Une approche quantitative des archives d'un projet numérique, Wikipedia (2001-2009).

Emmanuel Ruzé

UNE APPROCHE QUANTITATIVE DES ARCHIVES NUMERIQUES D'UN PROJET ENCYCLOPEDIQUE, WIKIPEDIA.

par Emmanuel Ruzé ©
Telecom ParisTech.

Nous présentons les archives numériques de Wikipedia, entre autres celles des listes de régulation des projets. Nous montrons qu'il existe paradoxalement des pratiques mémorielles au-delà des automatismes informatiques. Nous analysons ensuite la dynamique de la plus ancienne des listes, proposons une base de données utile issue de son dépouillement « manuel », et une approche comparatiste des différentes séries débouchant sur des éléments de périodisation du projet de ses origines de 2001 à 2009 inclus. Nous identifions des recherches à mener au-delà des quelques travaux à portée historique mentionnés.

Nous présentons ici, dans la lignée des numéros récents abordant la question de l'histoire de l'économie numérique et des systèmes d'information et de communication[1] un gisement d'archives considérable[2] mais encore exploité de façon incomplète par les chercheurs des sciences sociales qui s'intéressent pourtant à la dynamique des sites dits de « contenu ouvert »[3].

Nous rappelons que Wikipedia, célèbre encyclopédie collaborative, n'est pas seulement un site proposant du contenu numérique public, mais est aussi un projet industriel (utilisation d'un logiciel et de robots par exemple), ce qui implique des discussions techniques, des connaissances informatiques, et des formes de gouvernance.

Nous avons par ailleurs dépouillé la plus ancienne des archives des listes de discussions se trouvant dans ce gisement [4] et proposons à la curiosité des chercheurs la base de données démographique qui en est le fruit[5]. Elle donne une idée au minimum exploratoire et quantitative de la dynamique de régulation d'un grand collectif en ligne au moyen d'un artefact technologique numérique, en particulier de différentes périodisations et points saillants possibles.

[1] Nous renvoyons aux numéros 43, 55 et 60.

[2] L'URL est ici : https://lists.wikimedia.org/

[3] Cette recherche menée pour l'essentiel à Telecom ParisTech a été permise par un financement post-doctoral de l'Ecole polytechnique.

4 Source: http://lists.wikimedia.org/pipermail/wikipedia-l/

[5] S'adresser à l'auteur pour le fichier Excel: Emmanuel.Ruze@polytechnique.edu .

BREVE REVUE DE LITTERATURES

Cette présentation ne cache cependant pas que Wikipedia a sollicité l'attention d'un nombre significatif de chercheurs. Mais cela ne veut pas dire que les approches historiques soient exhaustives, loin de là, surtout sur une période significative. De plus, prendre du recul par la description de régularités est une approche qui comporte un intérêt en matière de vulgarisation.

Nous n'avons trouvé qu'un seul travail qui tente l'analyse des matériaux présentés ici, une thèse en cours en sciences politique[6]. En ce qui concerne les approches historiques du projet, quelques chercheurs[7] ont eu le mérite de mettre en évidence une rupture dans l'histoire de la communauté, en 2006 : elle fait suite à la période 2002-2006, où 50% des éditions sont le fait des administrateurs. Ce pourcentage déclinant nettement au-delà de 2006 permet aux auteurs de qualifier cette nouvelle période de « montée de la bourgeoisie ». D'autres travaux permettent l'identification de régularités (« patterns ») en matière de contributions individuelles[8]. Une analyse de réseau a été mise en œuvre (jusqu'à 2006 seulement) pour analyser l'aspect relationnel du wiki, mais non pour les listes de régulations[9]. Bref, le caractère marginal de ce type d'approche diachronique est courant[10]. Pourtant, la recherche suggère que les problèmes d'organisation du projet n'ont pas manqué[11], qu'il s'agisse de débats sur la viabilité du projet, de sa philosophie ouverte[12], ou de l'augmentation des conflits avec la taille du projet[13] ; or, une approche historique peut être utile sur ces questions. On dispose donc déjà d'éléments exploratoires, mais guère de perspectives sur les dynamiques d'ensemble, alors que de nouvelles statistiques accessibles suite à l'ajout d'outils par la communauté elle-même permettrait d'en savoir plus. Il manque par ailleurs une étude généalogique des débuts de la régulation de Wikipedia, en remontant aussi loin que possible dans le passé.

Nous reviendrons plus en détail à la fin de l'article sur les pistes de recherches possibles.

PRESENTATION DES ARCHIVES

Désormais en effet, les projets numériques basculent dans le territoire des historiens, qui doivent s'intéresser à l'écriture de leur histoire, à moins de la laisser (ainsi que l'archivage) justement à la collaboration des « amateurs »[14] des communautés. Des éléments d'une communauté peuvent mourir ou agoniser lentement, et donc basculer dans l'histoire. Notre propos n'est pas de porter un jugement, mais d'examiner des pratiques mémorielles qui peuvent sembler paradoxales.

6 Le travail en cours de Mayo Fuster Morell se trouve sur ce site : http://www.onlinecreation.info/ .

7 Kittur Aniket et al., T. "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie", *CHI 2007*, San Jose, CA, 2007.

8 Felipe Ortega et al., "Quantitative analysis of the Wikipedia community of users", *Proceedings of the 2007 international symposium on Wikis*, 2007.

9 Luciana S. Buriol, et al., "Temporal Analysis of the Wikigraph", *IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*, pp 45-51, 2006.

