Wikiconflits, un corpus extrait de Wikipédia : principe et méthode d'élaboration

Céline Poudat, Jin Kun, Thierry Chanier

Wikiconflits, un corpus extrait de Wikipédia : principe et méthode d'élaboration

Version 4

- Version 4, 1er septembre, Kun JIN et TC, ajout de section 3.1, 3.7 et le lien de téléchargement de wikiexporttool
- Version 3, 13 août, Kun JIN et TC, ajout du 3.5 sur contributeurs et de la section 8 pour discussion interne
- Version 2, 25 juin, Kun et TC
- Version 1, 29 avril 2014, assemblé par TC

Comment citer ce document:

Poudat, C, Jin, K., & Chanier, T. (2014). Manuel du corpus *wikiconflits (cmr-wikiconflits-tei-v1-manuel.pdf)* .In XX., Corpus Wikiconflits extraits de Wikipedia. Dans banque de corpus CoMeRe.org . Ortolang.fr : Nancy. [cmr-wikiconflits-tei-v1]

➢ Objectif de ce document

Le groupe *CoMeRe-nouvelles-acquisitions-Wikipédia* vise à constituer un corpus de pages Wikipédia qui sera adjoint au corpus CoMeRe afin d'améliorer sa représentativité. Son choix s'est porté autour de l'observation de pages ayant suscité des discussions conflictuelles autour de controverses dans les champs des sciences et des techniques. Dans une première partie, le groupe explique sa méthode de sélection des discussions conflictuelles. Dans les parties suivantes, le groupe *CoMeRe-LRL* expose la méthode de constitution du corpus en fonction des critères précédemment retenus et la façon de les transformer en TEI. Ce rapport se termine par la présentation de l'outil *WikiTool* développé lors des traitements. Mis à disposition sous licence libre il permet d'extraire des pages dans les fichiers de type Dump de Wikipedia.

Contenu

1. Partie 1 : conflits dans le Wikipédia francophone

Afin d'améliorer la représentativité du corpus CoMeRe en y adjoignant un corpus de pages Wikipédia, un groupe de travail restreint *Wikipédia - nouvelles collectes* a été constitué, avec N. Grabar (STL, Lille 3), C. Paloque-Berges (HT2S et DICEN, CNAM) et C. Poudat (Lattice, UP13). Après différents échanges autour de nos intérêts de recherche propres, nous avons décidé d'orienter la collecte autour des **pages ayant suscité des discussions conflictuelles** autour de controverses dans les champs des **sciences et des techniques**.

La procédure de sélection des pages a été réalisée en trois temps: nous nous sommes d'abord concentrées sur les lieux de Wikipédia les plus susceptibles de contenir des séquences conflictuelles et 1002 articles ont été présélectionnés pour examen et évaluation (Section 1). Au final, seuls neuf articles et thèmes conflictuels ont été retenus pour leur représentativité des conflits sévissant dans le domaine des sciences et techniques (Section 2). Le nombre peut paraître limité, mais il faut souligner que l'examen des pages et des conflits a été assorti d'une réflexion sur la nature et la structuration des données à retenir afin d'être en mesure de suivre un conflit, de sa genèse à sa montée et sa résolution le cas échéant. Nous avons ainsi choisi de de regrouper les pages potentiellement concernées par un conflit en *clusters de pages* autour d'un article (Section 3), ce qui a entraîné une réduction drastique du nombre d'articles sélectionnés étant donné la masse de données à extraire pour chaque cluster. Nous présentons également, dans cette dernière section, un ensemble de recommandations et de spécifications pour l'extraction des pages.

1.1. De la démographie du conflit dans Wikipédia à la présélection des pages

1.1.1. Les acteurs du conflit

Bien que les contributeurs de Wikipédia forment une constellation hétérogène d'individus, deux grandes catégories de participants peuvent être mises au jour dans la genèse et la résolution d'une dispute : les médiateurs, ou arbitres de la dispute (Figure 1, en vert), et ses impliqués potentiels (en bleu).

Si l'on s'en tient aux grands principes de Wikipédia, les conflits sont censés se résoudre par la discussion. Lorsque cette dernière échoue et que la dispute dégénère, deux solutions sont envisageables : la *médiation* ou le *procès*.

L'espace médiation (Wiki-médiation, 2014) propose ainsi un ensemble d'outils et de principes de communication et de résolution de conflits tandis que le *salon de médiation* (Wiki-salon, 2014) est le lieu dédié à la résolution de conflits du projet. Un contributeur dépose une demande de médiation en décrivant le conflit qui l'oppose à un (ou un ensemble de) contributeur(s) qui a (ont) naturellement droit de réponse. Un médiateur volontaire, généralement aguerri et fin connaisseur des grands principes du projet, répond et tente de trouver un compromis acceptable par les protagonistes du conflit. Par ailleurs, des médiateurs spécifiques peuvent être saisis: on peut ainsi demander à un utilisateur expérimenté d'être le *médiateur* d'un conflit, en lui remettant dans ce cas le *barnstar de médiation* sur sa page.

A noter qu'une liste de médiateurs expérimentés est disponible (Wiki-mediateurs, 2014), bien que celle-ci ne recense qu'un nombre très restreint de médiateurs (quatre au total).

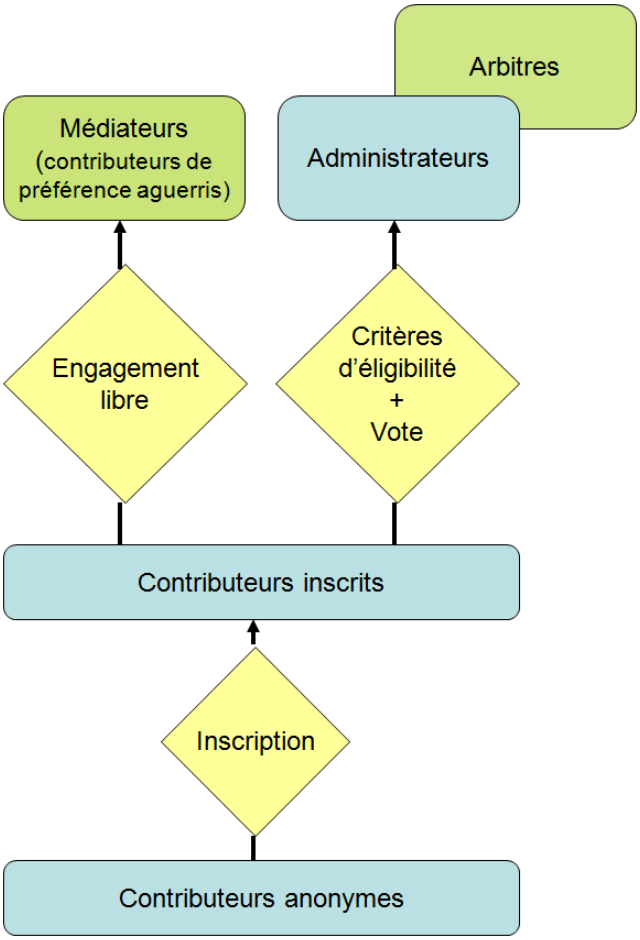

Figure 1.1: Acteurs du conflit dans Wikipédia

Plus formel, le *comité d'arbitrage* (CA), composé depuis la réforme de 2013 de cinq arbitres élus, est la plus haute instance de l'ordre judiciaire wikipédien. Il juge en dernier recours les affaires qui n'ont pu être réglées à l'amiable. La tenue d'un comité d'arbitrage est un événement de caractère exceptionnel. Depuis sa création en 2005, seule une centaine de contributeurs s'est vue impliquée dans un conflit nécessitant arbitrage plutôt que médiation.

Bien que tout contributeur puisse être impliqué dans une dispute, il faut également distinguer trois grandes catégories d'acteurs représentant trois niveaux d'engagement dans le projet encyclopédique (Figure 1.1, en bleu): les contributeurs *anonymes*, qui ne possèdent pas de compte Wikipédia, et qui ne sont identifiables que par leur adresse IP ; les contributeurs *inscrits*, qui recouvrent la population participante identifiée par un pseudonyme ; et enfin, les administrateurs, contributeurs aguerris et motivés, qui sont élus par la communauté des wikipédiens. A noter que les anonymes représentaient plus de 90% des contributeurs en avril 2006. Débutants ou participants épisodiques, les IP sont fortement associés aux vandales dans l'imaginaire collectif wikipédien, et leurs interventions sont étroitement surveillées (e.g. *Patrouille Recent Changes* – (Wiki-Prc 2014)), voire plus controversées : ainsi, la moitié des caractères insérés par les IP était supprimée en 2006. Les anonymes identifiés comme nuisances au projet sont rapidement bloqués par les

administrateurs, qui détiennent ce droit. Si les disputes impliquant des IP sont en général plus ponctuelles, c'est en partie parce que l'anonyme, en tant que « sans-papier » dans le projet, peut difficilement engager de réelles mesures juridiques par exemple : ainsi, les cinq anonymes que l'on recense dans les archives du comité d'arbitrage que nous avions observées en 2007 ont donné lieu à des plaintes non recevables (utilisateur Crodan contre X, et utilisateur 82.224.88.52 contre le « reste de la communauté »).

1.1.2. Les lieux du conflit

La dispute prend forme et se cristallise dans trois espaces principaux : (i) l'*article*, dans le cadre de désaccords, voire de guerres d'édition, qui peuvent se régler ou dégénérer par (ii) la *discussion*, voire donner lieu à (iii) un *arbitrage* devant le CA du projet.

(i) Dans le cadre de l'article, elle prend la forme d'actions d'édition antagonistes, dont l'expression la plus manifeste est la guerre de retours aux versions précédentes (ou révocations, ou `reverts`). 5% des éditions effectuées sur la base extraite en 2006 étaient ainsi des révocations, et 3% des articles avaient fait l'objet d'au moins trois reverts, ce qui peut être le signe d'un conflit. Un article ayant un taux important de révocations peut ainsi contenir une guerre d'édition. Par exemple, l'article *homéopathie* avait à l'époque un taux de 10% de révocations effectuées par 18 contributeurs distincts. Lorsque la guerre d'édition est manifeste et qu'elle ne peut être endiguée, un bandeau[1] signalant la présence d'un problème éditorial sur le texte peut être posé. Plusieurs types de bandeaux signalent la présence potentielle de conflits dans les textes, de l'admissibilité de l'article au désaccord avéré de pertinence ou de neutralité, en passant par l'article "incompréhensible" ou à "désacadémiser". Par ailleurs, certains articles peuvent être protégés (Wiki-blocecrit, 2014): l'article est alors temporairement, voire définitivement bloqué ou restreint en écriture. Plusieurs articles sont ainsi bloqués, du fait des thèmes-mêmes qu'ils abordent, religieux (*e.g.* Mahomet), politico-idéologique (*e.g.* Marine Le Pen) ou encore propices au vandalisme potache (articles sexuels, scatophiles).

(ii) Le désaccord peut également se manifester sous la forme discursive de la discussion, qui permet dans le meilleur des cas de coordonner les contributeurs et de rechercher un consensus. On recense de très nombreux lieux de discussion dans le projet : la moitié des vingt catégories de pages (ou `namespaces`) que l'on relève sont ainsi des pages de discussion (*e.g.* `discussion :portail, discussion:image`, *etc.*), et les plus remarquables sont associées aux articles (*discuter :article*) et aux utilisateurs (*discussion :utilisateur*). A noter que la pose d'un bandeau sur un article entraîne la création d'une page de discussion additionnelle, ce qui complexifie plus encore l'observation d'un conflit puisque les discussions peuvent advenir tant sur la page de discussion de l'article que sur celle du bandeau ou des utilisateurs impliqués.

[1] 94 modèles de bandeaux sont disponibles pour catégoriser les articles dans le projet, liste des catégories disponible sur (Wiki-bandeaux, 2014)

(iii) Enfin, le comité d'arbitrage (CA), haute instance de la justice wikipédienne, se charge enfin de trancher ces querelles et conflits d'édition entre contributeurs de l'encyclopédie. Composé de wikipédiens élus par la communauté et régulièrement renouvelés, le CA étudie et évalue la recevabilité des plaintes déposées par les contributeurs entrés en conflit. Si la plainte est déclarée recevable, une proposition de règlement du conflit est soumise au vote des arbitres et un jugement est rendu, qui peut décider du bannissement d'un contributeur.

1.1.3. Les lieux retenus en présélection

Une première présélection des lieux les plus susceptibles d'abriter un conflit a été effectuée. Un ensemble de pages a ainsi été arrêté en date d'octobre 2013 pour évaluation et intégration dans le jeu de données final.

Quatre ensembles de pages ont été présélectionnés: (i) les 73 pages faisant l'objet d'une procédure de médiation ont d'abord été retenues. Nous nous sommes ensuite limitées aux deux seuls bandeaux d'articles contenant le terme "désaccord", à savoir les pages faisant l'objet (ii) d'un **désaccord de neutralité** (214 pages) et (iii) d'un **désaccord de pertinence** (546 pages). Enfin, (iv) les 169 articles **protégés** ou **semi-protégés**[2] du projet ont également été évalués. Au final, 1002 pages ont été examinées [3]

1.2. Des conflits les plus représentatifs des domaines des sciences et techniques

Si nous avons veillé à sélectionner des articles de taille suffisante, ayant suscité discussion (nombre de topics suffisant, profondeur des fils), nous avons fait le choix de cibler un champ et des thématiques spécifiques correspondant à nos intérêts de recherche, afin notamment d'être en mesure d'avoir une certaine expertise ou une compréhension a minima des points de désaccord.

En ce sens, nous avons restreint notre sélection aux domaines des sciences et techniques, en privilégiant des sujets d'article controversés dans les champs académiques et d'ingénierie, ou qui, s'ils ont fait l'objet d'un débat public, ont suscité le recours important à des connaissances relatives à ces domaines. Notre sélection n'a pas seulement pour critère le sujet des articles, mais aussi le type de discussion engagée dans les coulisses de l'article. Dans le cadre d'un sujet controversé, la limite entre débat d'ordre scientifique et technique et débat d'ordre politique et idéologique est évidemment très poreuse (la controverse telle que définie par les *Science and Technology studies* se fonde d'ailleurs notamment sur cette incertitude). C'est pourquoi nous avons, parmi la première sélection des articles à conflit, ciblé ceux dont les discussions illustrent un recours explicite des participants à des types de raisonnement relatifs aux langages scientifiques et techniques et au discours de la preuve (débat sur une théorie, un concept, une méthode, une démonstration, une discipline…) -

[2] Pour ces deux notions, voir (Wiki-protect, 2014) et (Wiki-semiprotect, 2014)

[3] - tableur disponible ici https://docs.google.com/spreadsheet/ccc?key=0AnyqPxDqTJA7dE14UXhaUTlHZ1ZnWEdidngyMmJjVFE#gid=3 NON accessible

quelles que soient la qualité du raisonnement ou des arguments avancés, ou les prises de position politiques ou dérives idéologiques flagrantes les nourrissant.

Ainsi, les articles positionnés dans des thématiques éloignées des sciences et techniques ont été écartés d'emblée (e.g. personnalités politiques, entreprises, sport, religion).

L'examen des pages nous a permis d'observer différents types de conflits

- Pseudo-sciences
- personnalités controversées
- technosciences
- controverses publiques
- méthodologies
- légitimité, scientificité d'une discipline

Au final, huit pages et un thème conflictuel (dorénavant *clique*) ont été retenus(voir Tableau 1.1).

1.3. Articles, clusters et cliques de pages

Sur la base de notre connaissance du projet encyclopédique et après observation de quelques centaines de pages, il s'avère qu'un conflit en cours peut se prolonger sur plusieurs pages, i.e. démarrer dans une guerre de revert dans l'édition d'un article, puis dans la page discussion de l'article ou dans la page de l'un des utilisateurs concernés, voire donner lieu à un bandeau apposé sur l'article qui générera également une page de discussion spécifique.

Ainsi, parmi notre corpus, les pages de discussion des articles "Chiropratique", "Igor et Grichka Bogdanoff", "OGM", "Homoparentalité", "Psychanalyse" (qui sont aussi les sujets les plus controversés de notre corpus, à la fois médiatiquement et sur Wikipédia) présentent une structure complexe de plusieurs pages où se distribuent certains pans de la discussion. On prendra l'exemple de l'article sur les frères Bogdanoff, qui à défaut d'être représentatif en termes d'acteurs de la discussion (le conflit est nourri par l'intervention des personnalités - ou de leur relation de presse prenant leur nom pour intervenir sur WP) montre clairement ces jeux structurels d'extension du conflit. En effet, le conflit qui a lieu sur l'onglet discussion de l'article (http://fr.wikipedia.org/wiki/Discussion:Igor_et_Grichka_Bogdanoff) se poursuit d'abord dans deux des catégories listées par un bandeau "Autres discussions" en haut de cette page :

- rétrospectivement, dans les archives (http://fr.wikipedia.org/wiki/Discussion:Igor_et_Grichka_Bogdanoff/Archives) fréquent dans le cas de discussions pléthoriques comme celles évoquées plus haut ;
- sur une page "Neutralité" (http://fr.wikipedia.org/wiki/Discussion:Igor_et_Grichka_Bogdanoff/Neutralit%C3%A9)

Page	Lieu	Thème	Lien Wikipédia online	Commentaire
Chiropratique	NPOV	Pseudo- sciences	fr.wikipedia.org/wiki/Chiropratiqu e	Page de discussion avec archive et extension dans page de neutralité. Gros conflit sur l'interprétation politique et/ou scientifique d'un rapport critique de l'académie des sciences au sujet de la chiropratique, et plus généralement sur le choix sélectif des sources.
Igor et Grichka Bogdanoff	NPOV	Personnalités controversées	fr.wikipedia.org/wiki/Igor_et_Gric hka_Bogdanoff	Page de discussion avec archive et extension dans page de neutralité. Grosse intervention des frères Bogdanoff sur affaire controversée de leurs thèses (maths et physique) et procès CNRS. Confrontation rapports administratifs et scientifiques sur ces thèses.
OGM	Médiation	Technosciences , controverse publique	fr.wikipedia.org/wiki/Organism e_ génétiquement_modifié	Page de discussion comporte liens vers page "neutralité", "article de qualité", lumière sur", "à faire". Page pléthorique dont les derniers conflits concernent l'affaire Séralini.
Quotient intellectuel	Blocage	Méthodologies	fr.wikipedia.org/wiki/Quotient_int ellectuel	Conflits sur les modèles de calcul, mesure et critères du QI, discussions d'ordre épistémologiques sur la pertinence des modèles par rapport à des explications de type sociologique (débats qui débordent sur le sujet connexe de l'intelligence (et article associé)). Conflit sur rapports entre Qi et phrénologie/volume crânien. Question de la justesse des modèles statistiques et querelles sur lois mathématiques. Gros conflit sur théories génétiques/racistes controversées et verrouillage de l'article contre plusieurs défenseurs de ces théories (épisode de la « grosse vache stalinienne » en particulier, mais dispersé dans toute la page discussion), et conflit sur les méthodes "anti-scientifiques" et "abusives" des administrateurs (ça dégénère). Contient des "commentaires des lecteurs" (page spéciale)
Histoire de la logique	Pertinence	Histoire et épistémologie	fr.wikipedia.org/wiki/Histoire_de _la_logique	conflits sur l'interprétation de concepts (théorème complétude...) et sur les manques de cette histoire
Psychanalyse	Médiation	Légitimité, scientificité, Méthodologies	fr.wikipedia.org/wiki/Psychanalyse	Page de discussion comporte liens vers page "neutralité", "à faire", "archives". Page pléthorique, mais dont le conflit central concerne les pro-Livre noir de la psychanalyse contre les anti. Donc conflit sur sources et arguments, interprétation idéologique et anti-scientifique.
Eolienne	Pertinence	Technosciences	wikipedia fr..org/wiki/Éolienne	d'autorité sur discours : revendication posture d'expert, justification récurrente d'être ou non spécialiste (et quelle spécialité... "pour moi un mécanicien...", "ingénieur aérodynamicien" par ex), beaucoup de registre didactique (donnent des leçons de calcul de la puissance énergétique). 1er conflit sur meilleure description (forme du langage) pour explication unités de calcul ; 2ème conflit : question réchauffement climatique. Questions d'ingénierie (techniques alternatives énergétiques) et d'écologie (rapport au climat). Interaction se fait beaucoup sous la forme de l'échange complémentaire (essayer de comprendre ensemble, d'améliorer ensemble, félicitations réciproques), un peu moins sur mode de conflit (mais quand même présent, sous une forme relativement policée et bien argumentée). Tentative d'exhaustivité des facteurs à prendre en compte

