Wikipédia, objet de recherches : entre observations, expérimentations et co-constructions

Evelyne Broudoux

Evelyne Broudoux
DICEN, Cnam

Biographie :
Evelyne Broudoux est maitre de conférences en Sciences de l'information et de la communication au CNAM-Paris et membre de l'équipe de recherche DICEN. Ses travaux de recherche portent sur l'évolution des pratiques auctoriales, éditoriales et des mesures d'autorité à l'heure du web et du big data, dans les champs intéressés par la création et l'innovation.

Wikipédia, objet de recherches : entre observations, expérimentations et co-constructions

Introduction

L'exemple vivant d'agencement collaboratif de connaissances et de savoirs que constitue Wikipédia intéresse les chercheurs de nombreux domaines. Leur participation à l'organisation des savoirs de l'encyclopédie et leurs apports en contenus, les observations émises sur son fonctionnement éditorial, leur utilisation des systèmes de structuration pour tester des algorithmes aussi bien que des modèles façonnent des points d'accroche que nous nous proposons de décrire ici. Cet état de l'art de la littérature scientifique existante fournit ainsi une photographie des recherches sur l'encyclopédie à partir desquelles il est possible d'inférer des développements futurs.

Méthodologie

Une revue manuelle de la littérature scientifique produite autour de Wikipédia est-elle encore aujourd'hui possible ? En février 2012, Nicolas Jullien relevait 7029 articles citant l'encyclopédie dans la base Science Direct ; en janvier 2014 ce n'est pas loin du double qui est proposé dans la même base (12 488 articles) en effectuant une recherche simple, incluant le « *fulltext* ». Ce chiffre est *a priori* à pondérer puisqu'il n'en reste plus que 197 sur Science Direct si on sélectionne seulement les articles possédant dans leur titre, résumé ou mots-clés, le terme « wikipedia ».

Afin de ne retenir que les articles dont la thématique principale a trait à l'encyclopédie nous avons restreint pour une première communication[1] la recherche portant sur le seul terme « wikipedia » dans le champ « résumé », en anglais et en français et pris en compte les résultats de trois années de publication (2010 à 2012) à partir de différentes sources : 63 articles issus de revues mais aussi de conférences publiées ont ainsi été sélectionnés à partir de 11 articles sur Cairn, 38 articles sur Esmerald, 138 articles sur Base Francis, 5 articles sur ArchiveSic et 42 notices sur Mendeley.
Pour les besoins de ce chapitre nous avons ajouté quelques articles de 2009 et 2013 afin de compléter l'état de l'art mais présentons essentiellement une concentration de références de 2010-2011 issues des sciences et technologies de l'information et de la communication, période très riche en publication d'analyses et de résultats.

[1] http://www.iscc.cnrs.fr/spip.php?article1738

Deux principales directions de recherche

Un premier tri des références thésaurisées, a permis de déceler deux directions principales de recherches empiriques, expérimentales et théoriques dans les orientations prises par les articles publiés dans et autour de Wikipédia, si l'on ôte les mentions de l'encyclopédie destinées à apporter un éclairage rapide définitoire au cours d'un article.

La première émane des sciences et technologies de l'information et consiste à se servir des corpus engrangés sur Wikipédia comme matière première pour expérimenter des théories et des algorithmes liés au traitement automatique des langues, à la recherche d'information et à l'organisation des connaissances. Pour des raisons de périmètre nous ne mentionnerons ici que les principales branches de cette direction, bien que celle-ci mériterait à elle seule des études approfondies.
La seconde émane des sciences humaines et sociales et comporte l'observation qualitative et quantitative du phénomène éditorial à partir duquel l'encyclopédie se développe.
Du recoupement de ses deux directions émerge une troisième constituée par une approche qui regroupe des chercheurs plus spécifiquement investis dans des projets s'intéressant aux communautés open source collaboratives et intéressés par le phénomène Wikipédia en tant qu'exemple vivant de système d'information ouvert organisant des connaissances encyclopédiques créées de manière collaborative. C'est peut-être aussi la découverte principale de cet état de l'art que de remarquer que les chercheurs ayant pris le temps d'examiner l'encyclopédie sous de multiples facettes émettent des observations visant à la compréhension et au soutien des interactions entre collaborateurs aussi bien qu'à la conception de nouveaux systèmes d'organisation des connaissances.

Nous passerons sur la compréhension globale du phénomène qui impliquerait de replacer l'apparition et le développement de l'encyclopédie dans le contexte de la numérisation systématique du « monde » symbolique humain au début du XXIe siècle et ses conséquences : le développement de nouveaux systèmes de liaison de données et leur sémantisation, le web en tant que médium global générateur d'innovations, les systèmes d'enregistrement et de contrôle œuvrant dans l'ignorance des populations, dont les moindres faits et gestes sont scrutés en ligne. Cependant, une réflexion se fait jour prenant Wikipédia comme une réalisation concrète d'un projet de « biens communs », illustrée par une étude exploratoire des publications sur l'encyclopédie (Jullien, 2010).

