d'ouvrages ou d'articles de référence ou si vous connaissez des sites web de qualité traitant du thème abordé ici, merci de compléter l'article en donnant les références utiles à sa vérifiabilité et en les liant à la section « Notes et références » (modifier l'article, comment ajouter mes sources ?).	8
#53 803	WP n'est pas un blog (animal)	Blog Animal	Wikipédia n'est pas un blog ! Wikipédia est une encyclopédie, seul un professionnel est apte à fournir un avis vétérinaire ou des conseils d'élevage. Si vous souhaitez partager votre expérience personnelle vous pouvez contribuer utilement à un guide d'élevage sur Wikibooks. Il existe aussi de nombreux sites internet, hors de Wikipédia, qui vous offriront des conseils pratiques et des forums sur votre animal de compagnie favori. Chacun est en revanche cordialement invité à ajouter des données ou documents d'intérêt encyclopédique comme spécifié dans l'article zoologique type, en prenant soin d'en citer les sources en plaçant des annotations dans le texte. L'article idéal n'existe pas, mais il pourrait ressembler à celui-ci : Beagle, ou celui là: Chat persan.	4

#55 829	Nom de cultivar	Cultivar	Conformément au Code international pour la nomenclature des plantes cultivées, un nom de cultivar s'écrit toujours en caractères romains minuscules (jamais en italiques), avec une majuscule au début de chaque nom et doit être mis entre guillemets simples (ouvrant et fermant). Ces noms de cultivars sont en principe écrits dans une langue vivante (pas en latin, sauf avant 1959) et ne doivent pas être traduits, mais le cas échéant transcrits en alphabet latin. Exemple : Rose 'Zéphirine Drouhin' ou Rosa 'Zéphirine Drouhin'	2
#56 842	Salon de thé → projet Al Andalus	Andalus	Le Projet:Al-Andalus est un projet Wikipédia concernant al-Andalus, période de l'Histoire médiévale de la péninsule ibérique dite « mauresque ». Il est destiné à faciliter les travaux de l'ensemble des contributeurs, même occasionnels, qui souhaitent intervenir sur ce domaine.	15
#57 855	Bon thème national foot	Bon Foot	Projet:Football L'article Vendée Poiré-sur-Vie Football fait partie d'un ensemble d'articles concerné par la création d'un bon thème sur les clubs participant au Championnat de France de football National 2014-2015. Si vous souhaitez participer à ce projet, nous vous invitons à venir suivre l'évolution de ce dernier sur la page du projet.	14
#58 868	Aidez-moi	Aide	Modèle utilisateur demande de l'aide	50
#60 881	Bouillet	Bouillet	Pour information, voici l'article ADALBERT (S) extrait du Dictionnaire d'histoire et de géographie Bouillet Chassang paru en 1878. (Voir : (=> projet Wikipédia))  ADALBERT (S) évêque de Prague, l'apôtre des Prussiens, prêcha la religion en Bohême, en Hongrie et en Prusse, et périt martyr en 997, à 58 ans On le fête le 29 avril.  OU  Cet article comprend des extraits du Dictionnaire Bouillet. Il est possible de supprimer cette indication, si le texte reflète le savoir actuel sur ce thème, si les sources sont citées, s'il satisfait aux exigences linguistiques actuelles et s'il ne contient pas de propos qui vont à l'encontre des règles de neutralité de Wikipédia. Les informations données sont peut-être désormais erronées ou incorrectes : vous pouvez partager vos connaissances en améliorant ou en modifiant cet article.	4
#61 894	Rappel de critères	Critères	Ceci est un extrait de la recommandation « Wikipédia:Notoriété des personnalités politiques ».  Ces critères servent à évaluer le potentiel d'un article. Ils n'ont qu'une valeur indicative. Un article qui ne répond à aucun de ces critères, en plus d'être vérifiable, n'est pas forcément à supprimer. Et de même, un article remplissant une des conditions n'est pas toujours à conserver.  OU  Pour rappel les critères d'admissibilité des articles pour les musiciens sont donnés ci-dessous.	4