Tableau 1.1 : Liste des thèmes retenus

Mais elle s'étend également au-delà de l'espace éditorial propre au dispositif, déjà complexe, de l'onglet discussions lié directement à l'article :

sur la page de discussion de l'utilisateur "Igor et Grichka Bogdanoff" (homonyme, donc, du titre de la page leur étant consacrée), par le biais du lien "(d)" (pour "discussion")

accolé à son nom d'utilisateur (http://fr.wikipedia.org/wiki/Discussion_utilisateur:Igor_%26_Grichka_Bogdanoff);

mais aussi, bien sûr, sur la page d'autres utilisateurs ayant pris part au conflit, en particulier ceux qui y participent le plus activement (en créant du conflit ou au contraire en le modérant) ; c'est le cas par exemple de l'utilisateur "Noisetier", qui se place du côté des arbitres de médiation, et sur la page de discussion duquel (via le lien "d" pré-cité) on retrouve encore une structure d'archives dans laquelle on peut retrouver des débats sur le sujet (http://fr.wikipedia.org/wiki/Discussion_utilisateur:Noisetier/archive_mars-mai_2011, archive identifiée après avoir constaté que Noisetier a pris part au débat sur l'onglet discussion de l'article "Igor et Grichka Bogdanoff" en mars et avril 2011).

Pour ces raisons, nous avons choisi d'extraire l'ensemble des pages potentiellement susceptibles d'abriter un conflit autour d'une page - que nous appellerons désormais clusters de pages (1.3.1.). A noter que pour des raisons de représentativité qualitative de notre collecte, nous souhaitons collecter les frères et les filles de la page Mariage homosexuel, le conflit frappant cette page renvoyant en effet à la controverse publique ayant frappé - et frappant encore la France - ces derniers mois, autour du Mariage pour tous et de l'ouverture de la PMA et de l'adoption aux couples homosexuels. La sélection d'une page seule nous semblant insuffisante pour appréhender de manière pertinente le conflit dans sa globalité, nous envisageons d'extraire des clusters de pages en lien, que nous appellerons désormais clique de pages (1.3.2.).

Nous avons réfléchi à quelques pistes pour l'extraction des pages via les dumps Wikipédia (1.3.3.).

1.3.1. Clusters de pages

Pour chaque conflit sélectionné, quatre types de pages seront donc extraites:

- la page *article* et son historique wiki avec les diffs;
- la page *discussion* de l'article et son historique wiki avec les diffs;
- les pages *discussion* de l'ensemble des utilisateurs ayant contribué à l'article;
- les éventuelles autres pages *discussion* autour de l'article, accessible depuis un lien dans l'encadré *Autres discussions* de chaque page de discussion. A titre d'exemple, la page discussion de l'article **Homoparentalité** (Figure 1.1) renvoie également à trois autres pages de discussion: ***neutralité***, qui rassemble les discussions autour de la NPOV; ***A faire***, qui n'est pas une page de discussion à proprement parler, mais une todolist; et enfin, les ***Archives***, les discussions étant progressivement archivées.

Figure 1.2 : page de discussion de l'article homoparentalité

➢ Structure XML possible

Chaque article doit être contenu dans un même noeud XML lui-même divisé en trois noeuds ou sous-ensembles:

la page et ses diff dans l'ordre chronologique + les commentaires annotant chaque diff dans l'onglet **Afficher l'historique** (en ligne);

la discussion et ses diff dans l'ordre chronologique + les commentaires annotant chaque diff dans l'onglet **Afficher l'historique** (en ligne);

et un ensemble de pages additionnelles

- Les pages liées à la page discussion dans la rubrique **Autres discussions**, à l'exception de **A faire**
  - la page et ses diff dans l'ordre chronologique + les commentaires annotant chaque diff dans l'onglet **Afficher l'historique** (en ligne)
- Les pages **Utilisateurs** liées
  - Il faut donc récupérer les pages de tous les utilisateurs ayant participé aux pages susnommées; aspirer leur page personnelle (page utilisateur en l'état / discussion + historique avec diff)

1.3.2. Dumps Wikipedia

Le plus simple sera certainement de télécharger les dumps et de filtrer ensuite les pages qui nous intéressent.

- Une aide sur le Wikipédia hors connexion est disponible ici: http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Wikip%C3%A9dia_hors-connexion
- page dumps ici http://dumps.wikimedia.org/frwiki/latest/

Pour chaque projet, les fichiers suivants sont disponibles :

- pages-articles.xml.bz2 - révisions courantes, sans les autres espaces de noms . Inclut les modèles, exclut les pages de discussion et les pages utilisateur.
- pages-meta-current.xml.bz2 - révisions courantes, toutes les pages.
- pages-meta-history.xml.bz2 - toutes les révisions, toutes les pages.
- abstract.xml.gz - pages résumées.
- all_titles_in_ns0.gz - les titres des articles uniquement.
- des fichiers SQL pour les interwikis, les catégories et les modèles sont également disponibles.

2. Partie 2 : méthode d'extraction et sélection des thèmes choisis

2.1. License de Wikipédia : CC-BY-SA 3.0

Wikipédia est conçue pour être réutilisée et diffusée sous les règles de « Licence Creative Attribution - Partage dans les Mêmes Conditions 3.0 non transposé » (CC_BY_3.0, 2014) ; c'est en cela et seulement en cela qu'elle est une encyclopédie « libre ». Pour la réutilisation, la copie ou la modification de tout ou partie du texte d'un article, Wikipédia a demandé que (Wikipédia:Citation_et_réutilisation_du_contenu_de_Wikipédia, 2014)

- Indiquer que le contenu réutilisé, copié ou modifié est sous CC BY-SA 3.0,
- permettre l'identification des auteurs en donnant une adresse web vers l'article de Wikipédia, ou en donnant une liste des auteurs (paternité),
- indiquer si vous avez modifié le contenu original de Wikipédia,
- laisser tous les travaux dérivés sous la même licence (partage à l'identique).

2.2. Méthodes d'extraction de données recommandées par Wikipédia

Wikipédia demande ne pas crawler directement un grand nombre de page de la ligne étant donné que ceci va ralentir le serveur, sinon le serveur de Wikipédia va bloquer l'adresse IP de visiteur (Wikipedia:Database_download, 2014). Autrement, Wikipédia nous ouvert la porte pour exporter des données par différentes méthodes :

- **Exporter en ligne**, une méthode est plus pratique, elle nous permet d'extraire un petit nombre de pages et leurs versions historiques, mais le nombre de versions historique est limité à 1000.

  URL: http://fr.wikipedia.org/wiki/Special:Export

- **Dumps Wikimedia**, des fichiers en format XML contiennent le texte ou la métadonnée de toutes les pages actuelles ou leurs révisions historiques de Wikipédia. (Section : 2.3).

  URL: http://dumps.wikimedia.org/
  URL Français: http://dumps.wikimedia.org/frwiki/

2.3. Organisation des fichiers Dumps de Wikipédia

2.3.1. Identifiant des dumps

WMF (Wiki Media Fondation) publie régulièrement en dump les données de Wikipédia et de tous les projets de WMF. Un dump en format XML contient le texte ou la métadonnée de toutes les pages actuelles ou leurs révisions historiques. Pour identifier ces dumps, Wikipédia les a donné des identifiants : (Data_dumps/Dump_format, 2014):

- pages-articles.xml
  - Contient les versions de tous les articles, modèles et autres pages
  - Exclure des pages de discussion et des pages principaux d'utilisateur

- pages-meta-current.xml
  - Contient la version actuelle de tous les pages, compris les pages de discussions et pages principaux d'utilisateur.
- pages-meta-history.xml
  - Contient toutes les versions de tous les pages (le fichier est hyper gros). Ce sont ces dumps de la **Erreur ! Source du renvoi introuvable.**qu'on utilise dans le projet CoMeRe, ces fichiers dont les tailles décompressées sont entre 300Go – 500Go, une expérience obtenue après avoir comparé les fichiers décompressés entre ces deux différents types de compression : 7z et bz2, ils sont tous les mêmes.

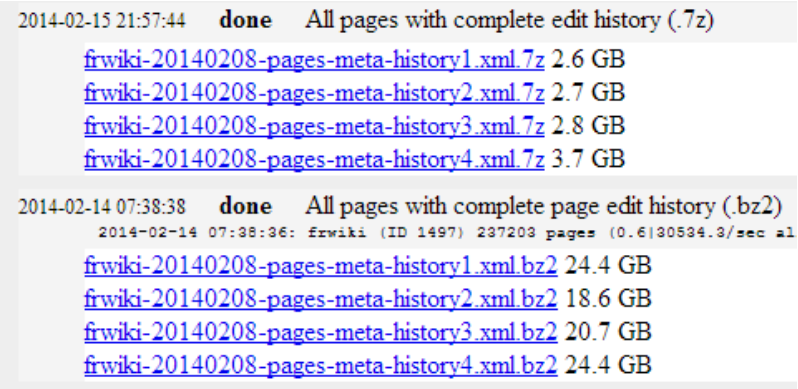

Figure 2-1 : Dump de Wikipédia et leur taille en Go

2.3.2. Structure des Dump

Quant à la structure de Wikipédia, un schéma (Wiki-xsd, 2014) et une représentation graphique (Figure 2-) sont présentés en ligne.

Ce schéma XSD illustre une structure complète de Wikipédia, y compris celle du celle du dump. Mais cette dernière est plus simple que celle de Wikipédia, par conséquent avec un éditeur XML « Oxygen » (Site_officiel_d'Oxygen, 2014) et un exemplaire de dump, nous avons créé un schéma Relax NG (RELAX_NG_Home, 2014) pour présenter la structure de dump (cf. Figure 2-2).

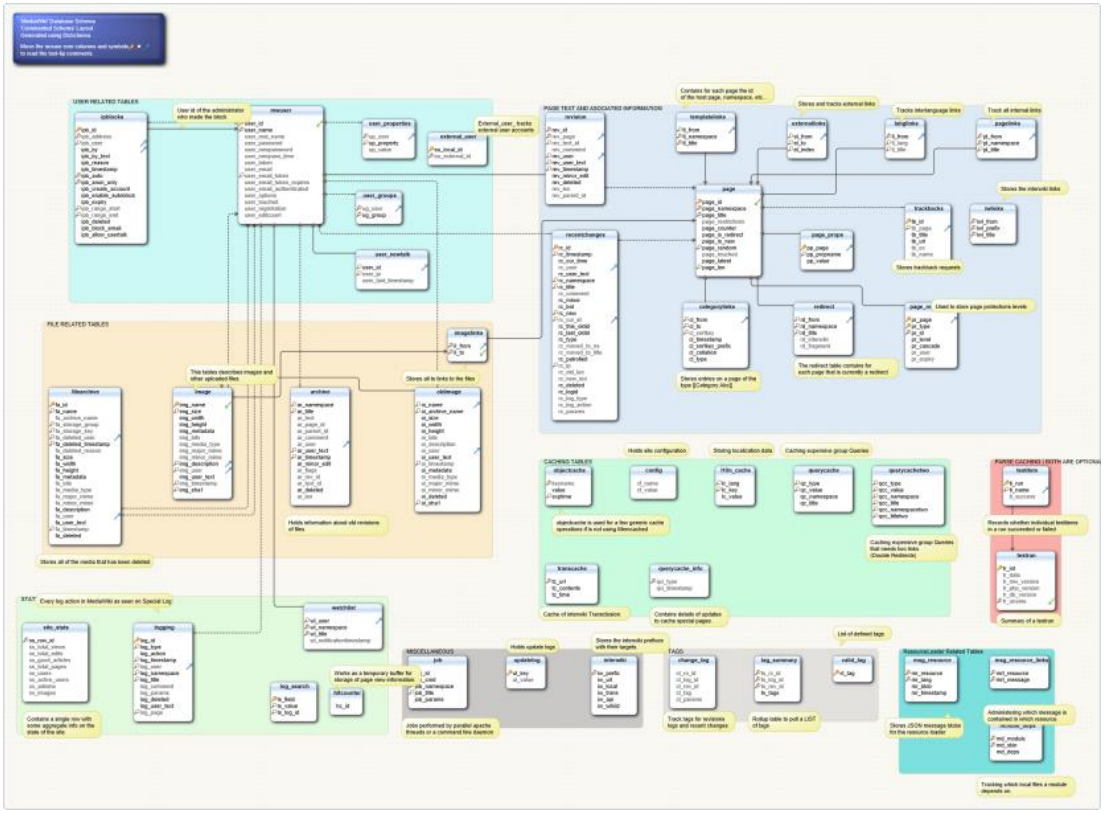

Figure 2-2 Schéma de l'organisation de Wikipédia, source : (Wiki-schéma, 2014)

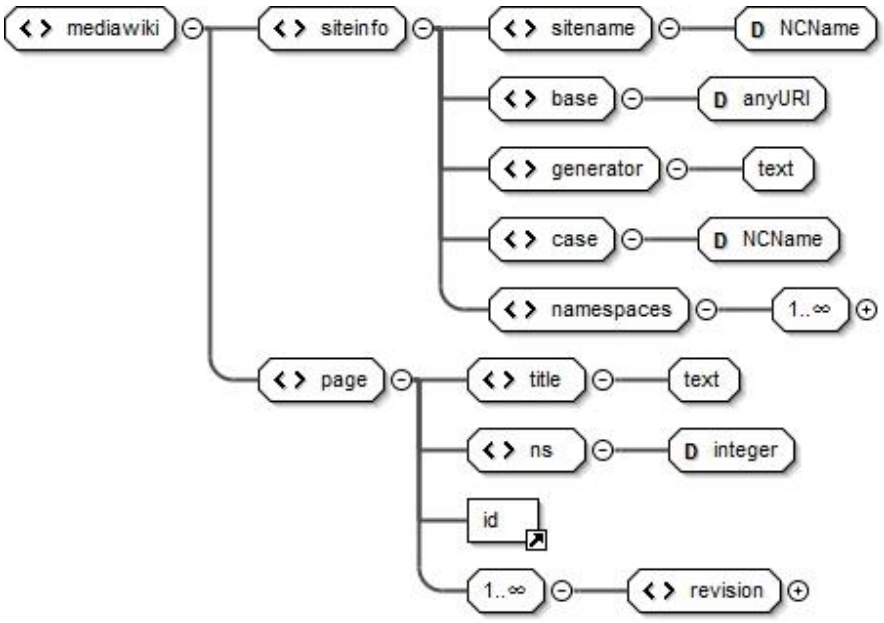

Figure 2-2, schéma XML des fichiers de type Dump

- **<mediawiki>** est la racine de fichier XML.
  - **<siteinfo>** contient des informations sur le site global, comme *Wikipédia*.
    - **< sitename>**, nom du site
    - **<base>**, URL de site

- **<générateurs>**, outil qui crée ce dump
- **<case>**, la valeur de cet élément 'first-letter' | 'case-sensitive' | 'case-insensitive'
- **<namespaces>** donne une indication pour distinguer le type de page (Section 2.3.3.).

■ **<page>** contient des informations de page et différentes versions historiques (Section 2.3.4.).

2.3.3. Nom d'espace

Nom d'espace (`Namespaces`) présente l'identifiant de type d'article, comme Wikipédia est un projet multi-langue, il est plus facile d'unifier tous les types d'article en chiffre. Nom d'espace dans le dump illustre comme <namespace> qui est un élément de fils de <siteinfo>. L'exemple (2.1) ci-dessous est une liste de noms d'espace de Wikipédia français. Cette liste est faite sous des conditions (Help:Namespace, 2012) :

- chaque espace de contenu s'associe à un espace de discussion ;
- chaque type d'espace s'associe à un identifiant qui est la valeur de l'attribut @key dans l'élément <namespace> ;
- le texte de <namespace> est une préfix de titre de page (Section 2.3.4) ;
- à partir des noms d'espaces principaux, chaque projet Wikipédia peut définir leurs propres noms d'espaces, c'est-à-dire qu'un espace n'existe pas forcément dans tous les projets de Wikipédia, il est donc numéroté à partir de 100.

```
(2.1)
<mediawiki xmlns="http://www.mediawiki.org/xml/export-0.8/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.mediawiki.org/xml/export-0.8/
http://www.mediawiki.org/xml/export-0.8.xsd" version="0.8" xml:lang="fr">
  <siteinfo>
 <base>http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil principal</base>
    <generator>MediaWiki 1.23wmf12</generator>
    <case>first-letter</case>
    <namespaces>
      <namespace key="-2" case="first-letter">Média</namespace>
      <namespace key="-1" case="first-letter">Spécial</namespace>
      <namespace key="0" case="first-letter" />
      <namespace key="1" case="first-letter">Discussion</namespace>
      <namespace key="2" case="first-letter">Utilisateur</namespace>
      <namespace key="3" case="first-letter">Discussion utilisateur</namespace>
      <namespace key="4" case="first-letter">Wikipédia</namespace>
      <namespace key="5" case="first-letter">Discussion Wikipédia</namespace>
      <namespace key="6" case="first-letter">Fichier</namespace>
      <namespace key="7" case="first-letter">Discussion fichier</namespace>
      <namespace key="8" case="first-letter">MediaWiki</namespace>
      <namespace key="9" case="first-letter">Discussion MediaWiki</namespace>
      <namespace key="10" case="first-letter">Modèle</namespace>
      <namespace key="11" case="first-letter">Discussion modèle</namespace>
      <namespace key="12" case="first-letter">Aide</namespace>
      <namespace key="13" case="first-letter">Discussion aide</namespace>
      <namespace key="14" case="first-letter">Catégorie</namespace>
      <namespace key="15" case="first-letter">Discussion catégorie</namespace>
      <namespace key="100" case="first-letter">Portail</namespace>
      <namespace key="101" case="first-letter">Discussion Portail</namespace>
      <namespace key="102" case="first-letter">Projet</namespace>
      <namespace key="103" case="first-letter">Discussion Projet</namespace>
      <namespace key="104" case="first-letter">Référence</namespace>
      <namespace key="105" case="first-letter">Discussion Référence</namespace>
      <namespace key="828" case="first-letter">Module</namespace>
      <namespace key="829" case="first-letter">Discussion module</namespace>
    </namespaces>
  </siteinfo>
  <!-- partie de page -->
</mediawiki>
```

2.3.4. Page et page historique

Une **page de Wikipédia** est composée d'un titre, des sujets et des textes correspondants. Comme la section précédente indiquée, Wikipédia contient plusieurs types de pages: page d'article, d'auteur ou de discussion etc. Le titre de chaque page correspond au type de page. Par exemple pour un sujet « Chiropratique » et le participant sur ce sujet « BonifaceFR », souvent à chaque page est associée une page de discussion :

- « Chiropratique », correspond au nom d'espace « 0 », pas de texte trouvé dans l'élément <namespace>, pas de préfix;
- « Discussion : Chiropratique », correspond au nom d'espace « 1 », donc le préfixe de titre est le texte de <namespace> est « Discussion » connecté avec le sujet par « : »;
- « Utilisateur : BonifaceFR », correspond au nom d'espace « 3 », le préfixe de titre est le texte de <namespace> est « Utilisateur » ;
- « Discussion utilisateur : BonifaceFR », correspond au nom d'espace « 4 », le préfixe de titre est le texte de <namespace> est « Discussion utilisateur».