10 Pour une autre revue de littérature, voir Ruzé Emmanuel, *Collaboration massivement distribuée et gestion du savoir en ligne. Le cas du wiki de la communauté WordPress (2003-2008)*, thèse de doctorat en économie et sciences sociales, Ecole polytechnique, 2009.

11 Viegas Fernanda et al., "Talk Before You Type: Coordination in Wikipedia", *40th Annual Hawaii International Conference on System Sciences*, 2007.

12 Goldman Eric, "Wikipedia's Labor Squeeze and its Consequences", *Journal of Telecommunications and High Technology Law*, Vol. 8, 2009.

13 Suh Bongwon et al., "Us vs. them: understanding social dynamics in Wikipedia with revert graph visualizations", IEEE Symposium on Visual Analytics Science and Technology, Sacramento, CA, 163-170, 2007.

14 Citation sans aucune connotation péjorative d'un concept usité.

Un début de politique mémorielle

Voici par exemple quelques liens aux intitulés explicites montrant qu'une communauté se préoccupe des problèmes de sa propre histoire et que contrairement à l'idée qu'on se fait des wikis, la gestion des données passées n'est pas automatique :

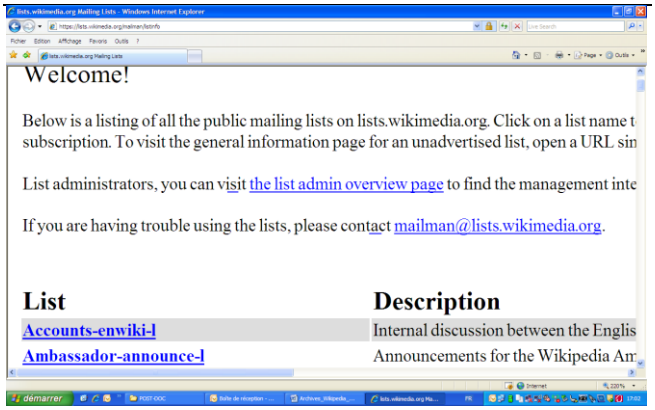
http://en.wikipedia.org/wiki/Help:Archiving_a_talk_page
http://en.wikipedia.org/wiki/Wikipedia:Historical_archive
http://meta.wikimedia.org/wiki/Historical_Wikipedia_pages
http://en.wikipedia.org/wiki/Wikipedia:History_of_Wikipedian_processes_and_people
http://en.wikipedia.org/wiki/History_of_Wikipedia
http://en.wikipedia.org/wiki/Category:Wikipedia_archives

La communauté se préoccupe donc d'écrire sa propre histoire, qu'il s'agisse des contributeurs ou des manières de composer une encyclopédie, de se souvenir des pages d'importance historique, d'archiver les données et les pages. On trouve ailleurs des souvenirs des premières années du projet[15].

Un bel exemple d'archives dont la catégorisation a évolué est celle de « Village Pump » (voir annexe n°2 pour une capture d'écran) : la régulation à un niveau plus général se fait sur cette partie du wiki nommée selon le modèle de la pompe des villages d'antan où s'échangeaient des informations et où se prenaient des décisions informelles. Il n'y avait pas de politique d'archivage au début, et par la suite une répartition des discussions entre différentes thématiques a été mise en place (« news », « policy », « technical », « proposal », « assistance », « miscellaneous »).

Les archives des listes de discussions

Les archives contiennent les archives de discussions sur Wikipedia, ce qui montre que la régulation de la communauté doit subir une politique d'archivage. Nous proposons par exemple en annexe 1 les archives de la liste de discussions que nous avons dépouillée, le dépôt descendant jusqu'à 2001. Voici à présent le haut de la page menant à l'ensemble des archives[16] **:**



Chaque ligne mène à un dépôt semblable à celui de l'annexe 1.

Quelques questions techniques

Des outils d'analyse récents ont été mis en place par la fondation MediaWiki pour permettre des retours vers le passé, qu'il s'agisse des contributeurs aux articles qui veulent avoir une vue d'ensemble d'un travail d'édition, ou des chercheurs. Mais ils sont spécifiques aux projets relevant de la fondation WikiMedia, et fonctionnant bien entendu sur moteur MediaWiki. Ils proposent des statistiques générales « en coupe », des statistiques diachroniques permettant d'identifier les phases importantes d'édition, l'évolution de la taille d'un article, les caractéristiques des éditions des 50 plus gros

[15] DiBona Chris, Cooper Danese, Stone Mark, *Open sources 2.0.: the continuing evolution*, O'Reilly Media, Inc., 2005. Voir le chapitre "The early history of Wikipedia and Nupedia: a memoir" (c'est à dire une "vision personnelle").

[16] https://lists.wikimedia.org/

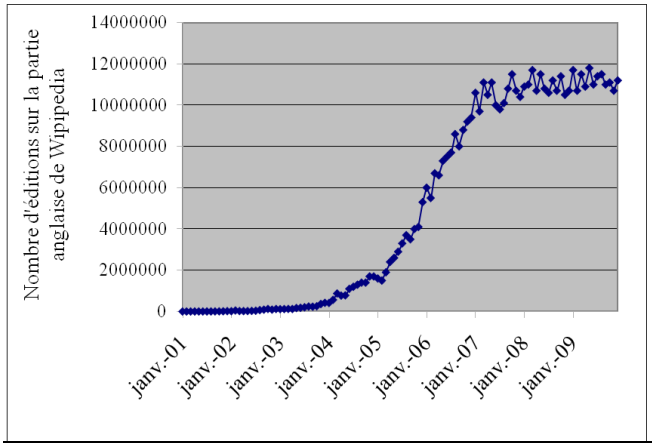
contributeurs. Nous renvoyons à la capture d'écran en annexe 4 pour un aperçu.