Wikipédia, objet et terrain d'expérimentations

L'encyclopédie se révèle un terreau fertile pour expérimenter des algorithmes et nous relevons plusieurs directions dans cette approche expérimentale de « terrain ». Les TAL sont les premières à se servir des corpus engrangés sur Wikipédia et de ses systèmes de classification pour effectuer des essais. Ainsi, une liste de termes définis comme ambigus par Wikipédia et disposant d'au moins cinq sous-thèmes sera utilisée pour tester le clustering sémantique d'un moteur de recherche sur mobile (Carpineto et al., 2009) ; l'utilisation de 160 « ontologies sociales » construites manuellement sous Wikipédia servira à la vérification de l'hypothèse de Sapir-Whorf pour la classification des langues

selon une approche théorique réseaux (Mehler et al., 2010) ; la détection d'opinion dans la construction des requêtes à facettes s'appuiera sur des concepts sélectionnés dans Wikipédia (Vechtomova, 2010) ; des méthodes de détection de réponses aux questions se serviront également de l'encyclopédie pour leurs essais (Grappy & Grau, 2011). Les exemples sont nombreux montrant que l'encyclopédie construit le plus vaste réseau sémantique organisant des connaissances à ce jour. A sa base sont les articles décrivant des concepts pris dans une structure de liens (catégories, hyperliens, redirection, etc.) vus comme des relations que l'on pourra spécifier et à partir desquelles des calculs pourront être effectués.

- Recherche d'informations

Les archives de l'encyclopédie étant téléchargeables, des études quantitatives s'appuient sur les « dumps » de Wikipédia pour extraire des corpus et réaliser des tests comme par exemple l'automatisation de résumés afin d'évaluer la performance d'une méthode purement statistique sur des données autres que celles constituées par les actualités (Bossard & Guimier, 2012). Les études approfondies en recherche d'informations (*information retrieval*) nécessitent des jeux de données en nombre massif pour exécuter des tests, comme celles cherchant à mesurer la taille des sources de données du web profond et qui se servent outre des données gouvernementales, de celles de Reuters et d'1,4 millions de documents du Wikipédia anglophone (Lu & Li, 2010). Toujours dans le domaine de la recherche d'informations, et plus particulièrement dans l'approche « filtrage collaboratif » des moteurs de recommandation, l'encyclopédie est envisagée comme un graphe dirigé connectant des concepts (les catégories de Wikipédia) mais dont les thèmes couverts par les articles sont reliés par des hyperliens possédant des relations sémantiques comme « équivalent à », associatives ou hiérarchiques : (Lee et al. 2011) appliquent ainsi différents modèles de filtrage collaboratif en utilisant des connaissances classées par domaine selon ce principe.
Les corpus téléchargés sont ensuite modifiés et adaptés : leur XMLisation en articles, sections et sous-sections va être utilisée pour tester le classement des entités dans les systèmes de question-réponse (Pehcevski, 2010) ou bien détecter les préférences des usagers en termes de stratégies de recherche d'informations (tâches de collecte d'informations et de recherche d'établissement de faits) (Pharo & Krahn, 2011). La versionnalisation des articles récupérable par l'historique des révisions permet à (Nunes et al., 2011) d'expérimenter de nouvelles mesures pour l'évaluation du poids des termes d'un document.

- Construction de connaissances et textmining

L'extraction de connaissances à partir de l'encyclopédie est un objectif suivi par de nombreux chercheurs informaticiens qui expérimentent de nouveaux modèles et c'est sans doute la recherche liée à l'information sémantique qui en tire immédiatement le plus de bénéfices, comme le prouvent de nombreux articles qui vont de la mesure du traitement lacunaire de l'information (Nadamoto and al., 2010) à la mesure de parenté sémantique basée sur l'analyse de co-occurrences de liens pour la construction de thésaurus (Ito and al, 2011). L'article de Nadamoto and al. (2010) vise à créer un moteur de recherche Web2.0 capable de proposer des informations inattendues à des usagers d'un réseau social, à partir de « trous de connaissances » détectés par comparaison avec Wikipédia. D'autres construisent, à partir des contenus, des systèmes de repérage

comme cette ontologie géographique disponible en six langues bâtie par (Ngo et al., 2012) à partir de la hiérarchie Wikipédia, en partant des divisions géographiques les plus grandes (les continents…) jusqu'aux plus petites (les hameaux…). Autant d'applications destinées à extraire des informations sur Wikipédia et à les utiliser comme jeux de données dans le développement d'outils de *textmining*.

Mais c'est l'emblématique projet DBpedia qui représente le mieux cette direction en tant qu'essai le plus abouti faisant converger les praticiens des systèmes d'organisation des connaissances avec ceux du web sémantique et des données liées. Au préalable, les « infoboxes[2] » de Wikipédia auront ouvert la voie en tant que première tentative de structuration des informations, avec des modèles à compléter utilisant les entités nommées et autres métadonnées ; leur extraction par moissonnage étant possible avec des algorithmes dédiés. Le projet DBpedia est d'extraire l'information structurée et de la transformer en une base de connaissances enrichie et interopérable qui pourra être interrogée ensuite par des requêtes complexes. (Morsey et al., 2012) présentent ainsi les dernières avancées du projet dont nous retiendrons trois objectifs : dépasser l'absence de coordination dans la gestion du modèle de l'infobox et ses appropriations particulières (ex : *birthplace* et placeofbirth ou *infobox_city* et *infobox_town*), produire des extractions des données structurées à la demande (actuellement dépendantes des archives mensuelles de Wikipédia) et faire le lien vers les communautés non informaticiennes, en particulier celle des bibliothécaires, en proposant outre l'interconnexion des bases de données en RDF, l'utilisation d'identifiants uniques et pérennes et la participation à des projets de développement d'outils comme l'exploration du web de données pour les usagers.