Wikipédia fournit une **page historique** (Help:Page_history, 2014) qui affiche le journal de modification depuis la création de cet article. On y accède en cliquant « Afficher l'historique » en haut de la page de l'article. Cette page contient une liste de révisions précédentes avec leurs horaires de modification, nom ou adresse IP d'auteur et résumé d'édition (cf.Figure 2-3).

Figure 2-3 : aperçu d'un historique des modifications d'un article

Dans le dump de Wikipédia, chaque article, ainsi que le contenu entier de chacune des révisions est inclus dans l'élément <page>. Dans chaque révision, se trouve le contenu entier de l'article (et non simplement les ajouts / suppressions) après modification par l'éditeur de la révision. Le schéma de <page> est donné en Figure 2-4. Il faut indiquer que ce schéma contient trois identifiants, chacun étant unique dans tout Wikipédia : un pour la page, un second pour la révision et un autre pour le contributeur / éditeur

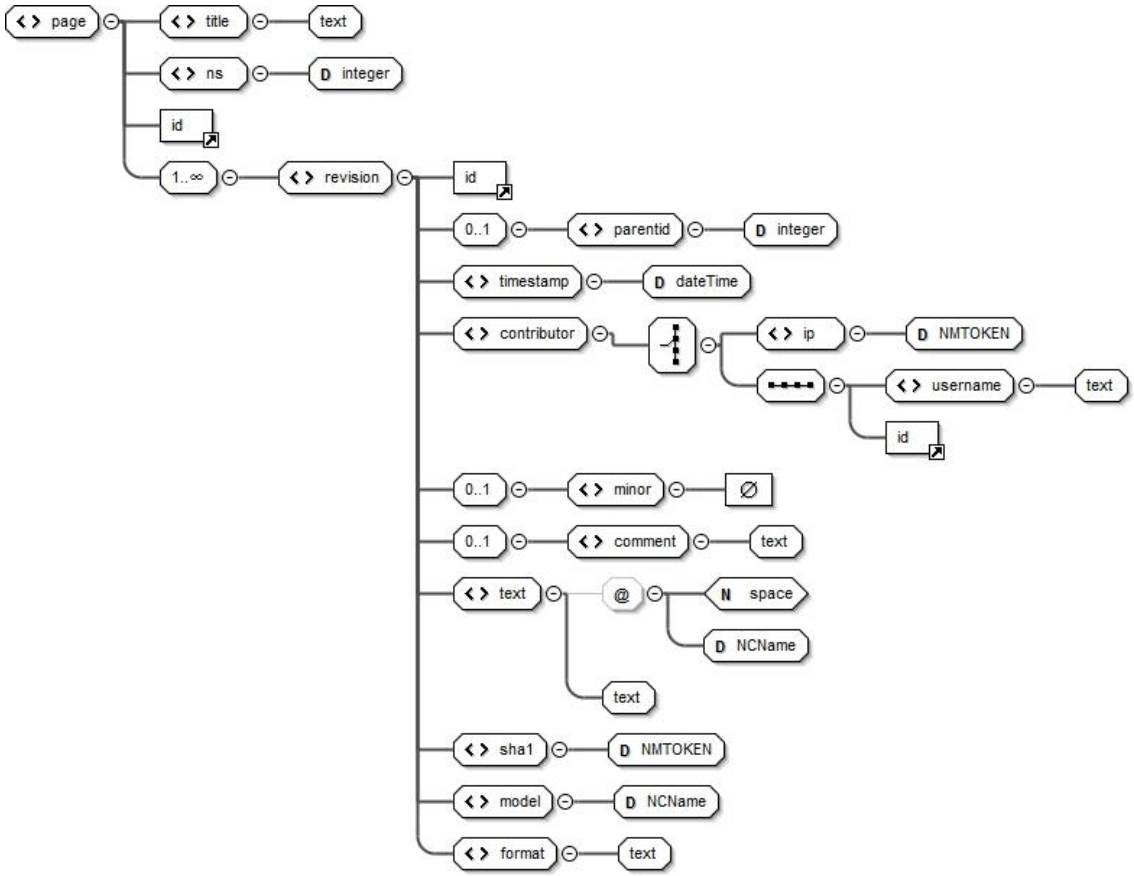

Figure 2-4 : schéma de l'élément <page>

- **<page>**
  - **<title>**, nom du page
  - **<ns>** ayant une valeur se correspond à la valeur de l'attribut @key de l'élément <namespace> dans <siteinfo>, il permet d'identifier le type de page.
  - **<id>**, identifiant de page.
  - **<revision>**, Chaque page contient le nombre de revision.
    - **<id>**, identifiant de revision
    - **<parentid>**, si elle n'est pas la première revision, donc la valeur de cet élément est la valeur <id> de la revision précédente. Si une revision ayant une valeur de <id> le plus grande dans une page, il est donc la version actuelle dans ce dump.
    - **<timestamp>,** date de publication de cette page.
    - **<contributor>**, information d'un utilisateur qui a publié cette revision.
      - **<id>**, identifiant d'utilisateur
      - **<username>**, nom d'utilisateur
      - **<ip>**, adresse IP d'utilisateur
    - **<minor>**
    - **<comment>**, n'importe quel texte, possible est le commentaire de modification.
    - **Text,** texte en format wikicode.
    - **<sha1>**, comme MD5, un identifiant de vérification.
    - **<model>**, un type de modèle de wiki.
    - **<format>**, un type de format de wiki.

L'exemple (2.2) illustre le schéma de la Figure 2-4.

```
(2.2)
<page>
    <title>Antoine Meillet</title>
    <ns>0</ns>
    <id>3</id>
    <revision>
      <id>5</id>
      <timestamp>2002-09-08T20:49:46Z</timestamp>
      <contributor>
        <username>Curry</username>
        <id>0</id>
      </contributor>
      <minor />
      <comment>*</comment>
      <text xml:space="preserve">'''Antoine Meillet''' (1866-1936) est le
principal [[Linguistes célèbres|linguiste]] français des premières décennies
du XXe siècle.

Étudiant à la [[Sorbonne]] à partir de 1885, il suit notamment les cours de
[[Michel Bréal]] au [[Collège de France]] et de [[Ferdinand de Saussure]] à
l'[[École Pratique Des Hautes études]]. En 1890, une mission d'un an dans le
Ca...
      </text>
      <sha1>gqxpuudnimhux7nr2c8xyoe2wsn209u</sha1>
      <model>wikitexte</model>
      <format>text/x-wiki</format>
    </revision>
    <revision/>
    <revision/>
</page>
```

2.4. Wikicode

Le texte de chaque révision est écrit dans un langage spécifique à Wikipédia. Ce langage, Wikicode (Wikitexte), indique la mise en forme, les liens, etc. C'est un langage de balisage léger qui définit la mise en forme de saisies de contenu d'utilisateur, le plus souvent utilisé pour écrire les pages wiki, il permet de décrire une page contenant des textes mis en forme à l'aide de textes et symboles d'Unicode. Wikipédia donne un guide pour rédiger le wiki avec une manière de syntaxe wiki (Aide_de_Wikipédia_sur_le_wikitexte, 2014).

- ❖ Bien que Wikipédia ait fourni des principes de rédaction, les contributeurs ne les respectent pas nécessairement. Ils peuvent, par exemple, mal utiliser le code, mal placer le texte, oublier la signature dans la discussion etc. La structure de l'article sera alors mal équilibrée.

2.5. Deux méthodes d'archivage

Une d'opération d'archivage est souvent effectuée quand une page de discussion est très lourde, très longue à charger et gêne la lecture. Cette opération peut se faire suivant deux méthodes :

- A) Déplacer toutes les versions historiques en changeant le nom de la page originale de "XXX" à "XXX/Archive 1 ". Puis créer une nouvelle page nommée "XXX", couper (c'est-à-dire copier et supprimer) les fils de discussions non terminées de la dernière version de page archivée, enfin coller les fils de discussions dans la nouvelle page. Cette méthode archive toutes les versions historiques, y compris celles des fils de discussions non encore terminées. Pour ensuite reconstituer l'historique complet, il faudra donc prendre en compte la page actuelle et celle archivée.

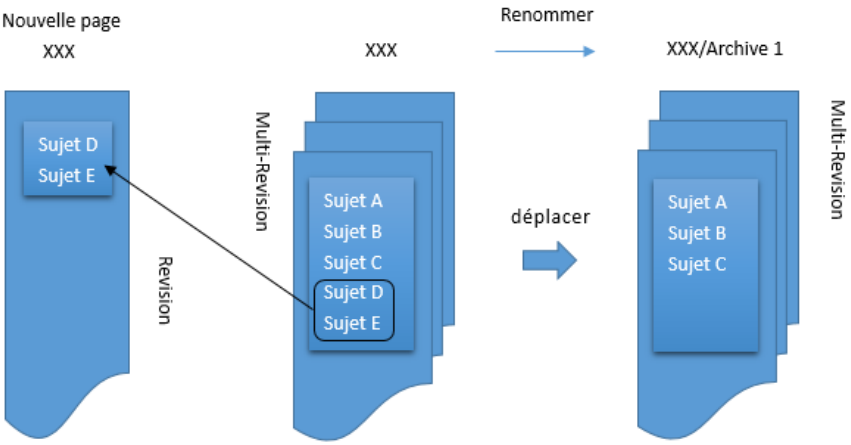

Figure 2-5 : première méthode d'archivage

- B) Cette méthode, au contraire de la précédente, ne déplace pas de version historique. L'éditeur coupe/colle les fils de discussions terminés de la page originale sur une nouvelle page nommé « XXX/Archive 1 ». Cette méthode garde toute les versions historiques. Ainsi, en principe, dans une seule page on veut retrouvera tout l'historique et compris les celui des nouvelles discussions.

Cependant, il peut arriver que des personnes continuent malgré tout à discuter dans l'ancienne page, devenue celle archivée !

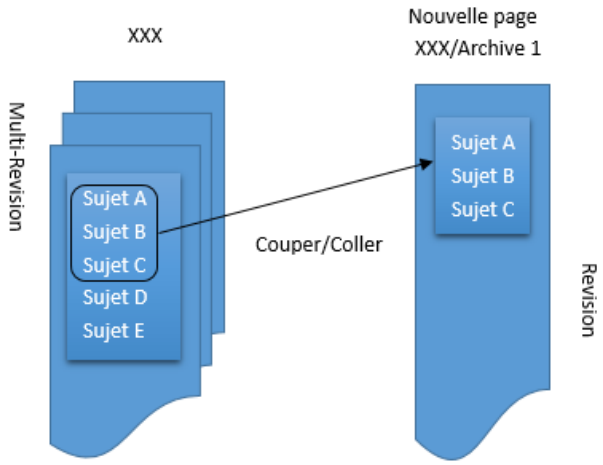

Figure 2-6 : seconde méthode d'archivage

Un autre problème peut également survenir : une discussion considérer comme terminée et donc archivée, peut être réouverte dans la nouvelle page. Il y aura alors doublon sur le titre du sujet entre la page principale et son archive.

2.6. Constitution de la version de "dépôt" du corpus Wikiconflits

Cette section explique comment a été constituée la version dit de "dépôt" du corpus Wikiconflits par le groupe CoMeRe-LRL. Une version dite de "dépôt" correspond dans les autres corpus CoMeRe à la version déposée par le compilateur (`compiler` suivant la terminologie OLAC) d'un corpus. Elle n'est donc pas en version TEI; Son format varie suivant les créateurs de la version de départ du corpus : XML maison, texte, Excel, etc. C'est à partir de cette version de dépôt que le groupe CoMeRe-LRL procède à la transformation en TEI.

Le lecteur constatera que le travail accompli correspond à la création d'un corpus. Il ne correspond pas à une simple extraction à partir de pages de Wikipedia. Il y a extraction, sélection, fusion et réarrangement de données. De plus des éléments supplémentaires sont traités de façon à enrichir le contexte d'élaboration des articles et discussions, en particulier une sélection des données sur les auteurs / éditeurs.

2.7. Liste de pages

Le tableau 2.1 rappelle les thèmes ou articles retenus pour le corpus Wikiconflits. A chaque article peuvent être associés une série de pages (avec leurs propres historiques) telles que listées dans le tableau 2.2.

Id	*Sujet*
1	Chiropratique
2	Eolienne
3	Histoire de la logique
4	Igor et Grichka Bogdanoff
5	Organisme génétiquement modifié
6	Psychanalyse
7	Quotient intellectuel

Tableau 2.1 : thèmes / articles retenus pour le corpus Wikiconflits

ID	*Page*
1	Page d'article
2	Page de discussion d'article
3	Page d'archive
4	Page d'autre discussion
5	Page d'utilisateur
6	Page de discussion d'utilisateur

Tableau 2.2 : type de pages associées à un article de Wikipédia

Étant donné que les pages dans la liste ci-dessus n'existent pas forcement pour tous les sujets, et qu'on a besoin d'extraire aussi d'autres pages touchant au même sujet, une liste contenant 38 pages est finalement sélectionnée manuellement (Annexes 7.1). Le tableau 2.3 en donne le nombre.

Id	*Sujet*	*Nombre de page*
1	Chiropratique	6 pages
2	Eolienne	3 pages
3	Histoire de la logique	2 pages
4	Igor et Grichka Bogdanoff	9 pages
5	Organisme génétiquement modifié	13 pages
6	Psychanalyse	6 pages
7	Quotient intellectuel	2 pages

Tableau 2.3 : nombre de pages retenues par thème.

2.8. Exportation avec SAX

L'exportation de corpus est basée sur le dump, du fait que la méthode en ligne n'autorise pas d'extraction de plus de 1000 pages ou révisions. Or les pages à sélectionner concernent beaucoup plus 1000 révisions. D'un autre côté les fichiers de type dump ont une taille très

importante une fois décompressés (plusieurs centaines de Go). Pour les traiter nous avons utilisé SAX.

SAX (anglais : *Simple API for XML*) est une interface de programmation pour de nombreux langages permettant de lire et de traiter des documents XML (Site_officiel_de_SAX, 2014).

SAX traite les documents, élément par élément, au fur et à mesure qu'ils sont rencontrés. Pour chaque élément (balise, commentaire, texte), la fonction de rappel correspondante est appelée (Simple_API_for_XML, 2014). Cela représente un avantage certains pour le traitement des gros fichiers au contraire des méthodes basées sur le DOM qui charge l'intégralité d'un document XML. Mais comme l'eau passe dans un tuyau, goutte par goutte, le temps de traitement est plus long que DOM.

Toutes les pages exportées gardent le format de Wikipédia décrit précédemment.

2.9. Pages de discussion et archive de pages de discussion

Nous avons présenté deux méthodes d'archivage sur Wikipédia. Du fait des problèmes indiqués, puis de la nécessité de reconstituer l'intégralité des fils de discussions, on a décidé de rassembler les fils de discussions présents dans les pages principales et les archives. Pour atteindre le but, il faut rechercher tous ces fils qui, en wikicode, commencent par un titre dans le format de (2.3)

```
(2.3)
== Titre ==
```

Un fil de discussion ne contient pas de balise signalant sa fin. La rencontre d'une autre balise sera alors considérée comme déterminant sa fin. Il faut également rechercher à la main les titres en doublons pour reconstituer un fil de discussions. In fine, nous avons rassemblé toutes les discussions dans un seul fichier pour un thème donné dans une structure comme (2.4). (2.5) l'illustre sur un exemple.

```
(2.4)
<list_title> : racine du fichier
<thread>: une fil de discussion
<title>: title
<content>: texte de discussion, attention: il faut garder le texte tel quel.
```

```
(2.5)
<list_title>
<thread>
        <title>== Title 1 ==</title>
        <content>{{Archive LANN|16636952}}
            * Débat précédent [[/1]]
        </content>
    </thread>
<thread>
        <title>== Title 2 ==</title>
        <content>{{Archive LANN|16636953}}
            * Débat précédent [[/2]]
        </content>
    </thread>
</list_title>
```

2.10. Page de participant

Reconstituer le contexte des discussions implique de retrouver les pages des principaux participants, tout comme leur page de discussions. La révision d'une page est accomplie par un et un seul éditeur à la fois. On appelle cette opération d'édition un « évènement (*Event*) ». Etant donné que chaque évènement s'associe à un utilisateur, nous avons exporté pour chaque page une liste d'évènements comme le montre l'exemple (2.6)

(2.6)

```
<event when="2002-09-30T19:06:09Z" who="Bam"/>
<event when="2002-09-30T19:10:38Z" who="Anthere"/>
<event when="2002-10-31T10:11:44Z" who="script de conversion"/>
<event when="2003-04-17T10:14:24Z" who="Olivier"/>
<event when="2003-08-11T23:55:13Z" who="Orthogaffe"/>
<event when="2003-08-11T23:56:27Z" who="Looxix"/>
<event when="2003-10-23T15:12:35Z" who="Ske"/>
<event when="2003-12-04T03:06:24Z" who="Alvaro"/>
```

Puis nous avons rassemblé toutes les listes de chaque sujet et construit une liste d'éditeurs associés à leur fréquence d'édition, comme illustré dans (2.7). Les statistiques sur l'ensemble des participants est dans l'annexe **Erreur ! Source du renvoi introuvable.**

(2.7)

```
197 Céréales Killer
165 Noisetier
140 YBM
100 Alain r
74  Keckel
67  Jean-Jacques Georges
63  Jean-no
60  Jean-Christophe BENOIST
59  Hégésippe Cormier
```

Suivant cette liste de participants, nous avons exporté toutes les pages correspondantes, soit au total 7928 pages (pour 3964 auteurs en comptant leurs pages et leur pages de discussions). Seules les participants ayant comptent (et non une simple adresse IP) ont été considéré ici.