Par ailleurs, on peut aussi trouver des formes de classifications et de dépouillements externes à la communauté[17] (voir annexe 3). Sur cette capture d'écran, on trouve une tentative de classement des listes de discussions (les données numériques sont au dessus, mais invisibles), avec trois niveaux, stratégique (« movement », « foundation »), par projet (« project »), opérationnel (« technically oriented ») ; le classement est discuté voire pour partie obsolète. Cependant, il donne une idée du contenu des discussions, qui rappelle peut-être des conceptualisations usitées dans des organisations plus classiques. On trouve ailleurs des tentatives multiples de traitement des données au point que Wikipedia devient même un terrain d'expérimentation d'algorithmes[18]. Un chercheur se doit de les prendre en compte, avec précautions.

LA DYNAMIQUE DU PROJET

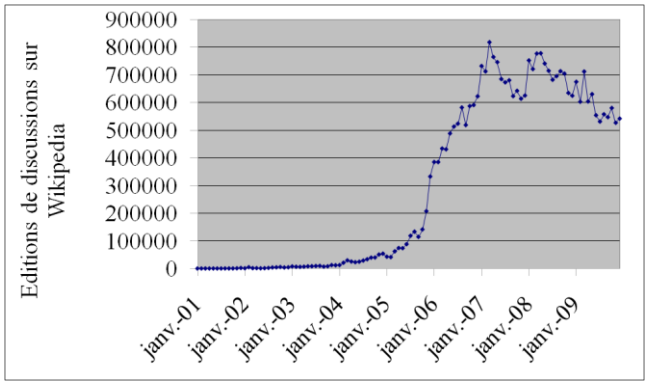
Cette partie est le résultat du traitement des données permis par la mise à disposition d'outils d'extraction et par le soucis de la communauté d'offrir des statistiques sur sa propre activité. Voici d'abord la dynamique des contributions (« éditions ») à la partie anglaise de Wikipedia elle-même :

Figure 1 : dynamique (non-cumulée) des éditions sur la partie anglophone de Wikipedia[19]



L'activité du projet suit une courbe en « S » assez classique, à ceci près que l'activité comporte des oscillations à partir de 2007. Mais ce n'est pas la seule activité sur Wikipedia : les contributeurs peuvent discuter entre eux du contenu sur des onglets spécialisés dits « de discussion », voisins des articles à contenu encyclopédique. Voici donc la dynamique des éditions de discussions sur les onglets :

Figure 2 : les éditions de discussions sur Wikipedia



L'année 2006 a été, de même que dans le cas des éditions sur les articles, une période de croissance

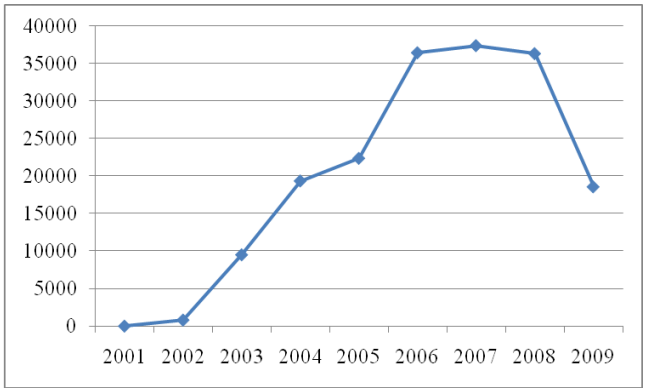
[17] Source:http://www.infodisiac.com/Wikipedia/ScanMail/index.html

[18] Voir par exemple Carlo A. Curino, et al., "Managing the History of Metadata in Support for DB Archiving and Schema Evolution", *Lecture Notes in Computer Science*, 2008, 5232, 78-88.

[19] Source : http://stats.wikimedia.org/EN/TablesWikipediaEN.htm#editdistribution

exponentielle. On remarque que les discussions sur les onglets ne sont pas la seule forme de régulation possible. Voici sur ce point le nombre d'éditions sur les pages de régulation de la communauté dites « Village Pump »[20] :

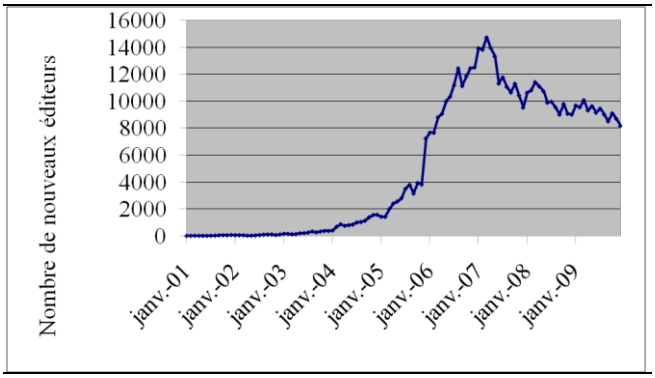
Figure 3: dynamique des éditions sur « Village Pump »



On constate donc un décalage par rapport aux séries chronologiques précédentes, la régulation de la communauté croît à partir de 2002. L'activité de régulation centralisée s'est faite plus en amont dans le temps.