Le processus éditorial

Le déroulement de la production des contenus, son encadrement par les wikipédiens, son soutien par des procédures automatisées, autant d'intérêts suscités par un phénomène éditorial qui ne dévoile pas facilement ses règles. Car la singularité de l'encyclopédie se trouve effectivement dans l'internalisation de ses modalités d'organisation. Si le système d'amélioration qualitative des contenus s'appuie bien sur des règles aménageant les activités d'écriture, celles-ci sont toujours négociées en interne et elles ne se réfèrent pas à des conventions collectives externes et on peut donc considérer le cadre éditorial de Wikipédia comme unique comparé aux cadres éditoriaux d'autres encyclopédies.
Une première série d'articles est consacrée à la description de cet échafaudage éditorial et en français on retiendra le dossier dirigé par D. Cardon du numéro 143 de la revue *Réseaux* et l'article de B. Jacquemin dans *Document Numérique*, sur la mise en place d'une gouvernance dans un dispositif collaboratif tel que Wikipédia, en 2011.

La première caractéristique visible pour le lecteur consultant l'encyclopédie en ligne est la double-face que constitue une « page » wiki : d'un côté les contenus publiés, de l'autre les discussions qui les concernent. Cette forme sociotechnique qui associe systématiquement la publication à la production ne se rencontre pas dans une chaîne éditoriale traditionnelle où différents acteurs se succèdent dans des rôles prédéfinis.

[2] http://fr.wikipedia.org/wiki/Aide:Infobox

La seconde caractéristique est invisible au lecteur, il s'agit de l'agencement organisationnel de l'encyclopédie dont la première brique est constituée par les cinq principes de base[3] fondant l'encyclopédie auxquels les « éditeurs » de contenus sont tenus de se référer :

- La « pertinence encyclopédique » est de rigueur : chaque article doit avoir un sujet précis, présenter la synthèse des connaissances dudit sujet et respecter les critères d'admissibilité des articles définis par la communauté.
- Les articles doivent satisfaire à la « neutralité des points de vue » : les informations doivent être vérifiables, les sources citées et les contenus ne doivent pas comporter des opinions générales masquant des partis pris idéologiques ou des flous masquant des imprécisions.
- Les contenus doivent être libres, publiés sous « licences GFDL/CC-BY-SA 3.0 » et leurs auteurs ne peuvent y exercer leurs droits ; toutefois des exceptions au droit d'auteur sont tolérées comme l'insertion d'images de monnaies ou de logos identifiant des marques.
- Le « savoir-vivre » est exigé : un code de bonne conduite explicite les principes à respecter, la cordialité doit perdurer pendant les conflits liés aux « guerres d'édition » de manière à éviter les « conflits de personnes ». La bonne foi doit être d'emblée supposée et les attaques personnelles évitées. L'accessibilité des contenus vise le plus grand nombre s'en toutefois s'adresser à des enfants. La règle des trois révocations (recommandation française) permet de discuter au lieu de révoquer. Les règles de blocage en écriture et de déblocage font aussi l'objet de définitions très précises.
- Enfin, le dernier principe, la « souplesse des règles » permet à chacun de conserver une grande liberté sur les contenus et sollicite l'audace des nouveaux entrants écrivant dans l'encyclopédie.

Cette division fonctionnelle de la page wiki nous conduit à traiter les articles en deux parties : ceux qui vont se concentrer autour des contenus publiés – leur qualité, leur typologie, leur crédibilité – et ceux qui vont s'intéresser aux étages du dispositif éditorial ainsi qu'aux différentes couches constituant le millefeuilles de la participation-collaboration jusqu'à la gouvernance globale de l'encyclopédie.

– Les contenus

Il existe deux temps dans les articles s'interrogeant sur la qualité des contenus produits : ceux antérieurs à 2008-10 vont s'interroger sur leur exactitude et leur fiabilité avec des études comparatives : autres encyclopédies, consistance du traitement des sujets entre Wikipédia et d'autres sources, services de Questions-Réponses en bibliothèques (service Oracle pour Wikipédia) et des recherches sur l'évaluation autoritative des sources et la crédibilité des contributeurs.

A partir de 2009-10, les études vont porter sur les différences des contenus entre les langues et les domaines de connaissance. Les représentations historiques alternatives et dominantes portées dans Wikipédia sont étudiées par (Luyt, 2011) qui compare le traitement de l'histoire de Singapour et celle des Philippines et repère pour ces

[3] http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Principes_fondateurs

dernières une tendance à s'écarter du modèle historiographique dominant. Le même auteur (Luyt & Tan, 2010) avait réalisé une étude basée sur la vérification des références et citations dans les articles traitant de l'histoire des pays sur une page de 2008 (50 sélectionnés aléatoirement sur les 249 pays représentés), comparée avec celles du *Journal of World History*, revue trimestrielle évaluée par les pairs depuis 1900. Après avoir fait le constat d'une dominante de quelques sources gouvernementales et médias d'informations généralistes et la rareté de sources académiques, l'auteur remarquait que les connaissances incluses n'étaient pas prouvées par les références externes choisies. Les difficultés d'accès - payant pour les bases de données ou interface-barrage pour les informations détenues par les bibliothèques – expliquent en partie ce manque de références à des savoirs expertisés. Cependant l'auteur reconnaît la nécessité d'une éducation à l'*information literacy* pour encourager la recherche d'informations fiables en même temps qu'inciter à la compréhension de la relativité contextuelle et sociale de la construction des connaissances.