Beaucoup de participants n'ayant réalisé qu'un nombre très restreint d'actes d'édition, compte tenu de leur nombre élevé, nous avons finalement décidé de garder seulement les participants qui ont une fréquence d'activité supérieure à 10 pour chaque sujet. Ceci représente au total 360 pages. Tous les calculs et sélections opérés sont listés dans le fichier `cmr-wikiconflits-freq_auteurs.xls` joint au corpus `cmr-wikiconflits`.

2.11. Rassemblement des modifications

Comme nous l'avons indiqué, Wikipédia stocke d'un côté pour un article la liste de toutes ces versions en texte complet. Il est donc difficile de déterminer où les modifications ont eu lieu et la redondance est très élevée. Par ailleurs nous avons indiqué que Wikipédia fournit pour chaque page un historique, qui ne contient que quelques éléments d'informations sur les modifications et les pointeurs sur les versions révisées. Lorsqu'un utilisateur désire connaître la teneur précise des révisions, il peut, sur la page d'historique sélectionner les versions qu'il désire comparer (cf. Figure 2-7)

Figure 2-7 : sélection pour comparaison entre deux versions

Un programme calcule alors dynamiquement les différences, comme le montre la Figure 2-8.

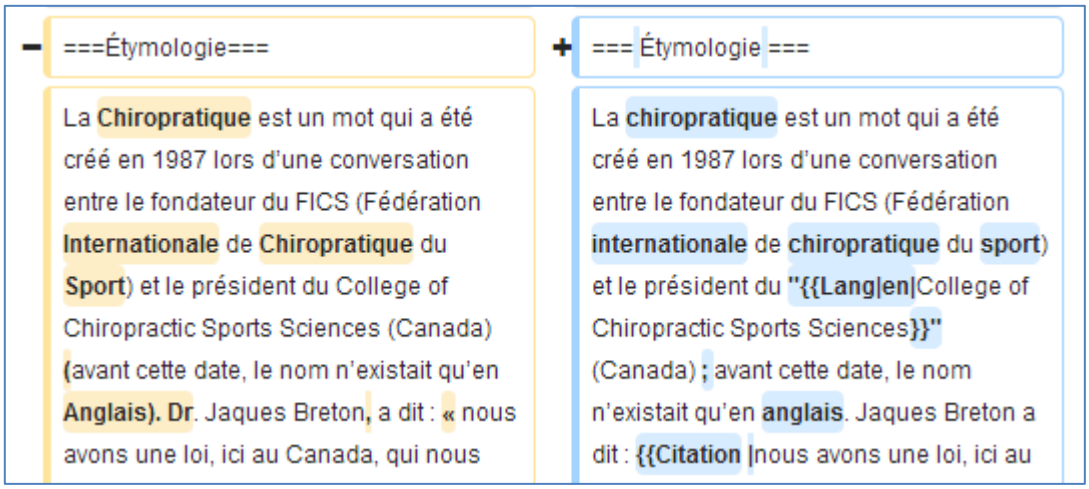

Figure 2-8 : comparaison entre deux versions alignées.

Le calcul n'étant que dynamique, il n'est pas possible de conserver ainsi les différences entre versions. Aussi avons-nous décidé d'utiliser un script développé par Stevage (JavaScript « EnhanceHIstory.user.js » voir (Broughton, 2008)).

Une fois le script chargé sous forme de plugiciel (« Tampermonkey »), dans le navigateur Chrome, il suffit d'aller dans une page historique. Un bouton supplémentaire « Show diffs » apparaît dans cette page (Figure 2-9).

Figure 2-9 : Bouton d'activation du calcul des différences entre versions

Cliquer sur ce bouton permet de générer une nouvelle page en format HTML qui contient toutes les différences (Figure 2-10).

Figure 2-10 : fichier HTML produit qui contient les modifications entre versions.

2.12. Organisation du corpus déposé.

Le corpus Wikiconflits créé par les moyens indiqués précédemment a été déposé dans le serveur Ajax de CoMeRe. La Figure 2-11 indique la structure du corpus dans le répertoire `depots`. Rappelons que ce répertoire correspond à celui où le collecteur d'un corpus vient le déposer avant passage en TEI. Donc ici, l'équipe du LRL agit en tant que premier déposant. Le travail ultérieur, par les autres membres de CoMeRe sera fait sur les formats TEI, donc dans le répertoire `tei-v1` et non `depots`.

❖ Le travail du **groupe Nouvelles Acquisitions** se faisant dans le répertoire tei-v1, la consultation du répertoire dépôt décrit dans cette section n'est pas essentielle.

On y retrouve un répertoire par thème, un pour les fichiers de statistiques et le dernier contenant les pages auteurs et les pages discussions des auteurs. La partie droite de la figure montre le contenu du répertoire `Eolienne` :

- `Eolienne.xml` : contient la page de l'article et inclut toutes les révisions de cet article. Son format est en wikicode.
- `Discussion-Eolienne_new.xml` : contient toutes les discussions associées à la page précédentes (i.e. la dernière version des discussions au jour d'extraction), c'est-à-dire celles qui figurent dans la page en ligne et celle contenues dans les discussions archivées, rassemblées par thème de discussions. Le fichier est en format wikicode également.
- `Discussion-Eolienne_Neutrialite-new.xml` : contient tous les fils de discussion associés au débat de neutralité (débat principal ou archivé). En wikicode.

- `Eolienne Historique des versions-wikipedia .html` : sauvegarde en format HTML de l'historique des discussions de la page article. Comme indiqué précédemment cet historique ne contient que des informations générales sur les révisions.
- Le répertoire `Eolienne-histo`.

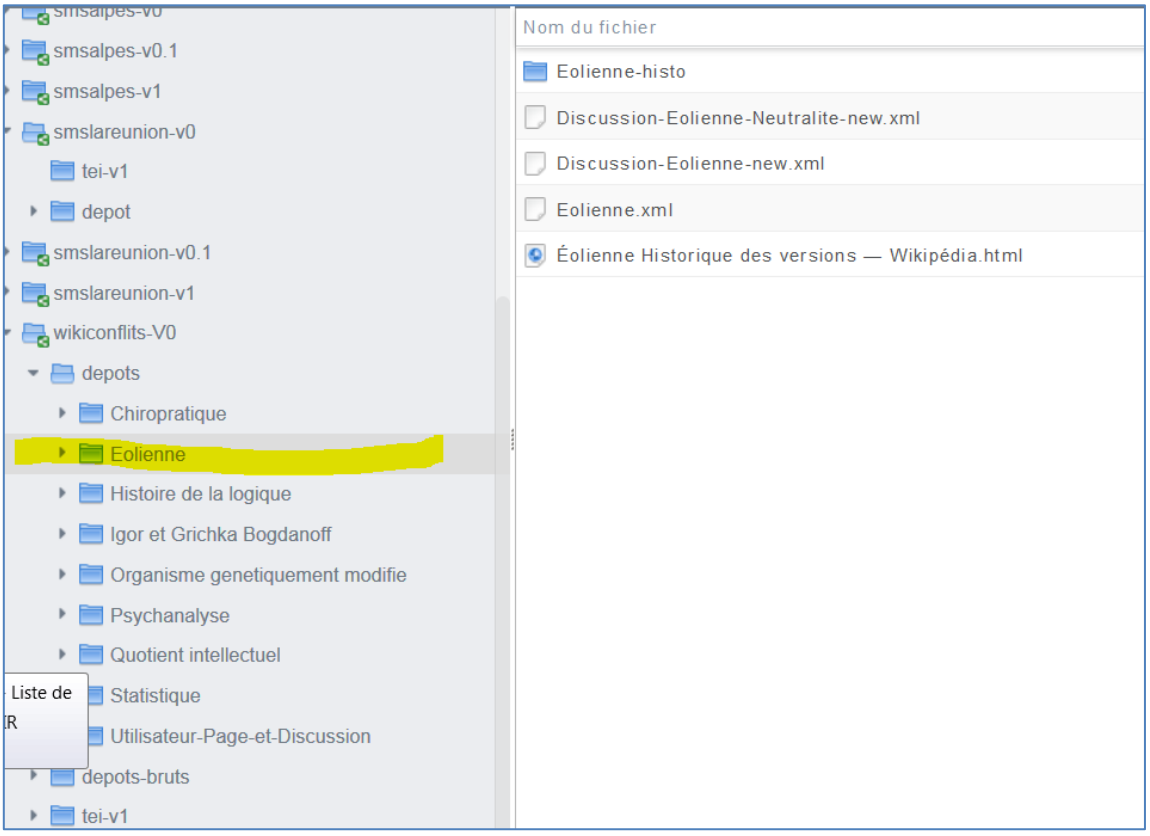

Figure 2-11 : Organisation des répertoires dans le serveur Ajax

Le contenu de ce dernier répertoire est indiqué dans la Figure 2-12.On y trouve 3 fichiers HTML renfermant **le détail des révisions** portant sur les 3 pages précédentes, à savoir celle de l'article, celle des discussions sur l'article et des discussions sur le conflit de neutralité. L'étude du codage des ajouts, suppressions de texte dans ces révisions sort du cadre du projet CoMeRe. Ces fichiers ont été générés de façon à permettre des recherches futures.

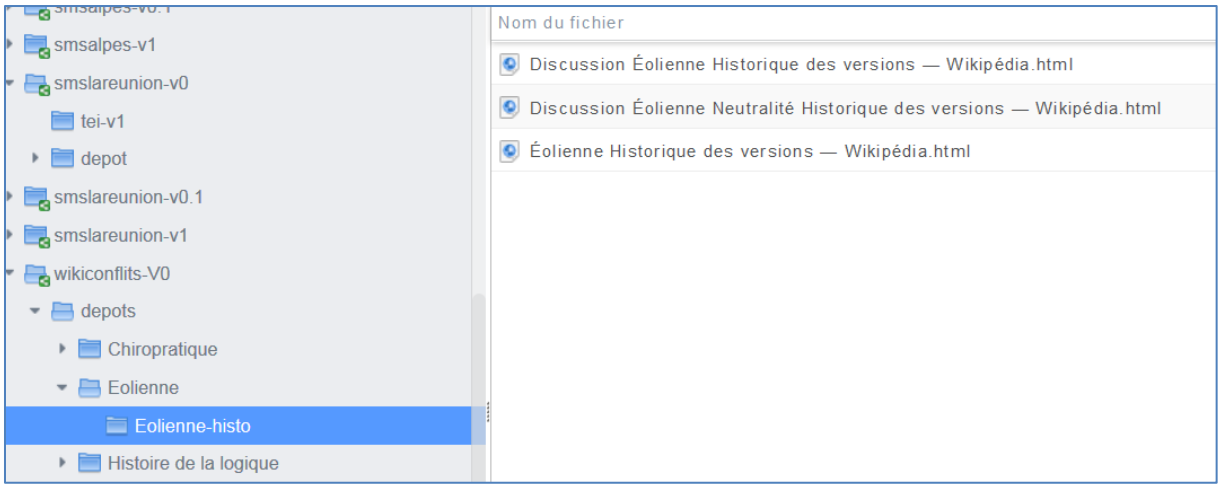

Figure 2-12 : detail du repertoire `Eolienne-histo`, maintenant copié dans répertoire `tei-v1`

Toutes les versions HTML ont été déposées en l'état avec le corpus en format TEI dans le répertoire `tei-v1` . Les mettre en format TEI dépasse le cadre du projet CoMeRe. Seules seront traduits en TEI les fichiers XML en format wikicode, ceux correspondant aux discussions (dernière version des discussions) et aux articles (toutes les versions). Ces fichiers HTML, tout comme les fichiers des auteurs et des pages discussions des auteurs ont un statut d'annexes par rapport aux fichiers TEI. Ils pourront être exploités dans des projets de recherche ultérieurs qui se concentreront sur l'étude des discussions.

2.13. Nouvelle nomenclature et organisation du répertoire tei-v1

La nomenclature des fichiers TEI a été fixée en juin 2014. La voici, avec un une explication sur l'organisation du répertoire `cmr-wikiconflits-V0 > tei-v1` dans lequel les collègues du **groupe Nouvelles Acquisitions** auront à travailler.

Ce répertoire a été organisé tel que dans la Figure 2-13 : contenu du répertoire `tei-v1`. On y reprend la nomenclature de la figure 2.15. Outre les répertoires sur chacun des sujets, figurent les fichiers XML et Excel lisant les contributeurs qui ont reçu un identifiant unique avec le répertoire Utilisateur-Page-discussion (renommé `Contributeurs-Page-Discussion`) contenant les pages des principaux contributeurs et leur page de discussion (non transformés en TEI) (voir détail dans section 3), le présent manuel dans sa dernière version et le schéma RNG de nos fichiers TEI.

Figure 2-13 : contenu du répertoire `tei-v1`

Par exemple tout ce qui concerne les éoliennes est dans le répertoire `cmr-wikiconflits-eolienne-tei-v1`. On y retrouve (voir Figure 2-14 : contenu du répertoire sur les éoliennes) un fichier TEI sur les discussions générales, un autre concernant la question de la neutralité. Le contenu de l'article étant important (plus de 100mo) a été découpé en deux sous-fichiers de taille plus réduite : `cmr-wikiconflits-eolienne_p1-tei-v1.xml` et `cmr-wikiconflits-eolienne_p2-tei-v1.xml`. Enfin les fichiers HTML présentant les historiques (voir section précédente) ont été copiés ici.

Figure 2-14 : contenu du répertoire sur les éoliennes

ident long	ident court	type	nom	nbc
cmr-wikiconflits-qi_discu-tei-v1	cmr-wiki-c001	wiki	Discussion-Quotient intellectuel	
cmr-wikiconflits-qi-tei-v1	cmr-wiki-c002	Wiki	Article Quotient Intellectuel	
cmr-wikiconflits-psychanalyse_discu-tei-v1	cmr-wiki-c003	wiki	Discussion-Psychanalyse	
cmr-wikiconflits-psychanalyse-tei-v1	cmr-wiki-c004	wiki	Article Psychanalyse	
cmr-wikiconflits-affaire_bogdanoff_discu-tei-v1	cmr-wiki-c005	wiki	Discussion-Affaire Bogdanoff	
cmr-wikiconflits-affaire_bogdanoff-tei-v1	cmr-wiki-c006	wiki	Article affaire Bogdanoff	
cmr-wikiconflits-bogdanoff-tei-v1	cmr-wiki-c007	wiki	Article Igor et Grichka Bogdanoff	
cmr-wikiconflits-bogdanoff_discu-tei-v1	cmr-wiki-c008	wiki	Discussion-Igor et Grichka Bogdanoff	
cmr-wikiconflits-bogdanoff_discu_neut-tei-v1	cmr-wiki-c009	wiki	Discussion-Igor et Grichka Bogdanoff-Neutralite	
cmr-wikiconflits-ogm_discu-tei-v1	cmr-wiki-c010	wiki	Discussion-Organisme genetiquement modifie	
cmr-wikiconflits-ogm_discu_neut-tei-v1	cmr-wiki-c011	wiki	Neutralite	
cmr-wikiconflits-ogm-tei-v1	cmr-wiki-c012	wiki	Article Organisme Génétiquement Modifié	
cmr-wikiconflits-debat_ogm_discu-tei-v1	cmr-wiki-c013	wiki	Discussion-Debat sur les organismes genetiquement modifies	
cmr-wikiconflits-debat_ogm-tei-v1	cmr-wiki-c014	wiki	Article Debat sur les organismes genetiquement modifies	
cmr-wikiconflits-eolienne_discu-tei-v1	cmr-wiki-c015	wiki	Discussion-Eolienne	
cmr-wikiconflits-eolienne-tei-v1	cmr-wiki-c016	wiki	Article Eolienne	
cmr-wikiconflits-eolienne_discu_neut-tei-v1	cmr-wiki-c017	wiki	Discussion-Eolienne-Neutralite	
cmr-wikiconflits-chiropratique_discu_neut-tei-v1	cmr-wiki-c018	wiki	Discussion-Chiropratique-Neutralite	
cmr-wikiconflits-chiropratique-tei-v1	cmr-wiki-c019	wiki	Article Chripratique	
cmr-wikiconflits-chiropratique_discu-tei-v1	cmr-wiki-c020	wiki	Discussion-Chiropratique	
cmr-wikiconflits-histologique_discu-tei-v1	cmr-wiki-c021	wiki	Discussion-Histoire de la logique	
cmr-wikiconflits-histologique-tei-v1	cmr-wiki-c022	wiki	Article Histoire dela Logique	

Figure 2.15 : nomenclature des fichiers TEI, articles et discussions.

3. Comment transformer en TEI les discussions et les articles ?

3.1. Format de la nouvelle structure des pages d'article

Dans cette partie, on va présenter le format de la page d'article en TEI. Les données de la construction viennent de deux fichiers indiqués dans la section 2.3.4 : page et page historique.

La base de traitement sur la page d'article est de garder toutes les révisions historiques. Vu que chaque page dans le dump est construite en contenant au moins une révision, cette dernière est représentée par un élément `<revision>`, et que plusieurs révisions correspond à plusieurs éléments, la structure principale en TEI est donc présentée en utilisant les éléments `<group>` et `<text>`, voir la figure 3.1.