Voici par comparaison les statistiques du nombre des nouveaux éditeurs mensuels, qui donne une idée de l'évolution démographique des contributions au contenu du projet:

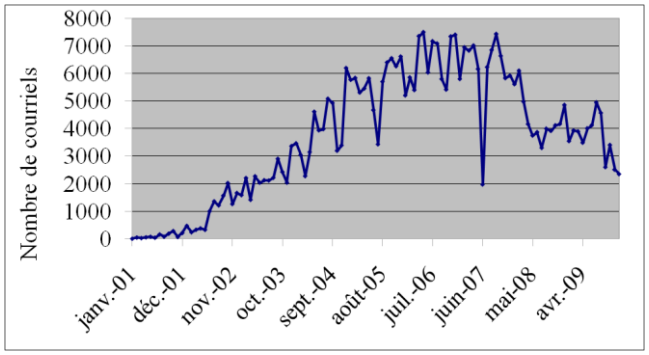
Figure 4: évolution démographique du nombre de nouveaux éditeurs sur Wikipedia



Un bref travail de comparaison entre les graphiques des figures 2 et 5, et une simple régression montre par ailleurs une relation structurelle stricte (R2=0,95) entre le nombre d'éditions de discussions et le nombre de nouveaux éditeurs, avec cependant quelques singularités début 2008.

Voici enfin, avec des données extraites des outils statistiques, la courbe de l'ensemble des listes de discussions de la fondation MediaWiki[21] :

Figure 5: agrégation des dynamiques des listes de discussions



On remarque encore que la dynamique ci-dessus est assez différente des précédentes concernant

[20] Source : http://en.wikipedia.org/wiki/Wikipedia:Village_pump .

[21] Source : http://www.infodisiac.com/Wikipedia/ScanMail/index.html

l'activité sur le wiki lui-même, mais est conforme à l'idée que des phases préparatoires de discussions furent nécessaires à la régulation du projet.

On constate donc que sur l'ensemble du projet, une structure assez nette se dégage :
- Une activité réduite de l'édition, du nombre de nouveaux éditeurs, et des éditions de discussions entre janvier 2001 et le début de l'année 2005.

- Une phase de croissance exponentielle en 2005 et 2006.
- Une phase de maturité avec un nombre d'éditions désormais stable, *mais* une baisse des discussions sur les listes, Village Pump, le wiki à mettre en relation avec une baisse du nombre de nouveaux éditeurs du wiki.
- Par comparaison, l'activité de régulation s'est effectuée plus en amont.

Nous proposerons plus bas une synthèse pour prise de recul.

QUELQUES RESULTATS EXPLORATOIRES EN REMONTANT EN AMONT DU PROJET

Présentation des archives dépouillées

Nous allons évoquer à présent les résultats de l'analyse d'une archive, celle de la liste précédemment mentionnée que nous avons dépouillée « à la main »[22].

Elle comprend, jusqu'en décembre 2009, 30986 courriels, ce qui est considérable. Cela suffit pour produire des résultats exploratoires. Des statistiques descriptives élémentaires "en coupe" montrent un fort débit et une dispersion notable de l'activité :

[22] Cela demande quatre mois de travail à raison de 35 heures par semaine approximativement. Cela permet de repérer les « doublons » dans la liste et de disposer des données mensuelles pour chaque individu qui y participe.

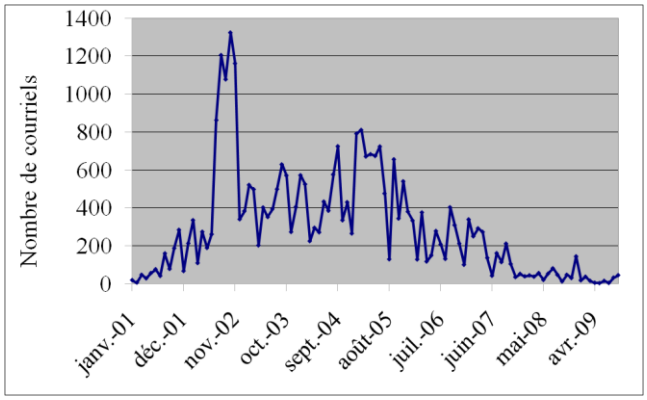
	Total	Moyenne	Ecart-type
Nombre total de courriels	30986	295,1	282
Nombre d'intervenants repérés	1178	49,2	31,3
Nombre de nouveaux		10,9	7,7
Nombre de fils de discussions	6903	63,9	68,1
Taille de la communauté		105,8	58,7

La base de données que nous proposons, trop conséquente pour être visualisée ici, permet d'identifier tous les intervenants, et le nombre de leurs interventions chaque mois, et donne une vision diachronique des statistiques ci-dessus.