Dans le domaine de la santé, l'évaluation des sources d'information est étudiée à travers le traitement de la controverse sur le dépistage du cancer du sein par (Hjørland, 2011) dans les encyclopédies Britannica et Wikipédia anglaises et danoises ainsi que l'Encyclopédie nationale danoise, la Wikipédia anglaise apparaissant la meilleure des quatre sources selon Hjørland. Un sociologue théorise une typologie de l'ignorance à partir du traitement de son livre sur les professions par un article de Wikipédia et 128 citations issus d'articles scientifiques, en considérant l'ignorance de l'amateur et de l'expert et en les croisant avec l'ignorance des faits, de la littérature et des compétences (Abbott, 2010).
Les différences de traitement des personnages célèbres en anglais et en polonais apportent un éclairage intéressant sur l'observation du fameux point de vue « neutre » de Wikipédia. (Callahan & Herring, 2011) se sont ainsi penchées sur les « biais » culturels reflétés par 60 entrées Wikipédia écrites dans la langue de deux pays, en examinant le traitement de 15 personnages américains en anglais et en polonais et le traitement de 15 personnages polonais en anglais et en polonais, dans des domaines en relation avec la célébrité (sports, politique, musique, cinéma, sciences ou religion) et en établissant des comparaisons sur le ton et les types d'information couverts (dont les controverses). L'étude se termine sur le fait que le terme « biais » sous-entendrait une intentionnalité dans l'objectif de convaincre ou de transformer qui ne correspond pas aux réalités observées. La croyance en une homogénéisation des contenus par des traductions automatiques risquerait de conduire à une domination du Wikipédia anglais qui servirait de modèle aux autres pour ce qui devrait être copié dans d'autres langues.

- Procédures « en étages »

L'organisation du travail d'écriture, d'édition, de vérification, de sauvegarde de l'encyclopédie est basée sur des processus formalisés qui vont de la définition précise des tâches à assumer par les personnes qui l'animent à l'établissement d'un ensemble de procédures destinées à encadrer et assurer la pérennité des apports.
Un ensemble de processus de régulation se déroulant dans des espaces plus ou moins formels de discussion (onglet discussion des pages, bistros) et des structures spécifiques (salon de médiation, comité d'arbitrage) construit la gouvernance de l'encyclopédie.

Bien que non explicite pour les nouveaux entrants contributeurs à l'encyclopédie, la gouvernance destinée à réguler les échanges sociaux liés à la confrontation des idées et des points de vue se révèle à travers les espaces dédiés à la discussion, dont le Comité d'arbitrage constitue le recours ultime. Ses membres élus sont administrateurs, bureaucrates ou bien simples wikipédiens. L'analyse de (Jacquemin, 2011) du processus d'élection du Comité d'arbitrage et de la représentation des administrateurs au sein de ce Comité indique qu'un nombre restreint de contributeurs concentre les pouvoirs les plus grands sur l'encyclopédie.
La participation à Wikipédia associe très étroitement les tâches de production de contenus à celles de discussion à leur sujet. Les phases de coordination des participants à l'encyclopédie entrant pour une bonne part dans leur motivation à investir les différentes tâches autour de l'édition, une vigilance s'exerce ainsi sur le bien collectif selon (Cardon et Levrel, 2009), à travers une régulation qui s'effectue à plusieurs étages.
Au premier niveau, la quasi-totalité des débats d'édition comporte une grammaire des fautes procédurales comprenant :

- Les fautes de délimitation relatives au périmètre des articles et aux thèmes associés
- Les fautes de composition relatives à la structure interne des articles
- Les fautes de *sourcing* qui rassemblent l'insuffisance du pointage vers des sources externes ou dénoncent le caractère original des énoncés proposés
- Les fautes d'équilibrage ou de neutralité liées au respect du NPOV et à la polyphonie des propositions
- Les fautes de civilité concernant les manquements aux principes de convivialité des discussions et à l'incapacité d'écouter les autres.

La typologie des modifications sur Wikipédia de (Dutrey et al., 2011) distingue les corrections des reformulations à faible ou à forte variation sémantique et autorise une différenciation entre correction factuelle et vandalisme. Le *trolling* en tant qu'activité est par ailleurs étudié avec la typologie de (Schachaf & Hara, 2010) qui rapproche le comportement de troll de celui de certains hackers. Afin de juguler le vandalisme dès les premières modifications de pages, un corps de patrouilleurs (RC) formé de volontaires est chargé de détecter les ajouts malveillants.

Au deuxième niveau, la capacité d'exprimer localement sa vigilance en alertant la communauté est sollicitée. Lorsque le consensus n'est pas atteint dans les pages discussion, des procédures d'alerte permettent de déplacer la controverse dans une page dédiée spécifiquement à la médiation. Un bandeau affecté à une page signale un des trois cas : les « pages à Supprimer » (PaS), les « désaccords de neutralité » et les « pages de feu ». Les débats sont internalisés dans le cas des PaS car seuls sont autorisés à donner leur avis ceux qui totalisent un certain nombre de contributions. De plus, seuls les administrateurs sont habilités à détruire les pages, il leur appartient donc de vérifier la validité des arguments exprimés. Les bandeaux d'alerte à la neutralité renvoient quant à eux à un ensemble d'exemples et de traitement de cas apparaissant sur la liste des articles non neutres (LANN).. Les « pages de feu » sont destinées à mettre fin aux de « guerres d'édition » qui sont repérables par des révocations mutuelles de modifications de pages entre deux ou plusieurs contributeurs. Un corps de « pompiers » formé à la médiation intervient en vérifiant essentiellement la validité procédurale des productions de connaissances sur Wikipédia sans prendre parti sur le fond du différend. Les Wikipompiers peuvent proposer des solutions soumises au vote et qui doivent être respectées ensuite sous peine de sanctions.