Avant la transformation

```
<page>
   <revision>
      <id>6241859</id>
      <timestamp>2006-03-24T02:40:04Z</timestamp>
      <contributor>
         <username>Gene.arboit</username>
         <id>33541</id>
      </contributor>
      <comment>création ébauche à partir de [[en:]] et [[Discuter:Nombre
réel]]</comment>
      <text xml:space="preserve"><!-- Texte principal --></text>
      <sha1>trrnoy9luv2yzovvhe831k37847wjcm</sha1>
      <model>wikitext</model>
      <format>text/x-wiki</format>
   </revision>
   <revision>
      <id>6243361</id>
      <parentid>6241859</parentid>
      …
   </revision>
   <revision/>
   …
</page>
```

Après la transformation

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader/>
   <text>
      <front/>
      <group>
         <text xml:id="cmr-wiki-c022-rev6241859">
            <front>
               <div>
                  <head>Version Historique : 6241859</head>
                  <docAuthor corresp="psn:cmr-wikiconflits-
p33541">Gene.arboit</docAuthor>
                  <docDate when="2006-03-24T02:40:04Z"/>
                  <extent>
                     <measure unit="Octet" quantity="7065"
type="text_size">7 065</measure>
                     <measure unit="Octet" quantity="+7065"
type="change_size">+7 065</measure>
                     <code lang="sha1">trrnoy9luv2yzovvhe831k37847wjcm</code>
                  </extent>
                  <desc>création ébauche à partir de en: et Discuter:Nombre
réel</desc>
               </div>
```

```
            </front>
            <body><p><!-- texte --></p></body>
         </text>
         <text xml:id="cmr-wiki-c022-rev6243361" prev="#cmr-wiki-c022-rev6241859">
         <text xml:id="cmr-wiki-c022-rev6243520" prev="#cmr-wiki-c022-rev6243361">
         …
      </group>
   </text>
</TEI>
```

Figure 3-1 Transformation d'article en TEI

On a vu aussi dans la **Erreur ! Source du renvoi introuvable.**Figure 3-1 Transformation d'article en TEI que d'autres éléments ont été transformés en TEI. La Table 3-1 montre une convention pour transformer tous les éléments de dump en TEI, et compris les éléments venant de page historique (HTML).

Dump	TEI	Origine	Note
<revision>	<text>	Dump	
<id>	@xml:id de <text>	Dump	Ajout d'un préfixe
<parentid>	@prev de <text>	Dump	Ajout d'un préfixe
<timestamp>	<docDate when="timestamp">	Dump	Dump possède un type de valeur utilisé aussi par TEI
<username> de <contributor>	<docAuthor>	Dump	
<id> de <contributor>	@corresp de <docAuthor>	Dump	Valeur voir la section 3.5
<text>	<body>	Dump	Wikicode
<sha1>	<code lang="sha1">	Dump	
	<mesure>	HTML	
	<desc>	HTML	

Table 3-1 Convention de transformation de la structure de page d'article

3.2. Format de la nouvelle structure des pages de discussion

Dans la section 2, nous avons expliqué comment ont été rassemblés à la main les différents morceaux d'une discussion autour d'un thème donné (eg Psychanalyse), morceaux éventuellement dispersés et dupliqués dans plusieurs archives. Ces fichiers en format Wikipédia ont la nouvelle structure (3.1).

(3.1)Nouvelle Structure, format wikipedia

```
<list_title>
<thread>
        <title>== Title 1 ==</title>
        <content>{{Archive LANN|16636952}}
                AAAA
        </content>
         <thread>
            <title>== Title 1.2 ==</title>
            <content>{{Archive LANN|16636954}}
                BBBB
            </content>
        </thread>
</thread>
<thread>
        <title>== Title 2 ==</title>
        <content>{{Archive LANN|16636953}}
            CCCC
        </content>
    </thread>
</list_title>
```

La page est segmentée en nombreux **sujets de discussion** (*thread* en anglais) ou sous-sujets de discussion. Chaque sujet de discussion ressemble à une section dans un article. Celui-là contient un « title » et « content ». Si l'on convertissait directement cette page en TEI, cela donnerait un contenu tel qu'en (3.2).

(3.2)TEI partie <text>

```
<text>
   <body>
      <div>
         <head>Title 1</head>
         Contenu à traiter
         <div>
            <head>Title 1.2</head>
            Contenu à traiter
          </div>
      </div>
      <div>
          <head>Title 2</head>
         Contenu à traiter
       </div>
   </body>
<text>
```

Apparemment, cette structure présente une macro structure avec des sujets de discussion. Mais elle ne suffit pas pour présenter l'interaction interne dans un sujet de discussion, par exemple : auteur, date, relation de réponse etc. Par conséquent, on a besoin d'une analyse plus en détail, un traitement au niveau des messages.

3.3. Format de rédaction des discussions

Nous avons jusqu'à présent parlé de `thème`, `sujet` et `sous-sujet` de discussions, en évitant d'employer l'expression `fil de discussion`. En effet, après avoir mené une étude de la façon dont les discussions sont construites par les participants, on a trouvé que **l'organisation des discussions Wikipedia ne ressemble pas à celle bien connue des forums de discussion,** où le logiciel oblige explicitement les participants à utiliser des fils de discussion et où chaque message est clairement identifiable et nécessairement rattaché à un et un seul père (un seul message d'un seul fil de discussion).

Tout d'abord, la notion même de `message` (qui va correspondre pur nous à un élément `<post>`) est difficile à identifier. Disons, de façon approximative qu'il s'agit d'un morceau de texte (éventuellement composé de plusieurs paragraphes) terminé par la `signature` de son auteur.

Wikipédia a bien recommandé un format de rédaction afin de normaliser message et fil de discussion de la façon suivante lorsqu'un participant écrit directement dans le wikicode (l'interface est alors différente de celle du lecteur final). :

- Ajouter deux points « : » devant un message pour répondre à un autre message. Si le message auquel on répond était précédé de deux points, alors ajouter deux points supplémentaires. Lors de l'affichage (en lecture pour l'internaute), à chaque caractère deux points correspondra une indentation supplémentaire du message de réponse.
- Ajouter une signature à la fin de message : nom d'auteur, date etc.

Si les auteurs respectaient de telles consignes, les fils de discussion seraient aisés à identifier, comme le montre la Figure .

Figure 3.2 Vue pour l'internaute de la page discussion avec le format recommandé

Malheureusement, les discussions sont organisées de façon très variées. Assez souvent les auteurs ne respectent pas ces consignes. La Figure 3.3 en donne une illustration. Une personne tape explicitement les graphies `Réponse :` au début de son texte puis semble signé en faisant appel à la marque d'indentation, seulement pour cette signature. Ici la signature n'indique qu'une adresse IP et la date. On hésite à savoir où se termine le texte du premier auteur. Celui qui répond intervient semble-t-il deux fois, sans respecter les formats et semble terminer par une indication de signature, `Curry` (pas au sens Wikipédia cependant). Si l'on examine le lien associé à ce dernier mot, on trouve, non une page d'auteur mais une page générale de Wikipédia (cf. Figure) ! Traiter automatiquement de telles pages pose donc problème.

Je ne trouve nulle part le terme chiropratique dans le dictionnaire mais bien chiropraxie, est-ce normal ?

Réponse : non, d'autant que les vrais termes Francophones sont "chiropratique" et "chiropraticiens" suivant la logique que l'on ne dit pas "un practeur de santé". Les Québecois, les Suisses et les Belges ne s'y sont pas trompé qui utilisent tous ces bons termes. L'Académie Française recommande d'ailleurs officiellement les termes "chiropratique" et "chiropraticien" comme celà est mentionné dans les bons dictionnaires.

> 212.195.212.116, 10 juillet 2005 à 14:45

Je trouve chiropractie ou chiropraxie mais pas chiropratique, des experts peuvent confirmer ? Shaihulud

Réponse : vous trouvez de mauvaises traductions littérales (des barbarismes linguistiques) du terme. Le vrai terme officiel en bon Français est chiropratique, à savoir "pratique des mains".

> 212.195.212.116, 10 juillet 2005 à 14:45

Je ne suis pas expert, mais nous devons avoir le même dictionnaire, car je ne trouve moi aussi que ces deux mots. Curry

Figure 3.3 Vue d'une discussion ne respectant pas le format recommandé

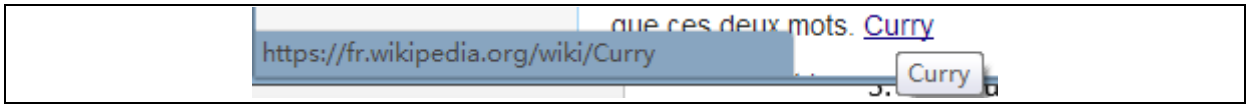

Figure 3.4 : HyperLien

3.4. Segmentation de messages : quels critères appliquer ?

L'objectif est de segmenter automatiquement chaque message (un premier texte sur un sujet chaque réponse et sous-réponse). Afin d'atteindre le but, on est obligé de définir des règles de segmentation. Etant donné les fortes variations observées, nous avons décidé de créer des règles dépendant du format recommandé par Wikipédia.

Par conséquent, la définition d'un message (`<post>`) est une suite de texte qui :

- commencée après un <title> fils de <thread> (délimiteur, voir le texte suivant) ou un autre <post> ;
- terminée par un délimiteur « fin de message » ou « fin de page » ;
- précédé un autre `<post>` ayant un niveau différent d'indentation.

Délimiteurs possibles:

- `<title>` de thread (utilisé pour segmenter les <thread>)
- Niveau d'indentation (suite de caractères deux points)
- Fin de message:
  - Ligne terminé par le module {{non signé}}
  - Ligne terminé par
    - `[[Utilisateur : XXX]]` avec ou sans Date
    - `[[Spécial:Contributions/xxx.xxx.xxx.xxx]]` avec ou sans Date
    - `[[Utilisateur : XXX]] [Discussions Utilisateur : XXX]` avec ou sans Date.
- Fin de page

Pour « fin de message », il faut prendre en compte aussi les formats des autres langues et la casse des lettres, par exemple :

- `[[utilisateur : XXX]]` avec ou sans Date
- `[[User : XXX]]` avec ou sans Date
- `[[user  talk : XXX]]` avec ou sans Date

3.5. Structure d'un message : <post>

En accord avec les travaux du groupe TEI-CMC (2014), les travaux déjà entamé par les collègues allemands sur Wikipédia (Margareta & Lunge, 2014) et dans la continuité des autres corpus CoMeRe, chaque message sera balisé par l'élément `<post>` est un élément principal utilisé dans tous les projets CoMeRe.

Les attributs et sous-éléments XML associés sont :

- `@xml:id` identifiant du post. Sa valeur est unique dans le fichier XML entier ;
- `@when-iso`, date de post, sa valeur est sous format ISO 8601 ;
- `@when-custom`, date personnalisée, on utilise pour présenter en cas qu'un post n'a pas de date, la valeur par défaut est « unknown » ;
- `@who`, auteur de post, sa valeur voir la section 3.6 ;
- `@n`, présente le niveau d'indentation ;
- `@ref` identifiant du post auquel le post courant répond. Lorsque le niveau d'indentation est d'indentation est « 0 », cet attribut n'apparaît pas ;
- `<p>` Un paragraphe de message / post contient un ou plusieurs `<p>` ;
- `<signed>`, signature de message. Un post contient un ou zéro élément `<signed>`.

La figure 3.5 donne un exemple de wikitexte segmenté en deux messages du fait des deux niveaux d'indentations différents : niveau 2 et niveau 3: et du délimiteur de fin de message. Afin de comprendre toutes les transformations opérées, on se reportera à la section sur le wikitexte.

Wikitexte

```
::L'article a déjà été l'objet d'un blocage partiel en raison d'une guerre d'édition. Il est raisonnable de venir discuter ici avant de faire des modifications de même nature qu'auparavant. Je vous suggère de faire valoir vos arguments ici, avant mise à jour de l'article, et d'obtenir un consensus. [[Utilisateur:Vincent Lextrait|Vincent Lextrait]] ([[Discussion utilisateur:Vincent Lextrait|d]]) 15 avril 2012 à 18:51 (CEST)

:::La guerre d'édition reste due à vos modifications non-neutres. les arguments sont exposés  sur cette page, je vous prie de réverter vous-même afin de revenir à la version du 12 avril 2012 à 18h11.[[Spécial:Contributions/82.229.70.91|82.229.70.91]] ([[Discussion utilisateur:82.229.70.91|d]]) 15 avril 2012 à 20:23 (CEST)
```

TEI

```
<post xml:id="cmr-wiki-c002-a83" when-iso="2012-04-15T18:51" who="psn: cmr-wikiconflits-p192844" n="2" ref="#cmr-wiki-c002-a82">
      <p>L'article a déjà été l'objet d'un blocage partiel en raison d'une guerre d'édition. Il est raisonnable de venir discuter ici avant de faire des modifications de même nature qu'auparavant. Je vous suggère de faire valoir vos arguments ici, avant mise à jour de l'article, et d'obtenir un consensus. </p>
     <signed><ref target="https://fr.wikipedia.org/wiki/Utilisateur:vincent_lextrait">Vincent Lextrait</ref>(<ref target="https://fr.wikipedia.org/wiki/Discussion_utilisateur:vincent_lextrait">d</ref>) 15 avril 2012 à 18:51 (CEST)
     </signed>
</post>


<post xml:id="cmr-wiki-c002-a84" when-iso="2012-04-15T20:23" who=" psn:cmr-wikiconflits-ip01383" n="3" ref="#cmr-wiki-c002-a83">
      <p>La guerre d'édition reste due à vos modifications non-neutres. les arguments sont exposés sur cette page, je vous prie de réverter vous-même afin de revenir à la version du 12 avril 2012 à 18h11.</p>
      <signed><ref target="https://fr.wikipedia.org/wiki/Spécial:contributions/82.229.70.91">82.229.70.91</ref> (<ref target="https://fr.wikipedia.org/wiki/Discussion_utilisateur:82.229.70.91">d</ref>) 15 avril 2012 à 20:23 (CEST)
      </signed>
</post>
```

Figure 3.5 : exemple de discussion en format wikitexte et sa traduction en TEI-CMC.

3.6. Traitement des contributeurs / auteurs

Un nouveau traitement a été fait (aout 2014) afin d'unifier les identifiants des contributeurs sur toutes les parties du corpus *cmr-wikiconflits-tei-v1*. Ce travail a été accompli suite aux discussions internes au projet CoMeRe, au vu des inconsistances apparaissant dans les pages Wikipédia et, enfin, pour répondre à un critère qualité habituel dans l'élaboration des corpus en linguistique (un et un seul identifiant pour le même participant dans tout le corpus). Tous les contributeurs ont été rassemblés et identifiés dans un fichier TEI unique (`cmr-wikiconflits-contributors-tei-v1.xml`). Chaque fichier TEI du corpus va donc explicitement renvoyer à ce document lorsqu'il s'agit d'identifier les contributeurs pour un article, comme pour une discussion.

3.6.1. Type d'identifiant

Chaque utilisateur dans ce corpus possède un identifiant unique, ça permet de servir une analyse concernant les utilisateurs entiers. Vu que l'utilisateur de Wikipédia soit une

personne inscrite, soit une adresse IP, et que la personne inscrite peut changer leur alias, et même que un alias inexistant, on les représente en utilisant trois types d'ID principaux et un type d'ID dirigeant l'auteur inconnu.

```
<person xml:id="cmr-wikiconflits-p6148">
  <persName>Kôan</persName>
  <persName>Ataraxie</persName>
</person>
```

a. Multi-alias

```
<person xml:id="cmr-wikiconflits-ip00002">
  <persName>84.100.8.208</persName>
</person>
```

b. Adresse IP

```
<person xml:id="cmr-wikiconflits-p_Chouca">
  <persName>Chouca</persName>
  <note>Utilisateur non pas trouvé dans les révisions correspondantes</note>
</person>
```

c. Alias inexistant

```
<person xml:id="cmr-wikiconflits-p_unknown">
  <persName>unknown</persName>
</person>
```

d. Auteur inconnu

Figure 3.6 : Types d'ID produit par le programme CoMeRe dans le fichier `cmr-wikiconflits-contributors-tei-v1.xml`

➢ Type I : auteur inscrit

Wikipédia distribue automatiquement à chaque utilisateur inscrit un identifiant unique dans le fichier Dump des révisions. (Voir la section 2.3.4) On prend cet identifiant en ajoutant une préfix du projet, `cmr-wikiconflits-pID.`

Un même auteur inscrit et reconnu par Wikipédia peut modifier son pseudonyme. Il apparaîtra alors avec deux noms d'auteur différents à deux moments différents (voir **Erreur ! Source du renvoi introuvable.** cas a). Nous avons rassemblé manuellement les deux pseudonymes derrière le même identifiant.

➢ Type II : auteur non inscrit avec IP

L'adresse IP n'a pas cette offre, on peut utiliser cette adresse IP comme identifiant. Mais si l'on ajoute le préfixe du projet, l'identifiant serait très long, on distribue à chaque adresse IP un nombre N unique (donc un identifiant construit par le programme CoMeRe), et le préfixe change vers `cmr-wikiconflits-ipN` .

➢ Type III : auteur inconnu

Cas peu fréquent. Cette situation peut apparaître dans deux cas différents :

- 1) un premier auteur a recopié une partie de texte / discussion provenant des thèmes non traités ici. Dans le texte recopié apparait un nom d'auteur, qui peut

exister dans Wikipédia, mais qui ne fait pas partie des auteurs contribuant à l'article du thème, ni à la page de discussion du thème. Il ne fait donc pas partie de notre liste d'auteurs

- 2) l'auteur peut être complètement inconnu du système Wikipédia et n'a pas été repéré par son IP. La Figure montre un cas où quelqu'un aurait pu utiliser le nom de Bogdanov. Une page Bogdanov peut alors avoir été créée, mais a ensuite été supprimée par d'autres utilisateurs qui considéraient qu'il ne s'agissait pas du véritable Bogdanov.

Dans ces deux cas, l'identifiant CoMeRe sera composé ainsi

`cmr-wikiconflits-p_NOM`, *NOM* étant la chaîne de caractère utilisée à l'emplacement auteur du wikicode.

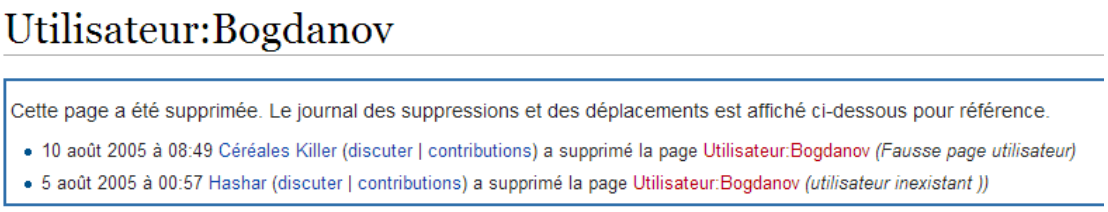

Figure 3.7 : utilisation erroné du nom Bogdanov en tant que auteur. Des utilisateurs ont alors supprimés les pages utilisateurs correspondantes.

➢ Type IV : signature non détectée

Nous avons indiqué que nos programmes n'avaient pas réussi à détecter les signatures d'environ 10% des cas (du fait, notamment du non-respect par les auteurs des conventions d'organisation des discussions). On a donc des posts sans auteur, dans ce cas on utilise un ID pour le représenter `cmr-wikiconflits-p_unknown`.

3.6.2. Liste unique de contributeurs et référence dans chaque fichier du corpus cmr-wikiconflits

Tous les utilisateurs ayant participé à tous les articles des thèmes de `cmr-wikiconflits-tei-v1` ou à toutes les discussions liées à ces thèmes sont rassemblés dans un seul et unique fichier TEI `cmr-wikiconflits-contributors-tei-v1.xml`. L'identifiant des contributeurs, ainsi que leur pseudonymes ont été rassemblés grâce aux traitements automatiques et manuelles décrits dans la section précédente.

Maintenant que l'on a une liste de personnes, on a besoin de les référer dans chaque corpus. TEI propose une méthode de référence dynamique, c'est que l'on déclare un préfixe dans `<teiHeader>` et utilise ce préfixe pour tous les références correspondantes. On a déclaré un préfixe `psn` pour représenter la personne, voir le détail de déclaration dans le site de TEI (Using Abbreviated Pointers, 2014).