La dynamique comparée des listes de discussions

Voici la dynamique de la liste de discussion dépouillée et étudiée :

Figure 6 : série chronologique de la plus ancienne liste de régulation



On peut à partir des courbes de tendance polynomiale et moyenne mobile (non reproduites ici) établir trois phases :

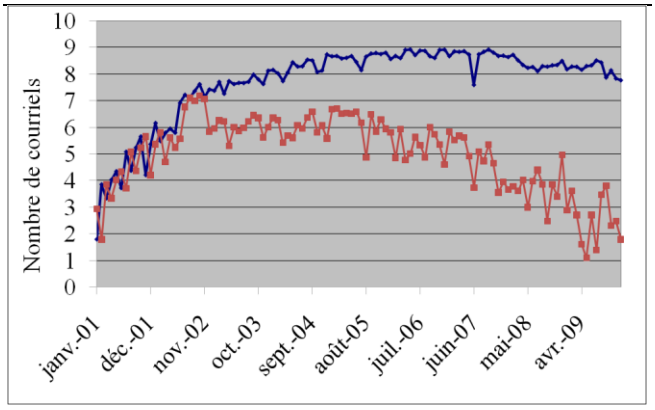
- de janvier 2001 à mai 2002, une phase de croissance lente
- de juin 2002-décembre 2002, une phase comportant un pic marqué.
- de janvier 2003 – août 05, un plateau avec des oscillations

- de septembre 2005, une phase de décroissance rapide, puis plus progressive[23].

L'activité cumulée est souvent décrite comme une courbe en « S », ce qui réduit la finesse de la description. Par exemple, plus l'activité de discussion devient intense, plus la variance de cette activité entre intervenants s'accroît (nous ne montrons pas le graphique).

On constate par ailleurs une différence de périodisation entre l'ensemble des listes et celle que nous étudions, plus visible si on compare à l'échelle logarithmique :

Figure 7 : comparaison de « wikipedia-l » avec les autres listes.



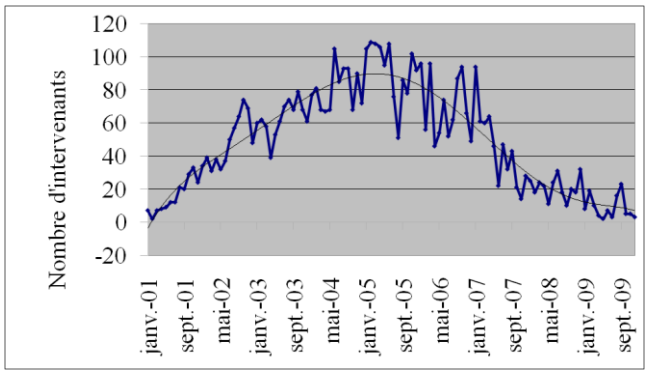
Les deux sommets, ceux de la liste étudiée et celle de la cumulée de toutes les listes ne correspondent pas. On observe une rupture début 2003, correspondant au lancement d'autres listes de coordination des autres aspects du projet. Cela correspond au pic observé en 2002-2003 sur la plus ancienne des listes, un point de rupture non répertorié méritant certainement plus ample examen.

[23] Cette rupture serait à mettre en relation avec celle trouvée par Kittur et al. .

QUELQUES ELEMENTS DEMOGRAPHIQUES

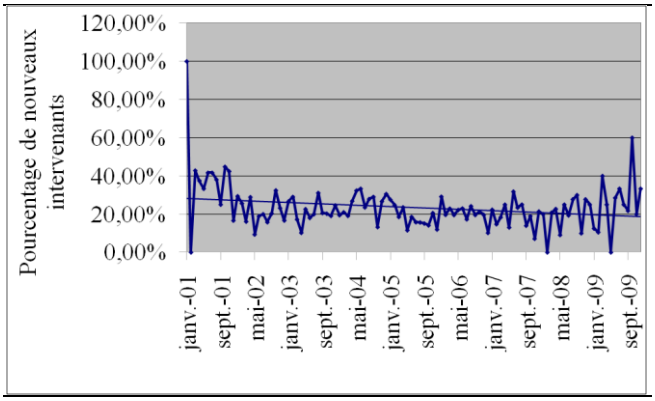
Voici à présent une série chronologique du nombre d'intervenants sur la liste, un des éléments de sa démographie :

Figure 8: évolution du nombre d'intervenants sur la liste



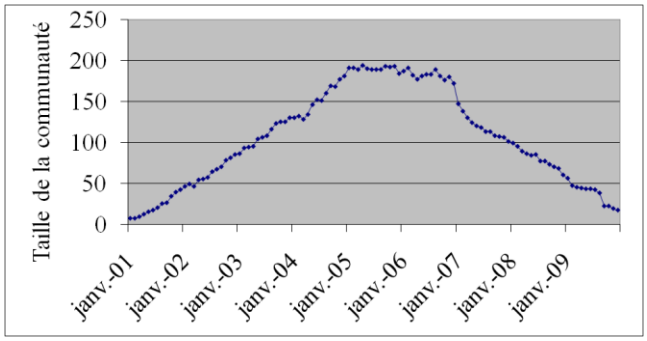
La périodisation en trois phases de la démographie de la liste correspond à celle de son activité, mais les sommets diffèrent. En revanche, lorsqu'on examine la dynamique des entrées et sorties sur la liste de discussion, on observe des régularités différentes. Tout d'abord, on observe une permanence d'intervention de nouveaux intervenants, autour de 20% des intervenants sur la liste chaque mois :

Figure 9 : évolution du pourcentage de nouveaux intervenants



On peut donc dire que le taux de renouvellement des intervenants est stable et que la liste a toujours attiré au point qu'on a toujours une proportion significative de nouveaux intervenants. En ce qui concerne ceux qu'on ne verra plus intervenir, on observe toujours un pourcentage régulier semblable (il y a toujours eu des départs), avec bien entendu une accélération à la fin (nous ne donnons pas le graphique). Soustraire les départs et les arrivées des intervenants permet de dégager cependant un « noyau » stable dans la communauté, dont il s'agit de déterminer la taille et son évolution. Voici ce qu'on obtient sur ce point[24] :