Le troisième niveau est constitué par le corps des administrateurs et le Comité d'Arbitrage. Les administrateurs sont habilités à protéger ou détruire une page, à bloquer des utilisateurs et interdire provisoirement des adresses IP. Les élus du Comité d'Arbitrage sont amenés à siéger en rassemblant un plaignant, un accusé et sept arbitres (parmi les 10 élus sur mandats de la wikipédia francophone). La plainte doit répondre à trois impératifs : la plainte concerne un autre contributeur, la nature du conflit doit être exprimée de manière concise, elle doit être également documentée et les preuves doivent être constituées par des hyperliens. Encore une fois, il s'agit de juger les discussions sur la forme et non sur leur fond et de rechercher une solution équitable au conflit en déterminant qui a eu tort.
Des cartographies sociales des conflits et querelles indicatrices de la négociation des points de vue montrent qu'une « compétence relationnelle » se constitue avec l'expérience au sein de Wikipédia et que l'élimination des membres qui en sont dépourvus constitue une part non négligeable du processus de régulation (Auray et al., 2009).

– Différences dans la participation

Wikipédia est vu par différents chercheurs comme un laboratoire testant la construction collaborative de savoirs et différents modèles de la participation sont explorés comme le modèle théorique du comportement de partage de connaissances. (Cho and al., 2010) sont parvenus à tester un modèle théorique intégratif de partage des connaissances à partir de l'interrogation de 223 wikipédiens. Leur étude examine les relations entre les motivations, croyances cognitives internes, facteurs relationnels sociaux et les intentions de partage de connaissances, selon la théorie du comportement planifié. L'analyse révèle que les attitudes, le sentiment d'auto-efficacité qui concerne le savoir, et une norme de réciprocité ont des relations directes et signifiantes avec les intentions de partage de connaissances. L'altruisme – en tant que motivation intrinsèque – se trouve relié positivement aux attitudes de partage du savoir, tandis que la réputation – en tant que motivation extrinsèque – ne se révèle pas un prédicteur signifiant de l'attitude. L'étude observe également que le facteur socio-relationnel du sentiment d'appartenance est relié aux intentions de partage de connaissances à travers différents facteurs motivationnels et sociaux comme l'altruisme, les normes subjectives, l'auto-efficacité et la réciprocité généralisée.

Des différences culturelles significatives sont révélées en croisant les attitudes comportementales et les situations géographiques. (Asadi S. et al, 2013) dans une étude auprès de 100 wikipédiens persans relèvent des différences notoires entre des résultats publiés concernant wikipédia en langue anglaise et leurs propres études en langue perse, sur les motivations de contribution et facteurs de découragements dans la participation à long terme.
L'analyse conduite par (Hara and al., 2009) a consisté à assimiler Wikipédia à des communautés de pratiques et a tenté d'en observer leurs différences culturelles à partir de quatre langues différentes, en termes de taille et de culture : « ouest » avec l'anglais et l'hébreu et « est » avec le japonais et le malais. Le corpus observé sur plusieurs années était constitué de 120 pages-discussions extraites aléatoirement dans les quatre langues, qui se différenciaient entre « discussion attachée à une page », « discussion attachée à un

usager » et « discussion attachée à une règle de wikipédia ». L'objectif était de repérer des normes comportementales dans ces différents types de pages-discussions à travers trois catégories rassemblant des actions précises encodées : normes d'écriture, partage de l'information et bien-être de la communauté.
Les similarités et les différences entre cultures et tailles trouvées ont été discutées. On n'en retiendra ici que deux principales, concernant l'art de communiquer dans la courtoisie (usage de la politesse) et les comportements attachés aux conflits et désaccords.

- Les comportements de politesse ont été observés sur les pages « discussions-usagers ». Ils sont plus fréquents sur les wikipédias de large taille plutôt que sur les petits mais aussi plus sur les wikipédias de l'« est » que sur les wikipédias de l'« ouest ».
- Plus la communauté est petite plus les désaccords sont visibles. Une augmentation notable des conflits est évidente au niveau des discussions attachées au « contenu des pages » pour les communautés de l'ouest, alors que peu de différences existent entre l'est et l'ouest pour les « discussions-usagers » et les « discussions-règles ».

En conclusion, c'est une véritable grille d'indicateurs de normes de comportement que les auteurs ont construite, applicable aux communautés de pratique partageant des connaissances à l'extérieur de limites organisationnelles.

La spécialisation renforcée des procédures wikipédiennes a conduit à la formation de rôles adoptés par les wikipédiens mais aussi à l'automatisation des tâches sous forme de scripts qui assurent des activités de surveillance à l'égard de l'apport des contenus. Une série d'articles en remarque la formalisation progressive et croissante. Halfaker et al. étudient depuis 2009 les raisons pour lesquelles les nouveaux éditeurs se découragent dans leur apprentissage et cessent rapidement de tenter d'alimenter l'encyclopédie. En 2013 leur dernier article fait le constat de l'incapacité depuis 2007 pour l'encyclopédie de retenir les éditeurs de bonne foi, au moment où confrontée à l'arrivée massive de nouveaux entrants, celle-ci s'est dotée d'outils semi-automatisés destinés à détecter le vandalisme ou les erreurs de débutants. Le taux de retour automatisé aux versions précédentes produit un effet négatif car les corrections ne sont plus discutées avec les nouveaux éditeurs et ceux-ci se retrouvent de fait éliminés du processus global de gouvernance de l'encyclopédie. L'effet de « légitimation de la participation périphérique » censé intégrer progressivement les nouveaux entrants est court-circuité par les automatismes qui empêchent l'observation directe des pratiques des wikipédiens confirmés. La formalisation accrue de l'approbation de nouvelles politiques produit un effet négatif aussi bien sur les nouveaux entrants que chez les éditeurs expérimentés qui tendent à privilégier les espaces d'expression moins formels pour engager de véritables discussions. Aussi les conséquences de la transformation progressive des recommandations d'écriture en injonctions étayées par des procédures bureaucratiques de vérification sont à évaluer. La « collaboration » ne risque-t-elle pas de muer en simple « participation » ?