```
(3.3)
    <encodingDesc>
      <listPrefixDef>
        <prefixDef ident="psn" matchPattern="([a-zA-Z0-9\-_]+)"
replacementPattern="../cmr-wikiconflits-contributors-tei-v1.xml #$1">
          <p>In the context of the project Wikiconflits, private URIs with the
prefix "psn"  point to <gi>person</gi> elements in the projects's cmr-
wikiconflits-contributors-tei-v1.xml file.</p>
        </prefixDef>
```

```
      </listPrefixDef>
    </encodingDesc>
```

Cette déclaration doit être faite dans le `<teiHeader>` de chaque corpus. Dans la partie <body> de ces fichiers. Pour une discussion, il s'agit de l'auteur d'un `<post>` se verra identifié par l'attrbut `@who` de la façon suivante (3.4).

```
(3.4)
<post xml:id="cmr-wiki-c020-a1" when-iso="2006-04-29T19:57" who="psn:cmr-
wikiconflits-p52690" n="0">
```

Pour les articles, l'identifiant de l'auteur apparaît dans `<docAuthor>` (3.5)

```
(3.5)
<docAuthor corresp="psn:cmr-wikiconflits-p33541">Gene.arboit</docAuthor>
```

3.6.3. Récapitulatif sur les auteurs / contributeurs

Le corpus `cmr-wikiconflits` contient donc trois types d'informations sur les contributeurs :

- `cmr-wikiconflits-freq_auteurs.xls` : statistiques sur tous les contributeurs, ceux qui ont été sélectionnés
- répertoire `Contributeurs-Page-Discussion` : liste des pages et des pages de discussion des principaux contributeurs (format Wikipédia)
- `cmr-wikiconflits-contributors-tei-v1.xml` : fichier listant ces contributeurs et leurs identifiants (format TEI)

3.7. Correction manuelle du traitement en TEI sur les discussions

3.7.1. Explications

Les traitements construits sur les principes de segmentation présentés en section découperont donc les discussions bien formées. Pour les autres, ne suivant pas les recommandations de Wikipédia, **il conviendra de faire un traitement à la main, à partir du résultat des segmentations opérées automatiquement, donc à partir des fichiers TEI.**

❖ Ce travail incombe au groupe CoMeRe sur les nouvelles acquisitions car il implique de lire les textes, de les interpréter correctement afin de rétablir les vrais fils de discussion, fils que les auteurs n'ont pas su établir correctement. Toutefois, nous recommandons, de commencer les vérifications et corrections par les discussions les mieux organisées (donc ne pas commencer avec Chiropratique). Par ailleurs, si une partie des discussions s'avère trop mal formées, on peut décider de ne pas la corriger à condition d'étiqueter la partie ignorée comme l'indique la section 3.7.3.

Voici quelques cas fréquents de discussions / messages mal formés :

- Utilisation de « `:` » pour une présentation de citation, qui aurait dû en fait être balisée par `<blockquote>` ;
- Réponse coupant le message précédant, voir l'exemple donné ci-dessus ;
- Message sans signature et même sans le module `{{non signé}}` ;
- Signature ne correspondant pas au format défini.

Pour retrouver la signature d'un post, le facteur de position peut être pris en compte avec les quatre cas listés dans le Table 3-2.

	Milieu de page	*Fin de page*
Sans signature depuis le départ	+	+
Avec blockquote	+	
Coupé par une réponse	+	
Non détecté par le système	+	+

Table 3-2 : aide au repérage de signature. Le symbole « + » indique la possibilité d'un post non pas détecté la signature

3.7.2. Premier exemple de correction

La qualité de la structure d'un fil de discussion varie suivant l'évolution du système de Wikipédia. Dans Chiropratique les discussions plus ancienne sont mal formées. En revanche, celles plus récentes sont mieux formées.

La **Erreur ! Source du renvoi introuvable.** donne un exemple de traitement automatique TEI correspondant à la Table 3-2. On notera que l'identifiant `psn:cmr-wikiconflits-p_unknown` contenu dans l'attribut `@who` de certains `<post>` du corpus `cmr-wiki-c020` correspond à un auteur factice. Cela veut dire que le traitement n'a pas réussi à identifier l'auteur du `<post>`. L'attribut @`when-custom` et sa valeur `"unknown"` indique la date de `<post>` n'est pas identifié.

```
Avant la correcte manuelle
<post xml:id="cmr-wiki-c020-a3" when-custom="unknown" who="psn:cmr-wikiconflits-
p_unknown" n="0">
 <p>Je ne trouve nulle part le terme chiropratique dans le dictionnaire mais bien
chiropraxie, est-ce normal ?</p>
 <p><hi rend="bold">Réponse</hi> : non, d'autant que les vrais termes Francophones
sont "chiropratique" et "chiropraticiens" suivant la logique que l'on ne dit pas
"un practeur de santé". Les Québecois, les Suisses et les Belges ne s'y sont pas
trompé qui utilisent tous ces bons termes. L'Académie Française recommande
d'ailleurs officiellement les termes "chiropratique" et "chiropraticien" comme
celà est mentionné dans les bons dictionnaires.</p>
</post>
<post xml:id="cmr-wiki-c020-a4" when-iso="2005-07-10T14:45" who="psn:cmr-
wikiconflits-p_unknown" n="1" ref="#cmr-wiki-c020-a3">
 <p>212.195.212.116, 10 juillet 2005 à 14:45</p></post>
<post xml:id="cmr-wiki-c020-a5" when-custom="unknown" who="psn:cmr-wikiconflits-
p4" n="0">
 <p>Je trouve chiropractie ou chiropraxie mais pas chiropratique, des experts
peuvent confirmer ? </p>
<signed><ref
target="https://fr.wikipedia.org/wiki/Utilisateur:shaihulud">Shaihulud</ref></sign
ed><p><hi rend="bold">Réponse</hi> : vous trouvez de mauvaises traductions
littérales (des barbarismes linguistiques) du terme. Le vrai terme officiel en bon
Français est chiropratique, à savoir "pratique des mains".</p>
</post>
<post xml:id="cmr-wiki-c020-a6" when-iso="2005-07-10T14:45" who="psn:cmr-
wikiconflits-p_unknown" n="1" ref="#cmr-wiki-c020-a5">
 <p>212.195.212.116, 10 juillet 2005 à 14:45</p>
</post>
<post xml:id="cmr-wiki-c020-a7" when-iso="2003-09-21T11:49" who="psn:cmr-
wikiconflits-p787" n="0">
 <p>Je ne suis pas expert, mais nous devons avoir le même dictionnaire, car je ne
trouve moi aussi que ces deux mots. <ref
target="https://fr.wikipedia.org/wiki/curry">Curry</ref></p>
 <p>Dans mon petit larousse illustré, on trouve les 3 termes, chiropratique étant
le terme canadien authentique.</p>
 <p> Peut-on dire que c'est une profession? N'est-ce pas plutôt une branche de la
médecine? </p>
<signed><ref
target="https://fr.wikipedia.org/wiki/Utilisateur:guillaume_bokiau">Guillaume
Bokiau</ref>21 sep 2003 à 11:49 (CEST)</signed>
</post>
```

Figure 3.8 : résultat du traitement d'une partie de discussion suivant le découpage de la Table 3-2

Pour corriger cela, servons-nous des pages concernant l'historique des discussions. La figure 3.9 montre une comparaison des deux premières révisions. La première révision (gauche) ne contient aucun texte. Les textes apparaissent pour la première fois dans la deuxième révision : `2002-09-30T21:53`. L'auteur de la révision est « Anthere ». En lisant ce texte, on en déduit que le texte peut se décomposer en quatre `<post>`. Dans les trois premiers `<post>`, impossible d'identifier la date de publication. On ajoute donc un préfixe `before_` de la date de révision : `before_2002-09-30T21:53`.

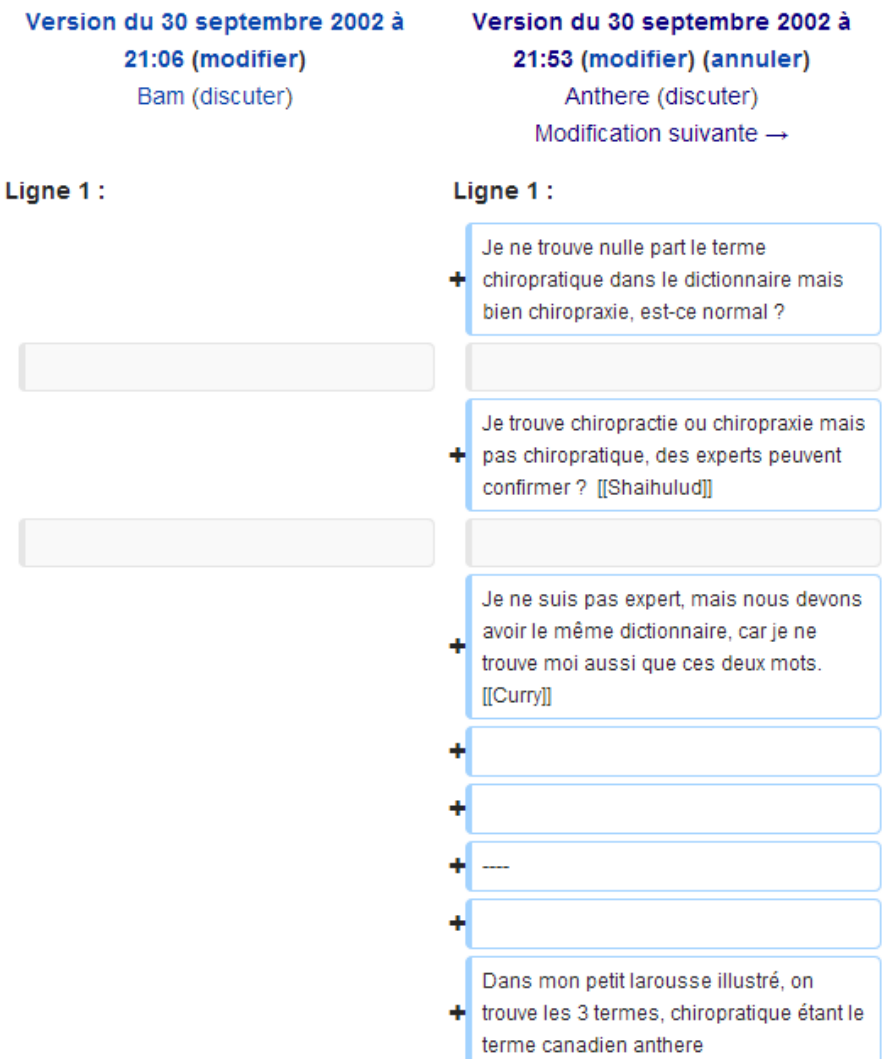

Figure 3.9 Comparaison des deux premières révisions

Les identifiants d'auteurs, `psn:mr-wiki-c020-ip00401`, `psn:cmr-wiki-c020-p4` et `psn:cmr-wiki-c020-p0` ont été repérés dans liste des auteurs contenu dans le `cmr-wikiconflits-contributors-tei-v1.xml`. Dans la figure 3.9, on ne trouve plus le mot « anthere » qui correspond à la signature de « Anthere », car dans la révision de `2005-07-12T19:13`, un utilisateur de IP `212.195.212.116` l'a remplacé par « authentique ». Ce texte est conservé jusqu'à la dernière version qu'on utilise. Dans ce cas, on met seulement l'auteur dans l'attribut @who. Voir la figure 3.10 le résultat final de la transformation manuelle

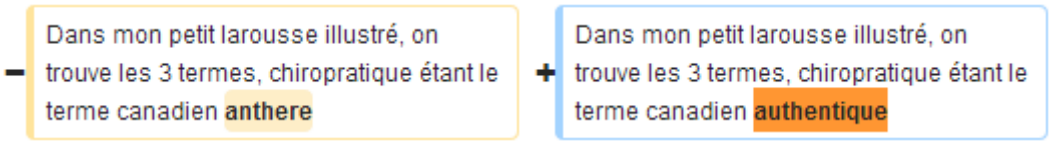

Figure 3.9 Capture d'écran de Wikipédia sur une erreur de modification

Après la correcte manuelle

```
<post xml:id="cmr-wiki-c020-a3" when-custom="~~unknown~~before_2002-09-30T21:53"
who="psn:cmr-wikiconflits-p_unknown" n="0">
 <p>Je ne trouve nulle part le terme chiropratique dans le dictionnaire mais bien
chiropraxie, est-ce normal ?</p>
</post>
<post xml:id="cmr-wiki-c020-a4" when-iso="2005-07-10T14:45" who="psn:cmr-
wikiconflits-ip00401" n="1">
 <p><hi rend="bold">Réponse</hi> : non, d'autant que les vrais termes Francophones
sont "chiropratique" et "chiropraticiens" suivant la logique que l'on ne dit pas
"un practeur de santé". Les Québecois, les Suisses et les Belges ne s'y sont pas
trompé qui utilisent tous ces bons termes. L'Académie Française recommande
d'ailleurs officiellement les termes "chiropratique" et "chiropraticien" comme celà
est mentionné dans les bons dictionnaires.</p>
~~</post>~~
~~<post xml:id="cmr-wiki-c020-a4" when-iso="2005-07-10T14:45" who="psn:cmr-~~
~~wikiconflits-p_unknown" n="1" ref="#cmr-wiki-c020-a3">~~
 ~~<p>~~<signed>212.195.212.116, 10 juillet 2005 à 14:45</signed>~~</p>~~</post>
<post xml:id="cmr-wiki-c020-a5" when-custom="before_2002-09-30T21:53~~unknown~~"
who="psn:cmr-wikiconflits-p4" n="0">
 <p>Je trouve chiropractie ou chiropraxie mais pas chiropratique, des experts
peuvent confirmer ? </p>
<signed><ref
target="https://fr.wikipedia.org/wiki/Utilisateur:shaihulud">Shaihulud</ref></signe
d>
</post>
<post xml:id="cmr-wiki-c020-a6" when-iso="2005-07-10T14:45" who=" psn:cmr-
wikiconflits-ip00401" n="1" ref="#cmr-wiki-c020-a5">
<p><hi rend="bold">Réponse</hi> : vous trouvez de mauvaises traductions littérales
(des barbarismes linguistiques) du terme. Le vrai terme officiel en bon Français
est chiropratique, à savoir "pratique des mains".</p>
~~</post>~~
~~<post xml:id="cmr-wiki-c020-a6" when-iso="2005-07-10T14:45" who="psn:cmr-~~
~~wikiconflits-p_unknown" n="1" ref="#cmr-wiki-c020-a5">~~
 ~~<p>~~<signed>212.195.212.116, 10 juillet 2005 à 14:45</signed>~~</p>~~
</post>
<post xml:id="cmr-wiki-c020-a7" when-iso="before_2002-09-30T21:53" who="psn:cmr-
wikiconflits-p0" n="1" ref="#cmr-wiki-c020-a5">
 <p>Je ne suis pas expert, mais nous devons avoir le même dictionnaire, car je ne
trouve moi aussi que ces deux mots. </p>
<signed><ref target="https://fr.wikipedia.org/wiki/curry">Curry</ref></signed>
</post>
<post xml:id="cmr-wiki-c020-a8" when-iso="2002-09-30T21:53" who="psn:cmr-
wikiconflits-p0" n="0">
 <p>Dans mon petit larousse illustré, on trouve les 3 termes, chiropratique étant
le terme canadien authentique.</p>
</post>
<post xml:id="cmr-wiki-c020-a9" when-iso="2003-09-21T11:49" who="psn:cmr-
wikiconflits-p787" n="0">
 <p> Peut-on dire que c'est une profession? N'est-ce pas plutôt une branche de la
médecine? </p>
<signed><ref
target="https://fr.wikipedia.org/wiki/Utilisateur:guillaume_bokiau">Guillaume
Bokiau</ref> 21 sep 2003 à 11:49 (CEST)</signed>
</post>
```

Figure 3.10 : résultat de la transformation manuelle

3.7.3. Indiquer les passages problématiques non corrigés

Le travail de correction manuelle pouvant s'avérer trop compliqué à certains endroits, il convient d'indiquer explicitement qu'une fraction de discussions est problématique et n'a pas été corrigé. Ainsi un traitement automatique pourra ignorer ce passage. Pour ce faire, on entourera le passage avec la balise `<div>` et son attribut `@subtype`, en lui donnant la valeur `ill-formed-discussion`, cf. figure 3.11.