Figure 10 : nombre de personnes présentes sur la liste (taille de la communauté)



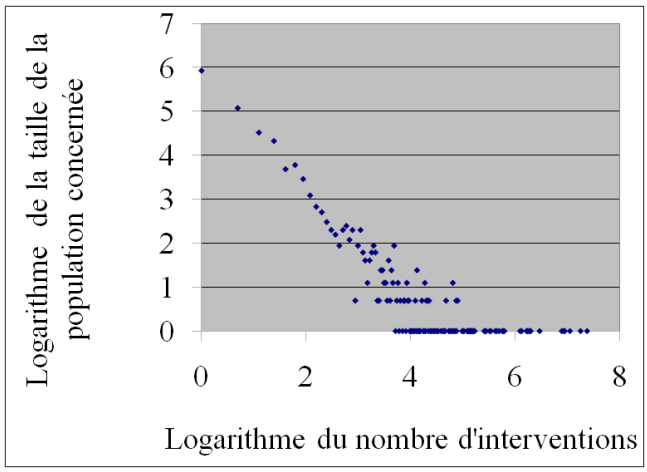
On distingue clairement trois phases bien distinctes caractérisant la taille de la communauté présente, phases invisibles précédemment, caractérisées par des tendances très régulières, quasiment linéaires (alors que cela n'était aucunement nécessaire), en "forme de volcan".

[24] Obtenir ce type de résultat suppose de travailler la base de données à la main afin de repérer avec des couleurs sur cette très grosse population le moment où chacun intervient pour la première fois et la dernière fois, et de décompter chaque mois le nombre d'individus ainsi « colorés ».

La structuration des interventions

Les calculs effectués en coupe en bout de séries chronologique et des mises en forme supplémentaires ont permis de mettre en évidence plusieurs lois de puissance comportant cependant des spécificités qu'il était bon de souligner, les originalités en la matière n'étant pas fréquentes sur le web et dans les grands collectifs.

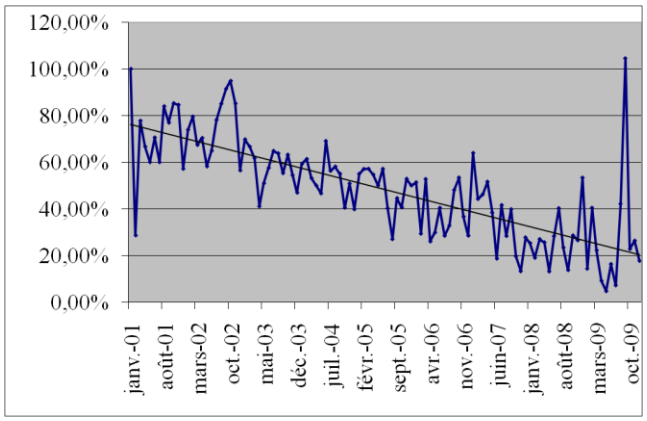
Figure 11 : distribution du nombre des interventions



On constate que la loi de puissance qui concerne le nombre d'interventions sur la liste n'est pas parfaite, comme le montre l'écart entre la droite de régression et une ligne imaginaire allant de 6 à 6 en abscisse et en ordonnée, ce qui est original sur la Toile. En ce qui concerne celle du nombre de mois avec intervention (i.e. avec une seule intervention), on observe que la distribution va de 6 en ordonnée à 4 en abscisse, ce qui n'en fait pas exactement une loi de puissance.

Quelle est la proportion d'intervenants qui interviennent effectivement ? Sur ce point, dont le calcul se fait en soustrayant les valeurs des figures 8 et 10, nous obtenons une tendance (ou « trend ») particulièrement marquée :

Figure 12 : baisse tendancielle de la proportion d'intervenants dans la communauté présente

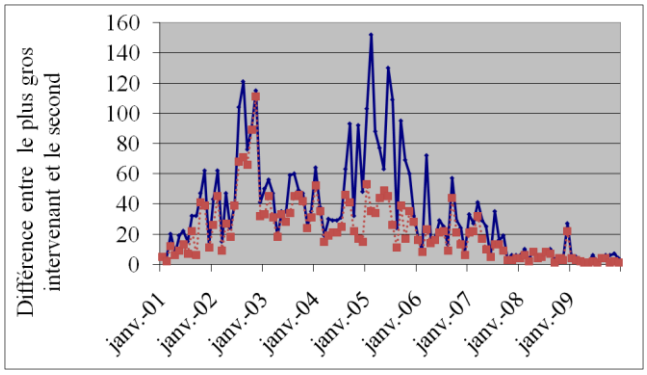


Là, on voit qu'au fur et à mesure de l'histoire de la communauté, celle-ci grandit de sorte que la proportion de la population qui intervient diminue régulièrement en pourcentage. Mais on peut aussi dire que le « trend » ne devient marqué qu'à partir de fin 2002. On note que sur la période "particulière" en 2002, on a un ratio plus fort qui « dépasse » de la courbe de tendance[25].