Le recul permis par plus de dix années d'expérience éditoriale wikipédienne a entrainé une série d'analyses concernant les points suivants :

- Les travaux autour du « point de vue neutre » (NPOV) et le rôle joué par l'ambiguïté dans l'appropriation du concept et dans l'avancement des travaux d'écriture. L'ambiguïté joue un rôle clé dans la manière dont les usagers interprètent et appliquent la politique de neutralité (Matei & Dobrescu, 2011).

- Les normes sociales contrebalancées par les différences individuelles ont été mises en évidence dans l'intention de contribuer, en particulier par le wikipédia de la Corée du Sud, où la pression des normes comportementales collectives est reconnue par les auteurs ; l'engagement volontaire et le sentiment d'auto-efficacité étant prédictifs de l'apport des contenus, dans le cas d'une enquête auprès d'étudiants (Park et al., 2012).
- Rappelons les conclusions de (Sundin, 2011) : Wikipédia n'est pas un « média alternatif », il s'inscrit bien dans les « médias établis » en procurant des savoirs de « seconde main » (Wilson, 1983), pointés par les ressources externes légitimant les contenus. Et c'est cette faculté à trouver les sources pertinentes justifiant les contenus publiés qui fait la force des « gardiens » de la vérifiabilité que sont les éditeurs. Cette responsabilité est d'autant plus importante que la stabilité des savoirs diffusés par l'encyclopédie en dépend.
- Les rapports entre éducation et wikipédia. Des initiatives multiples existent régulièrement suscitées ou encouragées par l'encyclopédie. Dans une approche interactionnelle de l'enseignement, (Ollivier C., 2010) a réalisé une expérience participative d'écriture dans l'encyclopédie en ligne avec des étudiants de FLE[4]. L'expérience a montré que le fait de proposer des tâches prenant place au sein d'interactions sociales non simulées conduisait à un recadrage du groupe apprenants-enseignant : l'enseignant étant considéré par l'apprenant comme une personne-experte l'aidant à réaliser une tâche fixée. Cette mise en œuvre de tâches ancrées dans la vie réelle fournissait aussi aux apprenants l'occasion de co-agir avec des personnes différentes, dans des relations interpersonnelles variées et réelles dépassant celles du groupe classe. Les apprenants ayant adapté leur discours et leur exigence de qualité à l'interaction sociale dans laquelle ils s'étaient engagés, la confrontation au réel a contribué à augmenter sensiblement leur motivation. *« Participer à un projet collaboratif tel que Wikipédia revient ainsi à valoriser les connaissances et compétences du sujet et conduit à ne pas le réduire au rôle d'apprenant ».*

Wikipédia co-construit par les chercheurs ?

Des contributions scientifiques à l'étude de Wikipédia dont l'extraordinaire variété n'a été qu'effleurée ici, nous retiendrons ceci : Wikipédia est un laboratoire pour les expérimentations dans au moins deux domaines, celui des données liées accessibles par le web et celui de la construction collective de savoirs.

– Données liées et structuration sémantique

Dans leur article sur l'extraction de données structurées de Wikipédia en temps réel, les auteurs du projet DBpedia lancent un appel direct aux bibliothécaires : *« How can librarians as well as DBpedia and Linked Data benefit from each other ? ».* Il s'agit bien de

[4] Français Langue Etrangère.

co-construire le web de données en proposant aux bibliothèques des adresses pérennes pour ajouter des informations à leurs ressources, proposer des extractions de données concernant livres, éditeurs, etc., fournir des infrastructures pour le web de données et des technologies sémantiques pour les chercheurs. Cette offre de collaboration qui se matérialise sous forme de projets en cours et à venir est représentative de la participation de l'encyclopédie à de multiples recherches : par ses corpus librement téléchargeables, Wikipédia soutient les expérimentations algorithmiques, la recherche d'informations (*information retrieval*), la structuration sémantique des données et bien d'autres domaines que nous n'avons pu explorer ici. En retour, les chercheurs quant à eux procurent des « conseils », des « propositions l'amélioration » à la fin de leurs articles, comme si l'ouverture de Wikipédia aux chercheurs (via les API pour les programmes, les archives pour les contenus, et la simple possibilité de participer) entrainait une forme de responsabilité voir d'engagement vis-à-vis de l'encyclopédie.
(Xiao & Askin, 2011) n'ont pas hésité pas à comparer le modèle de publication de l'*open access* avec celui des pages à contenus de qualité[5] de l'encyclopédie. Les processus de *peer-reviewing* ont été passés au crible avec les critères du *reviewing*, de la rapidité de publication, de la fiabilité, de l'auctorialité et du coût. Les conclusions de cette étude exploratoire sont que des avantages existent quand aux coûts de publication, au retour des critiques aux auteurs et à la possibilité de réaliser des corrections post-publication. Enfin, la structure de Wikipédia et ses openURLs est à même d'apporter une publicisation des résultats stable et pérenne. Il reste cependant à définir ce que serait une « publication scientifique » sur Wikipédia et à trouver l'accord académique des autorités quant à ces nouvelles pratiques potentielles.