```
<div subtype="ill-formed-discussion">
<post xml:id="cmr-wiki-c020-a3" when-custom="unknown" who="psn:cmr-wikiconflits-p_unknown" n="0">
 <p>Je ne trouve nulle part le terme chiropratique dans le dictionnaire mais bien chiropraxie, est-ce normal ?</p>
 <p><hi rend="bold">Réponse</hi> : non, d'autant que les vrais termes Francophones sont "chiropratique" et "chiropraticiens" suivant la logique que l'on ne dit pas "un practeur de santé". Les Québecois, les Suisses et les Belges ne s'y sont pas trompé qui utilisent tous ces bons termes. L'Académie Française recommande d'ailleurs officiellement les termes "chiropratique" et "chiropraticien" comme celà est mentionné dans les bons dictionnaires.</p>
</post>
<post xml:id="cmr-wiki-c020-a4" when-iso="2005-07-10T14:45" who="psn:cmr-wikiconflits-p_unknown" n="1" ref="#cmr-wiki-c020-a3">
 <p>212.195.212.116, 10 juillet 2005 à 14:45</p></post>

<post xml:id="cmr-wiki-c020-a5" when-custom="unknown" who="psn:cmr-wikiconflits-p4" n="0">
 <p>Je trouve chiropractie ou chiropraxie mais pas chiropratique, des experts peuvent confirmer ? </p>
<signed><ref target="https://fr.wikipedia.org/wiki/Utilisateur:shaihulud">Shaihulud</ref></signed><p><hi rend="bold">Réponse</hi> : vous trouvez de mauvaises traductions littérales (des barbarismes linguistiques) du terme. Le vrai terme officiel en bon Français est chiropratique, à savoir "pratique des mains".</p>
</post>

<post xml:id="cmr-wiki-c020-a6" when-iso="2005-07-10T14:45" who="psn:cmr-wikiconflits-p_unknown" n="1" ref="#cmr-wiki-c020-a5">
 <p>212.195.212.116, 10 juillet 2005 à 14:45</p>
</post>

<post xml:id="cmr-wiki-c020-a7" when-iso="2003-09-21T11:49" who="psn:cmr-wikiconflits-p787" n="0">
 <p>Je ne suis pas expert, mais nous devons avoir le même dictionnaire, car je ne trouve moi aussi que ces deux mots. <ref target="https://fr.wikipedia.org/wiki/curry">Curry</ref></p>
 <p>Dans mon petit larousse illustré, on trouve les 3 termes, chiropratique étant le terme canadien authentique.</p>
 <p> Peut-on dire que c'est une profession? N'est-ce pas plutôt une branche de la médecine? </p>
<signed><ref target="https://fr.wikipedia.org/wiki/Utilisateur:guillaume_bokiau">Guillaume Bokiau</ref>21 sep 2003 à 11:49 (CEST)</signed></post>
</div>
```

Figure 3.11: exemple de signalement d'un passage problématique de 4 <post> non corrigé

3.7.4. Exemple 2 de correction facile

La figure 3.12 donne un exemple visuel de discussion mal formée qu'il sera facile à corriger. L'auteur `82.229.70.91` a écrit un long texte (vérifié dans la page historique). Au milieu de cela, viennent s'insérer trois réponses de `Dakdada`. Ces inserts ont eu pour effet de découper

en TEI le texte en quatre parties, dans lequel 3 postes ont des auteurs non identifiés (figure 3.13 en haut). La correction manuelle est indiquer en figure 3.13 en bas.

Ernst qui a été mouillé dans différents scandales mettant en cause son intégrité, publie des reviews en étant plus créatif au niveau des titres chocs que rigoureux au niveau de la méthodologie.Ses productions sont basées sur ses croyances au point de faire dire à ses conclusions le contraire des études qu'il cite.Impossible de le prendre au sérieux si on a un minimum d'esprit scientifique. Ses travaux sont juste instrumentalisés par ceux qui ont envie de cracher sur la chiropratique pour des raisons qui leurs sont propres, mais qui n'ont rien à voir avec les bienfaits de la Science et les progrès de la connaissance.

> *Ad hominem* : stop. — **Dakdada** (discuter) 24 juillet 2012 à 16:49 (CEST)

Par ailleurs, comme le souligne le Pr Gagey, les études scientifiques dans ce domaine sont difficiles à mettre en place .Ce qui explique le manque de parution de très haute qualité(double-aveugle impossible).quand un assureur compare les résultats de traités pour le mal de dos d'1 million de patients sans couverture chiro et 700 000 personnes avec couverture chiro et découvre que

- Chiropractic care cut the cost of treating back pain by 28%.
- Chiropractic care reduced hospitalizations among back pain patients by 41%.
- Chiropractic care reduced back surgeries by 32%.
- Chiropractic care reduced the cost of medical imaging, such as X-rays or MRIs, by 37%.

Il ne fait pas une étude scientifique, il compare l'efficacité économique de deux approches.c'est une approche pragmatique plutôt que scientifique, mais c'est une étude "in vivo".
le fait que la Science n'ait pas produit d'étude "in vitro" capable d'expliquer le phénomène ne signifie donc pas que la chiropratique ne fonctionne pas, mais que à l'heure actuelle la Science n'est pas capable de fournir un modèle adéquat, ça ne veut pas dire que ça ne se fera pas.

> Avant de chercher une explication à un phénomène, il faut le mettre en évidence. Les études d'efficacité scientifiques montrent que c'est une entreprise vaine. L'efficacité économique n'a franchement rien à faire dans ce contexte, merci de nous épargner ça. — **Dakdada** (discuter) 24 juillet 2012 à 16:49 (CEST)

par ailleurs s'acharner à mettre en relief des effets secondaires à la limite de la visibilité statistique quand l'alternative ordinairement usitée les anti-inflammatoires fait des dégâts graves fréquents, c'est juste intellectuellement malhonnête. Et faire croire que "c'est pas mieux qu'un placebo" quand les études indiquent "superior to placebo" c'est juste totalement faux.Si Wikipedia continue dans ce registre, il faut alors que l'article soit intitulé "CROYANCES ET LEGENDES SUR LA CHIROPRATIQUE" mais ne pas prétendre rendre honnêtement compte des faits.

> C'est votre avis, qui n'a pas sa place sur Wikipédia. — **Dakdada** (discuter) 24 juillet 2012 à 16:49 (CEST)

En conclusion Ernst produit de la daube, se cite lui-même ("c'est vrai parce que je l'ai déjà dit") est repris par d'autres qui n'ont pas lu les études qu'il cite de travers et produit un machin m$rdique qu'il baptise pompeusement review, qui flatte vos croyances...l'article Wikipedia n'est déjà pas terrible, ça ne va pas améliorer la crédibilité.mais si vous préférez publier du faux ayant l'apparence de la qualité ,tant pis .c'est wikipedia qui perdra un peu plus en crédibilité.--82.229.70.91 (d) 24 juillet 2012 à 14:22 (CEST)

Figure 3.12 Exemple de discussion mal formée aisée à corriger.

```
Avant la correction
<post xml:id="cmr-wiki-c020-a197" when-custom="unknown" who="psn:cmr-
wikiconflits-p_unknown" n="3" ref="#cmr-wiki-c020-a188">
[…]</post>
<post xml:id="cmr-wiki-c020-a198" when-iso="2012-07-24T16:49" who="psn:cmr-
wikiconflits-p15305" n="4" ref="#cmr-wiki-c020-a197">
[…]</post>
<post xml:id="cmr-wiki-c020-a199" when-custom="unknown" who="psn:cmr-
wikiconflits-p_unknown" n="3" ref="#cmr-wiki-c020-a188">
[…]</post>
<post xml:id="cmr-wiki-c020-a200" when-iso="2012-07-24T16:49" who="psn:cmr-
wikiconflits-p15305" n="4" ref="#cmr-wiki-c020-a199">
[…]</post>
<post xml:id="cmr-wiki-c020-a201" when-iso="2012-07-24T14:22" who="psn:cmr-
wikiconflits-ip01383" n="3" ref="#cmr-wiki-c020-a188">
[…]</post>
```

```
Correction sur les entêtes et ajout d'un <join>
<post xml:id="cmr-wiki-c020-a197" when-iso="2012-07-24T14:22" who=" psn:cmr-
wikiconflits-ip01383" n="3" ref="#cmr-wiki-c020-a188">
<post xml:id="cmr-wiki-c020-a198" when-iso="2012-07-24T16:49" who="psn:cmr-
wikiconflits-p15305" n="4" ref="#cmr-wiki-c020-a197">
<post xml:id="cmr-wiki-c020-a199" when-iso="2012-07-24T14:22" who="psn:cmr-
wikiconflits-ip01383" n="3" ref="#cmr-wiki-c020-a188">
<post xml:id="cmr-wiki-c020-a200" when-iso="2012-07-24T16:49" who="psn:cmr-
wikiconflits-p15305" n="4" ref="#cmr-wiki-c020-a199">
<post xml:id="cmr-wiki-c020-a201" when-iso="2012-07-24T14:22" who="psn:cmr-
wikiconflits-ip01383" n="3" ref="#cmr-wiki-c020-a188">
<join result="post" target="#cmr-wiki-c020-a197 #cmr-wiki-c020-a199 #cmr-wiki-
c020-a201"/>
```

Figure 3-13 Mauvais découpage de <post> et correction

Si plusieurs <join> étaient nécessaires, on pourrait les regrouper avec l'élément <joinGrp>, qui porterait alors l'attribut @result.

```
(3.6)
<!-- join post par groupe -->
<joinGrp result="post">
 <join target="#cmr-wiki-c020-a197 #cmr-wiki-c020-a199 #cmr-wiki-c020-a201"/>
 <join target= …  />
 […]
<joinGrp>
```

3.7.5. Sources permettant d'effectuer les corrections manuelles et résultats

Le travail manuel consiste à vérifier et corriger les contenus des fichiers TEI concernant les discussions dans les sous-répertoires de `wikiconflits-V0 > tei-v1`. Par exemple, la figure 2.14 montre les fichiers TEI pour l'Eolienne, avec deux fichiers de discussion `(<teiheader>` incomplet). Pour en vérifier un, le télécharger en local. Effectuer les corrections, changer son nom (ajout d'un suffixe de version) et charger la nouvelle version en ligne dans le même répertoire. L'ancienne ne change pas.

- `cmr-wikiconflits-eolienne-discu-teiv1.xml` ancien nom du fichier (ne pas modifier)
- `cmr-wikiconflits-eolienne-discu-teiv1_1.xml` nom de la version corrigée.

Pour faire ce travail, qui nécessite de relire les discussions telles qu'elles apparaissent en ligne, le chercheur se reportera

- Essentiellement, aux pages en ligne dans Wikipédia. Ne pas oublier que les pages en ligne ne correspondent pas exactement aux discussions traités dans nos fichiers TEI puisque nous avons pris la peine, comme expliqué précédemment, de rassembler toutes les discussions sur un thème (archives comprises) dans un seul fichier
- Eventuellement, aux historiques stockés dans le répertoire de chaque article de tei-v1.
- En dernier recours, il pourra se reporter au contenu du répertoire `wikiconflits-V0 > depots` expliqué précédemment..

4. Analyse du wikitexte

4.1. Encodage et décodage

Le Wikicode (ou *Wiki Markup*) (Help:Wiki_markup, 2014) est un langage de balisage léger utilisé dans toutes les pages de Wikipédia (articles, discussion, etc.). Avec le langage HTML, ils constituent un type de texte, appelé wikitexte, celui que l'auteur voit en passant en position d'éditeur. Ce dernier est utilisé dans tous les projets de la Wikimedia Foundation (en anglais, le sigle est WMF), y compris Wikipédia.

Les fichiers Dump de Wikipedia sont des fichiers XML bien formés, écrits en wikicode. Tous les caractères illégaux (voir liste suivant) sont convertis en entités XML (terme *escape* en anglais).

Encodage d'entité XML	Décodage d'entité XML
&lt;	<
&gt;	>
&amp;	&

Tableau 4.1

Quand un contributeur rédige un texte sur Wikipédia, il n'a pas à se soucier du fait que le fichier résultant sera un fichier XML bien formé. Cet auteur peut taper du code HTML dans son texte. Le système devra alors le convertir en entités XML dans le Dump. Puis lorsque ce contenu est affiché dans un navigateur (en format HTML), le système effectuera partiellement un travail inverse. Le tableau 3.4, colonne de gauche, montre ce que l'utilisateur a sais dans le wikitexte (ici un mélange de XML et de HTML). Dans la colonne du milieu, le changement opéré par le système pour stocker la page. Dans la colonne de droite, on précise si le balisage de départ correspondant ou non à du HTML (car le système les traitetra ensuite différemment)

Décodage d'entité XML	Encodage d'entité XML	Balise de HTML
<xxx@xxx.com>	<xxx@xxx.com>	Non
title	title	Oui

Tableau 4.2

- **Suite de transformations effectuées par le système Wikipédia**

On a fait un test sur « comment le système Wikipédia peut identifier la balise HTML ». Le résultat a montré que le système n'interprète que la balise HTML. Voici un exemple dans le teste :

D'abord, on a saisi un texte contenant des chevrons, qui en HTML correspond à la balise pour la graisse :

```
<b><xxx@xxx.com></b>
```

Ensuite, on retrouve dans le dump de Wikipédia, le texte converti comme ci-dessous :

```
&lt;b&gt;&lt;xxx@xxx.com&gt;&lt;/b&gt;
```

Après, on vérifie dans le code source HTML du site Wikipédia, ce qu'est devenu ce contenu :

```
<b>&lt;xxx@xxx.com&gt;</b>
```

En final, on voit dans la page du navigateur, le texte est rendu en gras :

<xxx@xxx.com>

Seuls sont donc décodés et encodés les caractères correspondant à des balises HTML.

4.2. Stratégie pour traiter le wikitexte en TEI

L'objectif de traitement était transformer le wikitexte en TEI, c'est-à-dire que tous les types de forme de wikicode et de HTML doivent être convertis en TEI. Etant donné la très grande variété de wikicode et modules inventés par les auteurs, il s'est révélé impossible de traiter toutes les variantes d'utilisation du code pour les traiter automatiquement. Le faire reviendrait à développer seul un système d'une complexité équivalente au Wikicode, résultat du travail de centaines de personnes, pendant des années. Nous avons donc observé les formes / objets les plus utilisé(e)s dans le corpus nous concernant et avons essayé de les traduire en TEI, dans la mesure où ces formes étaient régulières. Les autres ont été laissées en l'état et leurs contenus seront entourés de la balise TEI `<code>`.

4.2.1. Première étape : détection des formes / objets

La détection est faite à l'aide d'un module de Python : mwparserfromhell (The_Earwig, 2014). Ce module repère différentes formes de wikicode et de HTML, y compris le texte dans les chevrons. Il produit un objet Python avec des attributs qui correspondant à chaque type de forme. En voici deux exemples (4.1) et (4.2).

(4.1)
a) `[[Foo|Bar]]` wikitexte de départ. Il s'agit d'un objet « Wikilink » contenant des attributs « text » et « title ».
b) Sortie Python à l'aide du programme mwparserfromhell:

	Objet	*Text*	*Title*
Valeur	Wikilink	Foo	Bar

c) Notre traitement en TEI

<ref target=" https://fr.wikipedia.org/wiki/Foo">Bar</ref>

(4.2)
a) `== Foo ==` wikitexte de départ, correspondant à un objet « Heading » contenant des attributs « level » et « title »
b) Sortie Python à l'aide du programme mwparserfromhell:

	Objet	*Level*	*Title*
Valeur	Heading	2	Foo

c) Notre traitement en TEI

<head>Foo</head>

4.3. Traitement d'objets du wikicode en TEI

Dans cette section, on présente une suite de traitements de conversion sur les objets détectés. Il y a 7 objets différents : `heading, wikilink, external_link, commentaire, module, tag` et `text`.

4.3.1. Heading

Heading contient deux attributs, titre et niveau. Il indique un début de section. l présente donc ces deux informations d'une section. Sa structure est comme cet exemple : `==titre==`, une suite de pairs de symbole `=` est mise autour du texte du titre.

Composant	*Objet*	*Title*	*Niveau*
Valeur	Heading	Titre	2

Le format en TEI :

Type	*Format wikitexte*	*Format TEI*
Heading standard	== Titre ==	<head>Titre</head>

4.3.2. Wikilink

Wikilink présente le lien interne de Wikipédia, on met les doubles crochets autour de texte pour constituer la structure, ex. `[[Foo]]`. Le texte correspond au titre d'une page Wikipédia. Mais d'autres variations sont possibles, comme en témoigne le tableau 4.3.

Type	*Format wikitexte*	*Format TEI*
Lien normal	[[Bonjour]]	<ref target="https://fr.wikipedia.org/wiki/Bonjour">Bonjour</ref>
Lien renommé	[[Bonjour|French Hello]]	<ref target="https://fr.wikipedia.org/wiki/Bonjour">French Hello</ref>
Lien renommé automatique	[[Bonjour(fr)|]]	<ref target="https://fr.wikipedia.org/wiki/Bonjour(fr)">Bonjour</ref>
Lien lie à page d'une autre langue	[[en:Bonjour]]	<ref target="https://en.wikipedia.org/wiki/Bonjour">Bonjour</ref>
Lien lie à un autre projet WMF	[[Wiktionary:Bonjour]	<ref target="https://fr.wiktionary.org/wiki/bonjour">Bonjour</ref>

Tableau 4.3

Le système de conversion n'est pas complet. Ainsi ce wikicode `[[en:bonjour]]` devrait être traduit par :

```
https://en.wikipedia.org/wiki/Bonjour
```

Il sera traduit par notre système ainsi :

```
<ref target="https://fr.wikipedia.org/wiki/en:bonjour">en:bonjour</ref>
```

Cette référence à l'encyclopédie anglaise est parfois notée également ainsi (d'autres variations sont encore possibles):

```
{{en}} [[bonjour]]
```

4.3.3. External_link

Le lien externe est mis dans une paire de crochets pour la structure. Une espace sépare le lien de son titre. Cet « external_link » peut se constituer aussi sans crochet, mais c'est un format non recommandé.

Type	*Format wikitexte*	*Format TEI*
Lien normal	[http://www.wikipedia.org Wikipedia]	<ref target="https://www.wikipedia.org">Wikipedia</ref>
URL Simple (non recommandé)	http://www.wikipedia.org	<ref target="https://www.wikipedia.org">http://www.wikipedia.org</ref>

Tableau 4.4

4.3.4. Commentaire

Commentaire est un type de texte invisible quand le texte est mis en ligne. Il n'existe que dans le wikicode. Son format correspond à celui du XML.

Type	*Format wikitexte*	*Format TEI*
Commentaire standard	<!-- texte invisible -->	~~texte invisible~~

Tableau 4.5

4.3.5. Module

Il existe des milliers de modules dans le système de Wikipédia. A chaque module correspond un traitement particulier assuré par le système Wikipédia (par exemple, affichage de texte d'un format particulier comme les bandeaux dans le fichier visualisé dans le navigateur). Sa structure est du type `{{u|bonjour le monde}}` », avec le nom de module, suivi des paramètres (ici un seul).

Composant	*Objet*	*Nom*	*Param 1*	*Param 2*	*…*	*Param n*
Valeur	Module	u	bonjour le monde			

Tableau 4.6

Du fait qu'il existe des milliers de modules différents, nous avons décidé de ne pas les traiter, mais de faire figurer les contenus correspondant dans la balise `<ode>`

Type	*Format wikitexte*	*Format TEI*
Module	{{u|bonjour le monde}}	<code lang="wikicode"> {{u|bonjour le monde}} </code>

Tableau 4.7

4.3.6. Balises

Ces balises correspondent à du code HTML ou du Wikicode. Compte tenu du grand nombre de balises / étiquetage possibles, seul un petit nombre a été pris en compte (cf. tableau 4.9).

Type	*Format wikitexte*	*Format TEI*
Tag	<u>bonjour</u>	<hi rend="underlining">bonjour</hi>
	bonjour	<hi rend="bold">bonjour</hi>
	bonjour	<hi rend="italic">bonjour</hi>
	''bonjour''	<hi rend="bold">bonjour</hi>
	'''bonjour'''	<hi rend="italic">bonjour</hi>
	bonjour	<emph>bonjour</emph>
	…	…

Tableau 4.9

4.4. Exemple avec fichier TEI simplifié

L'exemple (4.3) donne un extrait simplifié de fichier TEI correspondant à une page de discussion. Dans le `<teiheader>` seuls figurent le titre et un extrait de listes d'auteurs / contributeurs ; dans le corps, un seul sujet de discussion avec un seul message contenant deux paragraphes.