#62 907	Événement en cours	Event	Cet article concerne un événement en cours.  Ces informations peuvent manquer de recul, ne pas prendre en compte des développements récents ou changer à mesure que l'événement progresse. Le titre lui-même peut être provisoire. N'hésitez pas à l'améliorer en veillant à citer vos sources. Dernière modification de cette page le 25 mars 2015, à 17:07.	-
#63 920	Ébauche	Ebauche	Cet article est une ébauche concernant la mer.  Vous pouvez partager vos connaissances en l'améliorant (comment ?) selon les recommandations des projets correspondants.	10
#64 933	Formation	Formation	Article en cours de modification  Un utilisateur en formation dans le cadre du projet de la classe de FLE (Kansai Gaidai) travaille en ce moment même sur cet article ou cette section. Ne pas modifier pour éviter les conflits de versions.	11
#65 946	Wiki Africa	Wikiafrica	Cet article a été créé ou agrandi par Festival de Cinéma Africain de Cordoue-FCAT dans le cadre du projet Share Your Knowledge (it) développé au sein de WikiAfrica (it).	2
#66 958	Article numéro n	Ne	L'article Sheryl Crow est le 90 000e article de Wikipédia en français, créé le 14 mars 2005. (voir l'historique)	12

Tableau 4: Synthèse 2 des balises significatives

- Soit la balise contient des informations qui se situent normalement dans un élément catDesc mais de type différent de autre : les balises wiki projet. Ces dernières portent des informations qui doivent être déplacées dans :
  - catDesc @type= « discipline »
  - catDesc @type= « avancement »

Dans ce cas la balise est supprimée, et les informations sont remises dans les bons éléments.

## 5. Balises observées relevant des pages parallèles

Lors de la phase d'observation, certaines balises mal constituées ont été relevées, mais la division du corpus en pages principales et parallèle n'a pas permis de les modifier car elle n'étaient présentes que dans les pages parallèles. Ces balises sont synthétisées dans le Tableau 1 à titre indicatif.



Num	Catégorie	Mot clef attribué	Texte dans wikipédia
#7 215	Page soupçonnée de violation de copyright	PCP	
#8 228	Article non neutre	LANN	
#9 241	Article rejeté en AdQ	Rejet AdQ	
#13 293	Article déchu en AdQ	Déchu AdQ	
#15 319	Article promu en AdQ	Promu AdQ	
#16 332	Article mis en lumière sur page d'accueil	Lumiere	
19	Article supprimé	Article supprimé	
#19 371	Article promu en BA	Promu BA	
#33 556	Suppression à revoir	Suppression à revoir	<p>Des éléments significatifs sur l'admissibilité de l'article ont été apportés depuis la proposition de suppression.</p> <p>Nous invitons les contributeurs ayant émis une opinion à son sujet à consulter la discussion en cours et à modifier leur opinion s'ils estiment que cela est nécessaire.</p>
41	Contenu issu d'une autre source sous licence GFDL	GFDL	<p>Une partie de cet article est issue de l'ancien article « Sélections » sous licence GFDL qui a été transformé en simple redirection. Consultez l'historique de la page originale pour connaître la liste de ses auteurs.</p> <p>OU</p> <p>Le contenu de cet article est issu en tout ou partie du travail de Alain Legrand. La permission de distribuer ce travail sous la GFDL (et donc par extension sous la licence Creative Commons paternité partage à l'identique) a été reçue sur OTRS, via le ticket no 2007022210011921.</p> <p>OU</p> <p>Tout ou partie de cet article est issu de l'ancien article « Homoncule (œuvre de fiction) » sous GFDL qui a été transformé en simple redirection.</p>
#24 436	Article rejeté en BA	Rejet BA	
#25 449	Article déchu en BA	Déchu BA	
45	Confirmation de suppression après vote → sur page thématique suppression	Suppression	Article supprimé par GL (d) 1 décembre 2009 à 02:18 (CET) Consensus
46	Confirmation de conservation après vote → sur page thématique suppression	Conservation	Article conservé par GL (d) 3 juillet 2009 à 22:31 (CEST) Pas de consensus pour la suppression.

#54 816	Protection	Protection	<p>Cette page est protégée.</p> <p>Cette protection ne constitue pas obligatoirement une approbation de la version actuelle. Vous pouvez proposer une modification en page de discussion. Lorsqu'un compromis sera trouvé en page de discussion, vous pourrez demander la modification auprès des administrateurs. · Journal des protections.</p>
------------	------------	------------	---

*Tableau 1: Synthèse 3 des balises présentes uniquement au sein des pages de discussion parallèles*