Ce type de phénomène peut comporter plusieurs explications, par exemple celle selon laquelle l'augmentation de la taille de la communauté limiterait la prise de parole et permettrait la formation d'une élite, visible sur la loi de puissance précédente. Encore une fois, une telle régularité n'est absolument pas nécessaire. La relation entre les plus gros intervenants se constate aisément :

[25] Remarque : le pic de la fin de la série est très certainement dû à un biais dû à 13 individus manquants sur 1178, mais il est marginal dans une série chronologique comportant une tendance baissière bien marquée. Nous ne le prendrons pas en compte a priori.

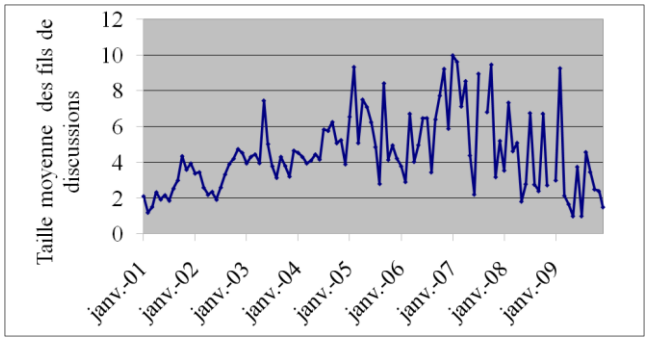
Figure 13 : nombre d'interventions des deux plus gros intervenants



Le plus gros intervenant correspond à la courbe du dessus. La structuration des interventions des 3e, 4e, 5e est quasi-identique à celle du 2e, et nous ne les avons pas ajoutées pour ne pas surcharger le graphique. Elles sont corrélées, seules les interventions du plus gros intervenant sur la liste se distinguent en 2005.

Enfin, nous présentons l'intensité des discussions (i.e. la taille moyenne des fils, le nombre moyen de courriels par fil), il commence de façon cyclique avant de devenir chaotique à partir de mi-2005, c'est-à-dire le sommet de la série chronologique étudiée:

Figure 14 : évolution de la taille des fils de discussions



On constate donc que l'intensité des discussions devient plus irrégulière, avec des pics de

discussions intenses à mesure que l'activité globale augmente et que l'activité sur l'encyclopédie entre dans une phase exponentielle. Il y aurait là une curiosité à creuser.

SYNTHESE

A partir des données statistiques mise en forme, nous avons pu mettre en évidence une baisse de l'activité de régulation, qu'il s'agisse des éditions de discussions, listes de discussions et Village Pump à partir de 2007, baisse qui ne se retrouve pas dans les statistiques des éditions et sur le wiki. Cette rupture (baisse activité de coordination) n'a pas été observée dans les travaux antérieurs qui décrivaient alors une croissance principalement axée sur ce type d'activité[26].

On a mis en évidence plusieurs autres ruptures dans la dynamique de la communauté. En 2003, un pic d'activité de la plus ancienne des listes correspond à une phase de décentralisation de la régulation sur d'autres listes. On observe aussi un bond exponentiel du nombre d'éditeurs et d'éditions de discussions début janvier 2006, rupture qu'on ne retrouve pas dans l'activité des listes de discussions ; faut-il y voir une forme de décentralisation supplémentaire ou un autre aspect de la « montée de la bourgeoisie », selon l'expression de Kittur et al. ?

On constate aussi que les deux chronologies diffèrent : la croissance du nombre d'éditions de contenu s'est déclenchée en 2004 avant de croitre de façon exponentielle, ce qui ne correspond pas à une phase originale sur la liste étudiée et sur l'ensemble des listes de régulation.

Un autre constat s'impose à partir des données sur l'ensemble du projet : il est difficile de dire fin 2009 s'il y a un ralentissement du projet ou juste une maturité de son organisation. Nous proposons une périodisation utile (quoique simple) de l'ensemble du projet utile à tout chercheur ou enseignant désirant s'emparer du sujet dans une perspective historique. Une approche quantitative sur la plus longue période possible manquait.

Nous avons examiné *l'articulation* entre la dynamique de la liste la plus ancienne et d'autres parties de la communauté (l'ensemble des articles, les discussions sur le wiki, le reste des listes de discussions), ce qui intéresse la théorie des organisations. Il est possible que la communauté soit passée d'un mode de coordination centralisé à un autre, plus décentralisé (listes, discussions) au fur et à mesure du développement du projet. Les modalités historiques d'une telle transition demanderaient d'ailleurs davantage d'examen. Mais nuançons : une communauté d'intervenants est toujours *présente*, jusque fin 2006, mais elle intervient de moins en moins à partir de fin 2002, mais avec des périodes de discussions intenses voire chaotiques. Un approfondissement dans cette voie relancerait dans une perspective plus globale le débat sur la « managérialisation » des communautés[27], mais à un niveau plus macro.

Nous pouvons aussi examiner le cycle de vie d'une liste de régulation « pour lui-même » et dégager des spécificités:

- Le pic d'activité de 2003 est singulier et se retrouve lorsqu'on extrait d'autres variables.
- Nous avons pu dégager la démographie et la dynamique de la communauté des intervenants présents et la spécifier par comparaison avec les interventions effectives.
- Nous avons articulé l'intensité des discussions avec la dynamique de l'encyclopédie.
- Nous avons pu observer des formes de leadership significatives.

[26] Viegas Fernanda et al., *ibidem*.

[27] Pour ce qui concerne les formes de management de projet dans les communautés en ligne, voir Ung Hang et Dalle Jean-Michel, « Project management in the Wikipedia community », *WikiSym* '10, July 7-9, 2010, Gdansk, Poland.