```
(4.3)
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="tei cmr.rng" type="application/xml"
schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Discussion:Chiropratique</title>
         </titleStmt>
         <publicationStmt>
            <p>Publication Information</p>
         </publicationStmt>
         <sourceDesc>
            <listPerson> les contributeurs ne sont plus identifier dans chaque fichier,
voir modif opérées en aout 2014 dans section 3.5
               <person xml:id="cmr-wiki-c020-p0">
                  <persName>
                     <addName>unknown</addName>
                  </persName>
               </person>
               <person xml:id="cmr-wiki-c020-p1">
                  <persName>
                     <addName ref="https://fr.wikipedia.org/wiki/Utilisateur:Nicorama"
                        >Nicorama</addName>
                  </persName>
               </person>
            </listPerson>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <div>
            <head>Controverses</head>
            <post xml:id="cmr-wiki-c020-a1" when-iso="2006-04-29T19:57" who="#cmr-wiki-
c020-p1" n="0">
            <p>Si l'on ne considère a priori la discipline que sous l'angle des
"manipulation vertébrales", on pourrait n'y voir à première vue qu'une forme
particulière de <ref
target="https://fr.wikipedia.org/wiki/Superkinésithérapie">superkinésithérapie</ref>,
mais ce serait là faire preuve de myopie.</p>
               <signed><ref target="https://fr.wikipedia.org/wiki/Utilisateur:nicorama"
>Nicorama</ref> 29 avril 2006 à 19:57</signed>
            </post>
            <post xml:id="cmr-wiki-c020-a2" when-custom="unknown" who="#cmr-wiki-c020-
p0" n="0">
               <p>Cependant les Docteurs en Chiropratique constituent la première
profession de santé manuelle doctorale dans le monde. Dans de nombreux Etats et pays,
ils sont les seuls à être autorisés à pratiquer les soins manuels vertébraux....</p>
            </post>
         </div>
      </body>
   </text>
</TEI>
```

5. Wiki Tool

Wiki Tool, un outil du projet CoMeRe. Il permet d'exporter la page venant de dump de Wikipédia. L'outil, pratique à utiliser, ne requiert pas de connaissance de programmation. Cet outil sera bientôt mis en ligne sous licence libre.

```
Lien de téléchargement : http://sourceforge.net/projects/wikiexporttool
```

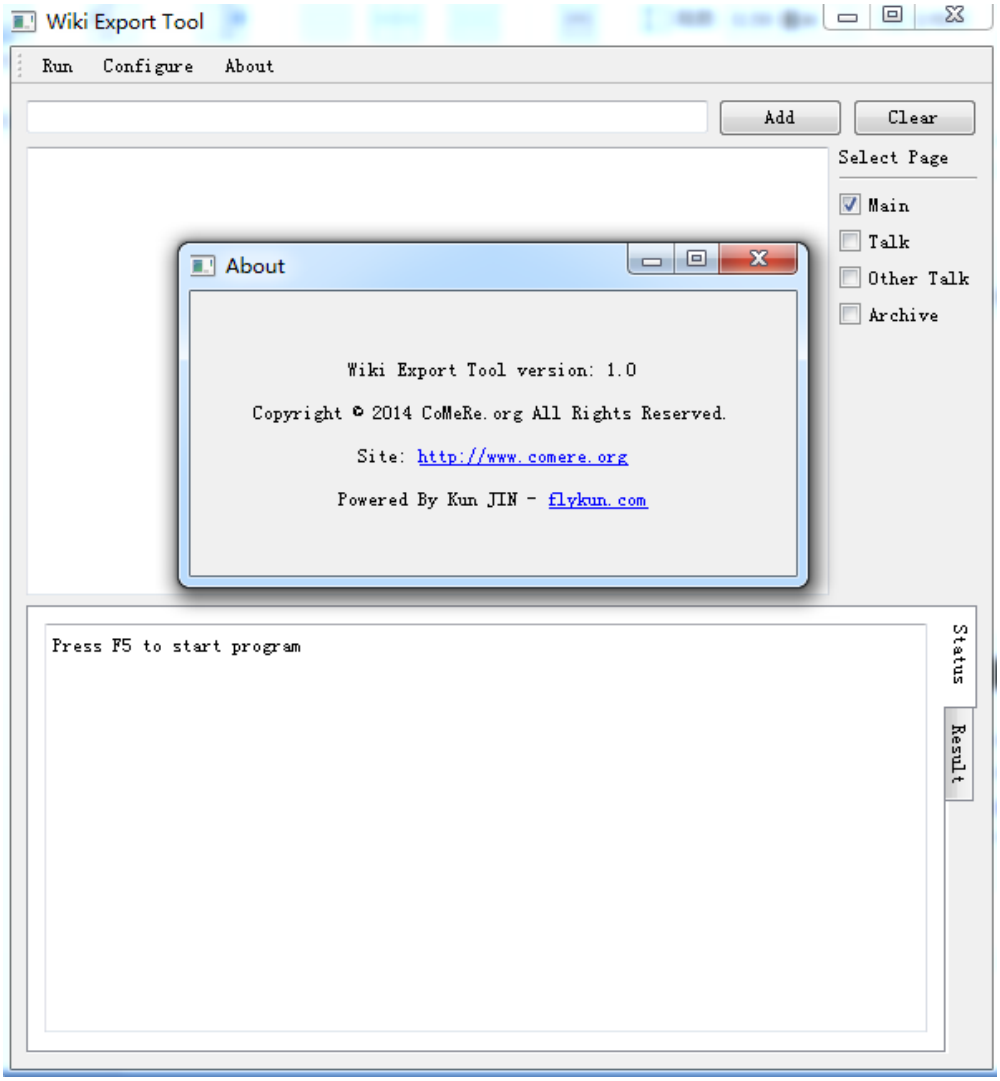

Figure 5-1 Wiki Tool main

5.1. Manuel

Afin d'exporter la page venant de dump de Wikipédia, on a besoin de 5 étapes seulement :

1. Télécharger l'outil avec la version correspondante à votre système ;
2. Préparer le dump de Wikipédia dont la page voulez-vous exporter ;
3. Préparer une liste de noms de pages, tel qu'indiquer dans les figures suivantes :

Figure 5-2 Nom "Chiropratique"

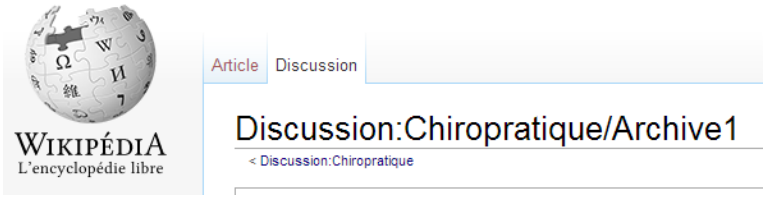

Figure 5-3 Nom "Discussion:Chiropratique/Archive1"

4. Charger le dump (menu« Configure »), qui supporte deux types de dumps : bz2 et xml

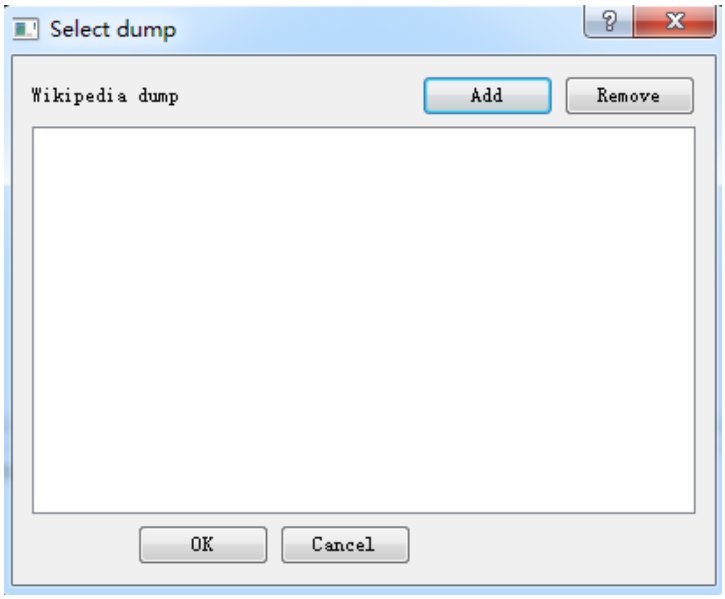

Figure 5-4 Charge de dump

5. Ajouter le nom de la page sélectionnée,

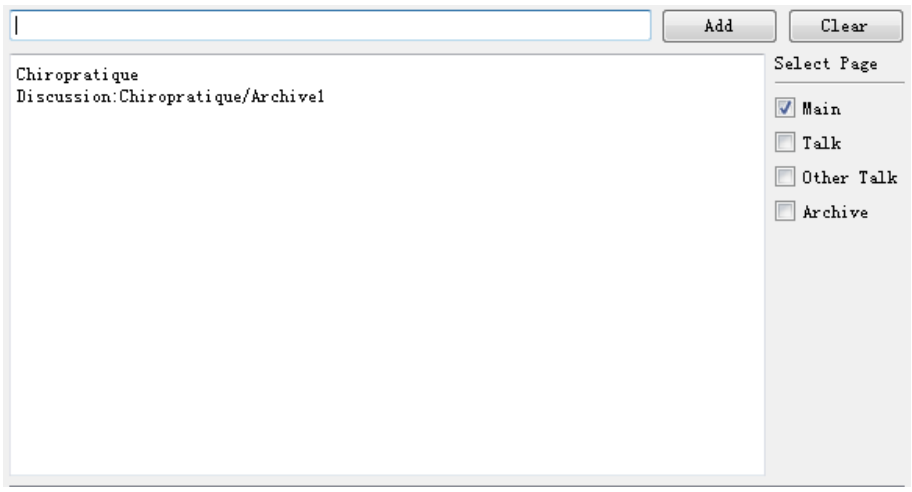

Figure 5-5 Ajouter les noms de page

6. Sélectionner les pages demandées.
   a) Main, n'ajoute pas de préfixe
   b) Talk, ajouter le préfixe « Discussion : »
   c) Other Talk, spécialement pour le Wikipédia français, s'il est croché, l'outil va chercher automatiquement les pages :
      i. Discussion:Chiropratique/Neutralité
      ii. Discussion:Chiropratique/Suppression
      iii. Discussion:Chiropratique/Droit d'auteur
      iv. Discussion:Chiropratique/Article de qualité
      v. Discussion:Chiropratique/Bon article
      vi. Discussion:Chiropratique/Lumière sur
   d) Archive, fonctionne seulement pour les pages comme :
      i. Discussion:Chiropratique/Archive1
      ii. Discussion:Chiropratique/Archive2
      iii. …

Si votre liste de noms de pages est mélangée déjà différentes types de pages, crochez seulement « main », sinon l'outil prendra beaucoup plus de temps pour chercher tous les pages.

7. Appuyer « Run » ou « F5 ».

6. Références

Aide_de_Wikipédia_sur_le_wikitexte. (2014). Wikepedia. 1 (2014). *Aide de Wikipédia sur le wikitexte* : http://fr.wikipedia.org/wiki/Aide:Syntaxe.

Broughton, John (2008). "Who Did What: Page Histories and Reverting - Advanced Techniques". In *Wikipedia: The Missing Manual*.: O'Reilly.

CC_BY_3.0. (2014). *Licence Creative Commons*. http://creativecommons.org/licenses/by/3.0/.

Data_dumps/Dump_format. (2014). Forzat des Dumps. Wikepedia. https://meta.wikimedia.org/wiki/Data_dumps/Dump_format.

Help:Namespace. 2012. Explications sur les noms d'espace. Wikimedia. http://meta.wikimedia.org/wiki/Help:Namespace.

Help:Page_history. (2014). Excplications sur les historiques. Wikipedia. 2014. http://en.wikipedia.org/wiki/Help:Page_history.

Margaretha, E & Lüngen, H. (2014). "Building Linguistic Corpora from Wikipedia Articles and Discussions". [Article soumis] Revue JLCL

RELAX_NG_Home. (2014). Site présentant la structure de documents XML de type RELAX. RELAX NG Home. http://relaxng.org/.

Simple_API_for_XML. (2014). Wikipédia. https://fr.wikipedia.org/wiki/Simple_API_for_XML.

Site_officiel_de_SAX. (2014). Site officiel de SAX. http://www.saxproject.org/.

Site_officiel_d'Oxygen. (2014). Oxygen. http://www.oxygenxml.com/.

TEI-CMC (2014). *Wiki SIG Computer-Mediated Communication*. [document]. Consortium TEI-C. http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication .

Wikipédia:Citation_et_réutilisation_du_contenu_de_Wikipédia. (2014). Wikipédia. 1 2014. http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Citation_et_r%C3%A9utilisation_du_contenu_de_Wikip%C3%A9dia.

Wikipedia:Database_download. 2014. Wikipedia. 3 2014. [Citation : 11 3 2014.] http://en.wikipedia.org/wiki/Wikipedia:Database_download.

Wiki-bandeaux (2014). http://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Bandeau_d%27article

Wiki-blocecrit (2014). http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Blocage_en_%C3%A9criture

Wiki-mediateurs (2014). Liste médiateurs expérimentés : http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:M%C3%A9diation/M%C3%A9diateurs_exp%C3%A9riment%C3%A9s

Wiki-mediation (2014). Définition de l'espace médiation Wikipedia. http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:M%C3%A9diation

Wiki-PRC (2014). Volontaires surveillant les malversations : http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Patrouille_RC

Wiki-protege (2014) http://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Article_prot%C3%A9g%C3%A9

Wiki-salon (2014). http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Le_salon_de_m%C3%A9diation

Wiki-semiprotect (2014). http://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Page_en_semi-protection_longue

Wiki-schéma (2014). Schéma de l'organisation de Wikipédia : http://upload.wikimedia.org/wikipedia/commons/3/36/Mediawiki_database_Schema.svg

Wiki-xsd (2014) schéma XSD de toute la structure Wikipédia : http://www.mediawiki.org/xml/export-0.8.xsd

Using Abbreviated Pointers. (2014). TEI-C: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SAPU

7. Annexes

7.1. Liste de pages à extraire (janvier ou février 2014)

Page	Lieu	Thème	URLs Wikipédia online
Chiropratique	**NPOV**	**Pseudo-sciences**	• **Page article** : http://fr.wikipedia.org/wiki/Chiropratique • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Chiropratique • **Archives de la page discussion (avant juillet 2008) :** http://fr.wikipedia.org/wiki/Discussion:Chiropratique/Archive1 • **Page discussion NPOV** : http://fr.wikipedia.org/wiki/Discussion:Chiropratique/Neutralité • **Page discussion NPOV (débat précédent)** : http://fr.wikipedia.org/wiki/Discussion:Chiropratique/Neutralité/2 • **Archives de la page discussion NPOV (version archivée du 8 mai 2007)** : http://fr.wikipedia.org/w/index.php?title=Chiropratique/Neutralité&oldid=16636952
Igor et Grichka Bogdanoff	**NPOV**	**Personnalités controversées**	• **Page article** : http://fr.wikipedia.org/wiki/Igor_et_Grichka_Bogdanoff • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Igor_et_Grichka_Bogdanoff • **Archives de la page discussion (de 2005 à fin 2008) :** http://fr.wikipedia.org/wiki/Discussion:Igor_et_Grichka_Bogdanoff/Archive_1 • **Archives de la page discussion (de 2009 à fin 2012) :** https://fr.wikipedia.org/wiki/Discussion:Igor_et_Grichka_Bogdanoff/Archive_2 • **Page discussion NPOV** : http://fr.wikipedia.org/wiki/Discussion:Igor_et_Grichka_Bogdanoff/Neutralité • **Page discussion NPOV (débat précédent)** : http://fr.wikipedia.org/wiki/Discussion:Chiropratique/Neutralité/2 • **Archives de la page discussion NPOV (version archivée du 7 décembre 2006)** : http://fr.wikipedia.org/w/index.php?title=Igor_et_Grichka_Bogdanoff/Neutralité&oldid=12364516 **A noter, l'existence d'une autre page portant explicitement sur l'affaire qu'il faudrait également récupérer** • **Autre article « Affaire Bogdanoff » :** http://fr.wikipedia.org/wiki/Polémique_autour_des_travaux_des_frères_Bogdanoff • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Affaire_Bogdanoff • **Archives de la page discussion (archives jusqu'au 21 octobre 2006, « discussions virulentes ») :** http://fr.wikipedia.org/wiki/Discussion:Polémique_autour_des_travaux_des_frères_Bogdanoff/Archive_1

OGM	**Médiation**	**Technosciences, controverse publique**	• **Page article** : http://fr.wikipedia.org/wiki/Organisme_génétiquement_modifié • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié • **Archives de la page discussion (avant juillet 2008, 7 archives) :** ○ **Archive 1 (octobre 2003 → octobre 2007) :** http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié/Archive_1 ○ **Archive 2 (novembre 2007 → 17 janvier 2008) :** http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié/Archive_2 ○ **Archive 3 (17 janvier 2008 → 21 février 2008) :** http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié/Archive_3 ○ **Archive 4 (22 février 2008 → 11 mars 2008) :** http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié/Archive_4 ○ **Archive 5 (15 mars 2008 → 29 avril 2008) :** http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié/Archive_5 ○ **Archive 6 (29 avril 2008 → 10 mai 2011) :** http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié/Archive_6 ○ **Archive 7 (28 septembre 2011 → 4 septembre 2013) :** http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié/Archive_7 • **Page discussion NPOV** : http://fr.wikipedia.org/wiki/Discussion:Organisme_génétiquement_modifié/Neutralité • **Archives de la page discussion NPOV (version archivée du 4 février 2007)** : http://fr.wikipedia.org/w/index.php?title=Organisme_génétiquement_modifié/Neutralité&oldid=13859950 ○ Noter que la page est quasi vide, alors que l'historique témoigne d'une activité intense – les versions précédentes de la page devraient permettre de récupérer les points de débat intéressants. **A noter, l'existence d'une autre page portant explicitement sur l'affaire qu'il faudrait également récupérer (scission en mai 2008, mentionnée sur la page discussion de la page controverse)** • **Page article** : http://fr.wikipedia.org/wiki/Controverse_sur_les_organismes_génétiquement_modifiés • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Débat_sur_les_organismes_génétiquement_modifiés
Quotient intellectuel	**Blocage**	**Méthodologies**	• **Page article** : http://fr.wikipedia.org/wiki/Quotient_intellectuel • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Quotient_intellectuel
Histoire de la logique	**Pertinence**	**Histoire et épistémologie**	• **Page article** : http://fr.wikipedia.org/wiki/Histoire_de_la_logique • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Histoire_de_la_logique
Psychanalyse	**Médiation**	**Légitimité, scientificité, Méthodologies**	• **Page article** : http://fr.wikipedia.org/wiki/Psychanalyse • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Psychanalyse • **Archives de la page discussion (avant février 2013, 3 archives) :** ○ **Archive 1 (avant 2006)** : http://fr.wikipedia.org/wiki/Discussion:Psychanalyse/arch1 ○ **Archive 2 (2006-2008)** : http://fr.wikipedia.org/wiki/Discussion:Psychanalyse/arch2 ○ **Archive 3 (2008-février 2013)** : http://fr.wikipedia.org/wiki/Discussion:Psychanalyse/arch3 • **Page discussion NPOV** : http://fr.wikipedia.org/wiki/Discussion:Psychanalyse/Neutralité
Eolienne	**Pertinence**	**Technosciences**	• **Page article** : http://fr.wikipedia.org/wiki/Éolienne • **Page discussion** : http://fr.wikipedia.org/wiki/Discussion:Éolienne • **Page discussion NPOV** : http://fr.wikipedia.org/wiki/Discussion:Éolienne/Neutralité