– Wikipédia et la construction des savoirs

La dynamique de développement de la communauté Wikipédia est rendue observable par l'auto-archivage de sa propre activité. En se souciant de fournir les données de son histoire (Ruzé, 2011) et des outils pour les analyser, elle suscite un intérêt constant des chercheurs de multiples horizons regroupés en projets, quelquefois en groupes d'intérêt spécifiques comme SIGWP[6] (*Special Group on Wikipedia Research*) qui laisse les résultats des travaux en accès libre comme un visualisateur de thésaurus permettant de parcourir l'encyclopédie[7].
Wikipédia, encyclopédie organisatrice de communautés de pratique, testant un modèle collectif de production de savoirs attire les chercheurs engagés dans les savoirs communs ou « *knowledge commons* », investis dans la conception de systèmes participatifs, qui voient dans l'encyclopédie la réalisation de processus collaboratifs par des communautés dans des projets en ligne ouverts. (Jullien, 2012) dans son état de l'art examine ainsi les processus et patterns de la participation, l'organisation, la structure et la gouvernance de ce projet sociotechnique.
Les conclusions de chercheurs intéressés par la diffusion des savoirs – qui concernent aussi bien l'éducation que les organisations utilisant des wikis pour partager des connaissances – tendent à procurer des conseils précis comme (Cho and al., 2010) visant à reconnaître la participation et construire les cadres encourageant la réciprocité et le sentiment d'appartenance.

[5] http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Contenus_de_qualit%C3%A9
[6] http://sigwp.org/en/index.php/Main_Page
[7] http://sigwp.org/en/index.php/Wikipedia_Thesaurus_Visualizer

Du côté des Humanités, des chercheurs encouragent les professionnels de l'information à la production des contenus avec les Wikipédiens pour diversifier les représentations historiques (Luyt, 2011).
Comme (Matei & Dobrescu, 2011) souhaitant que la « conversation » initiée sur le NPOV et l'ambiguïté devienne un « dialogue » rigoureux sur la nature des processus culturels et sociaux reflétés par les travaux collectifs bâtissant l'encyclopédie, nous reconnaissons le caractère fascinant de cet objet d'étude co-construit.

Conclusion

Bien qu'internalisées, les règles, procédures et politiques de gouvernance de Wikipédia sont accessibles au public comme la totalité des échanges archivés autour de la fabrication des contenus.
Cette ouverture de l'encyclopédie à la création des contenus tout autant qu'à leur organisation et leur étude ne cesse de susciter l'intérêt des chercheurs qui saisissent l'occasion d'expérimenter, d'observer et de décrire. Cette activité rejaillit à son tour sur l'encyclopédie qui participe à de multiples projets d'animation scientifique. Une diversité de pratiques fertilise les échanges entre acteurs dans et hors l'encylopédie. Nous retiendrons le symbole récent de l'intégration d'un wikipédien géographe à l'Université de Berkeley, chargé d'enseigner la rédaction scientifique pour faciliter la lecture pour tous[8] de l'encyclopédie.

BIBLIOGRAPHIE

ABBOTT Andrew, « Varieties of ignorance », *American Sociologist*, 2010, n°41, p. 174–189. DOI : 10.1007/s12108-010-9094-x

ASADI Saeid et al., « Motivating and discouraging factors for Wikipedians: the case study of Persian Wikipedia », *Library Review*, 2013, Vol. 62, issue 4/5, p. 237-252.

AURAY Nicolas et al., « La négociation des points de vue » une cartographie sociale des conflits et des querelles dans le Wikipédia francophone, *Réseaux, 2009,* n° 154, p. 15-50. DOI : 10.3917/res.154.0015.

BOSSARD Aurélien, GUIMIER DE NEEF Emilie, « Le résumé par classification » Principes et applications, *Document numérique*, 2012 /2, vol. 15, p. 11-39. DOI : 10.3166/DN.15.2.11-39

CALLAHAN Ewa, HERRING Susan, « Cultural bias in Wikipedia content on famous persons », *Journal of the American society for information science and technology*, 2011, 62(10), p. 1899–1915. DOI: 10.1002/asi.21577.

[8] Une « chaire Wikipedia » créée à l'université de Berkeley. Url : http://www.actualitte.com/education-international/une-chaire-wikipedia-creee-a-l-universite-de-berkeley-48917.htm

CARDON Dominique, LEVREL Julien, « La vigilance participative. Une interprétation de la gouvernance de Wikipédia », *Réseaux*, 2009, n° 154, p. 51-89. DOI : 10.3917/res.154.0051

CARPINETO Claudio et al., « Mobile Information Retrieval with Search Results Clustering : Prototypes and Evaluations », *Journal of the American society for information science and technology*, 2009, 60(5), p. 877–895. DOI: 10.1002/asi.21036.

CHO Hichang, CHEN MeiHui, CHUNG Siyoung « Testing an integrative theoretical model of knowledge-sharing behavior in the context of Wikipedia » *Journal of the American society for Information Science and Technology,* 2010, n° 61(6), p. 1198–1212.

DUTREY Camille et coll., « Typologie des modifications dans les révisions de Wikipédia », Notes et documents LIMSI n° : 2011-01.

GRAPPY Arnaud, GRAU Brigitte, « Validation du type de la réponse dans un système de questions réponses », *Document numérique*, 2001/2, Vol. 14, p. 125-147. .

HALFAKER Aaron, GEIGER R. Stuart, MORGAN Jonathan, RIEDL John, « The rise and decline of an Open Collaboration System: How Wikipedia's reaction to sudden popularity is causing its decline », *American Behavioral Scientist*, 2013, n°57(5), p. 664-688, 2013.

HARA Noriko, SHACHAF Pnina, FOON HEW Khe, « Cross-Cultural Analysis of the Wikipedia Community », *Journal of the American society for information science and technology,* 2009, 61(10), p. 2097–2108.

HJØRLAND Birger, « Evaluation of an information source illustrated by a case study : effect of screening for breast cancer », *Journal of the American society for information science and technology*, 2011, n°62(10), p. 1892–1898. DOI: 10.1002/asi.21606

ITO Masahiro et al., « Semantic relatedness measurement based on Wikipedia link co-occurrence analysis », *International Journal of Web Information Systems*, 2011, vol. 7, n° 1, p. 44-61.

JACQUEMIN Bernard, « Autorégulation de rapports sociaux et dispositif dans Wikipedia », *Document numérique,* 2011, vol. 14, n°3, p. 57-79.
URL : http://www.cairn.info/resume.php?ID_ARTICLE=DN_143_0057

JULLIEN Nicolas, « What we know about Wikipedia. A review of the literature analyzing the project(s) ». Working Paper, SSRN: http://ssrn.com/abstract=2053597

LEE Jae-Won et al., « A probabilistic approach to semantic collaborative filtering using world knowledge », *Journal of Information Science*, 2011, n°37 (1), p. 49-66. CILIP, DOI: 10.1177/0165551510392318.

LU Jianguo, LI Dingding, « Estimating deep web data source size by capture–recapture method », *Information Retrieval,* 2010, n°13, p. 70–95. DOI 10.1007/s10791-009-9107-y.

LUYT Brendan, « The nature of historical representation on Wikipedia : Dominant or alterative historiography? », *Journal of the American society for information science and technology*, 2011, n°62(6), p. 1058–1065. DOI: 10.1002/asi.21531

LUYT Brendan, TAN Daniel, « Improving Wikipedia's credibility : references and citations in a sample of history articles », *Journal of the American society for information science and technology*, 2010, n°61(4), p. 715–722. DOI: 10.1002/asi.21304

MATEI Sorin Adam, DOBRESCU Caius, « Wikipedia's "neutral point of view": settling conflict through ambiguity », 2011, *The Information Society*, n°27, p. 40–51.

MEHLER Alexander et al., « Geography of social ontologies: Testing a variant of the Sapir-Whorf Hypothesis in the context of Wikipedia », *Computer Speech and Language*, 2011, n°25, p. 716–740.

MORSEY Mohamed et al., « DBpedia and the live extraction of structured data from Wikipedia » in *Program : electronic library and information systems,* 2012, vol. 46 No. 2, p. 157-181. URL : www.emeraldinsight.com/0033-0337.htm

NADAMOTO Akiyo et al., « Extracting content holes by comparing community-type content with Wikipedia », *International Journal of Web Information Systems*, 2010, vol. 6 No. 3, p. 248-260.

NGO Quoc Hung et al., « Using Wikipedia for extracting hierarchy and building geo-ontology », *International Journal of Web Information Systems*, 2012, vol. 8, No. 4, p. 401-412.

NUNES Sérgio et al., « Term weighting based on document revision history », *Journal of the american society for information science and technology,* 2011 ? n° 62(12), p. 2471–2478.

OLLIVIER Christian, « Ecriture collaborative en ligne : une approche interactionnelle de la production écrite pour des apprenants acteurs sociaux et motivés », *Revue française de linguistique appliquée,* 2/2010, vol. XV, p. 121-137.
URL : www.cairn.info/revue-francaise-de-linguistique-appliquee-2010-2-page-121.htm.

PARK Namkee et al., « Factors influencing intention to upload content on Wikipedia in South Korea : The effects of social norms and individual differences », *Computers in Human Behavior*, 2012, n° 28, p. 898–905.

PEHCEVSKI Jovan et al., « Entity ranking in Wikipedia: utilising categories, links and topic difficulty prediction », *Information Retrieval,* 2010, n°13, p. 568–600. DOI 10.1007/s10791-009-9125-9

PHARO Nils, KRAHN Astrid, « The effect of task type on preferred element types in an XML-based retrieval system », *Journal of the American society for information science and technology,* 2011, 62(9), p. 1717–1726.

RUZE Emmanuel, « Nouvelles des archives. Une approche quantitative des archives numériques d'un projet encyclopédique, Wikipédia », *Entreprises et histoire*, 2011/2, n° 63, p. 86-99. DOI : 10.3917/eh.063.0086

SCHACHAF Pnina, HARA Noriko, « Beyond vandalism: Wikipedia trolls », *Journal of Information Science*, 2010, n°36, p. 357-370.

SUNDIN Olof, « Janitors of knowledge : constructing knowledge in the everyday life of Wikipedia editors », *Journal of Documentation*, 2011, vol. 67, No. 5, p. 840-862.

VECHTOMOVA Olga, « Facet-based opinion retrieval from blogs », *Information Processing and Management*, 2010, n°46, p. 71–88. DOI:10.1016/j.ipm.2009.06.005.

WILSON Patrick, *Second-hand Knowledge: An Inquiry into Cognitive Authority*, Greenwood Press, Westport, CT and London, 1983.

XIAO Lu, ASKIN Nicole, « Wikipedia for academic publishing: advantages and challenges », *Online Information Review*, 2012, vol. 36, No. 3, p. 359-373. URL : www.emeraldinsight.com/1468-4527.htm