Conclusions methodologiques, pistes de recherche et difficultes

Nous avons montré qu'une communauté de cette taille s'est préoccupée de sa propre histoire, de l'archivage d'une partie significative de son activité, et propose même depuis peu des outils d'analyse qui éviteraient aux chercheurs un travail fastidieux. Cependant, l'automatisation de l'archivage des données comporte des limites, comme le montre notre travail exploratoire sur l'une des archives. Partir du postulat d'analyser des séries chronologiques même simples et d'adopter une perspective historique permet de dégager des périodisations différenciées et de soulever de nouvelles questions. Une telle approche permet de dégager d'autres éléments que les habituelles lois de puissance[28] et courbes en « S ». Elle peut être le point de départ d'un examen historique d'autres types de préoccupations plus politiques : que s'est-il réellement passé à telle ou telle période charnière et qu'est-ce que la comparaison des données nous dit ? Beaucoup reste à faire.

De nombreuses pistes sont en effet ouvertes pour traiter de façon plus approfondie de telles archives. Notre base de donnée permet déjà cependant de repérer à quel moment de l'histoire de la communauté on trouvera les « verbatims » de tel ou tel intervenant. D'un point de vue méthodologique, d'autres approches sont attendues, comme une analyse de réseau bien entendu, une comparaison des dynamiques des listes entre elles, une comparaison des biographies de différents intervenants à la liste de discussions, identifiés désormais dans la base de données proposée… Il est par ailleurs nécessaire d'articuler plus finement les données des archives des listes de discussions et les données extraites automatiquement par la communauté, comme nous l'avons un peu fait. Sur ce point, la quantification ne pose pas vraiment de difficultés en matière de critique des sources, contrairement à une analyse qualitative de matériaux de cette ampleur[29]. Il faut faire le lien avec l'histoire du projet connue des « amateurs », ce que nous n'avons pas fait ici pour assumer une approche quantitative et éviter les biais liés à la prise en compte non-réfléchie du savoir que la communauté a accumulée sur elle-même.

Au-delà de Wikipedia, une analyse des relations entre le projet encyclopédique et la fondation MediaWiki pourrait être envisagée de façon fructueuse à partir d'une analyse conjointe des caractéristiques des listes (démographies comparées des projets[30], moments politiques clefs…). Avec des données de cette taille, l'usage d'outils d'extraction de données et d'analyse de données qualitatives seraient certainement utiles. Enfin, pour prendre du recul, se pose aussi la question de l'articulation avec des projets institutionnels passés ou en cours (Autograph, PROSODIE) dans les autres sciences sociales.

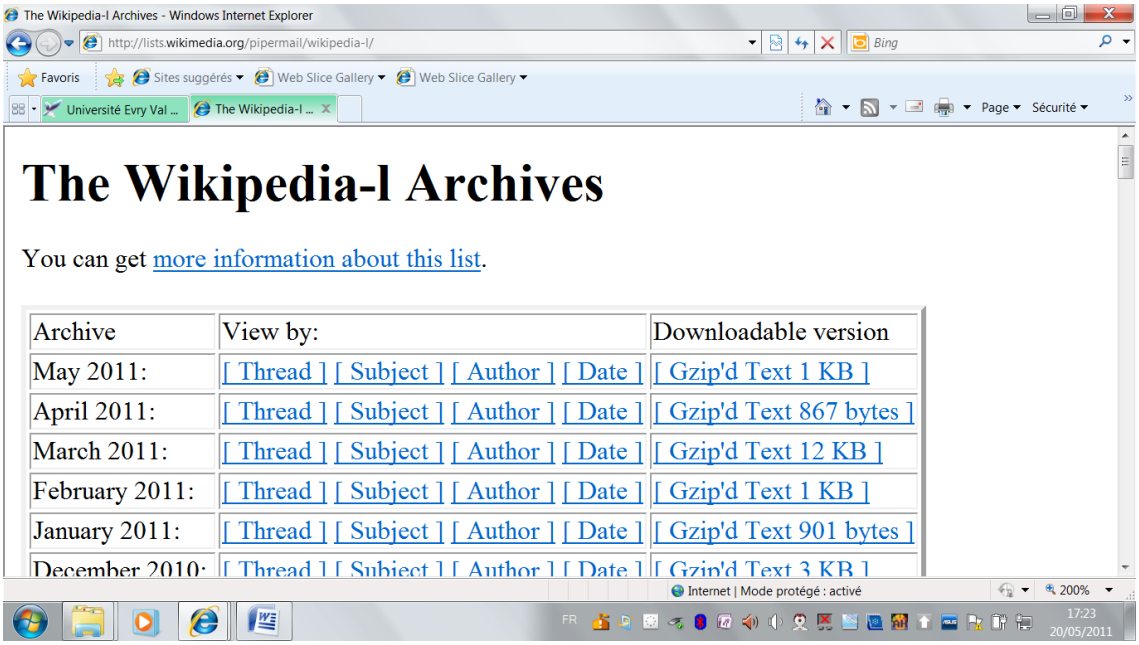
[28] Holloway, Todd et al., "Analyzing and visualizing the semantic coverage of Wikipedia and its authors", *Complexity*, 12: 30–40, 2007.

[29] S'agissant de la question de l'analyse critique des sources, nous renvoyons à : Ruzé Emmanuel, « Traiter les archives de la Toile. Une histoire d'un système d'information dans une communauté, WordPress (2003-2008) », *Entreprises et Histoire*, 55, 74-89, 2009.

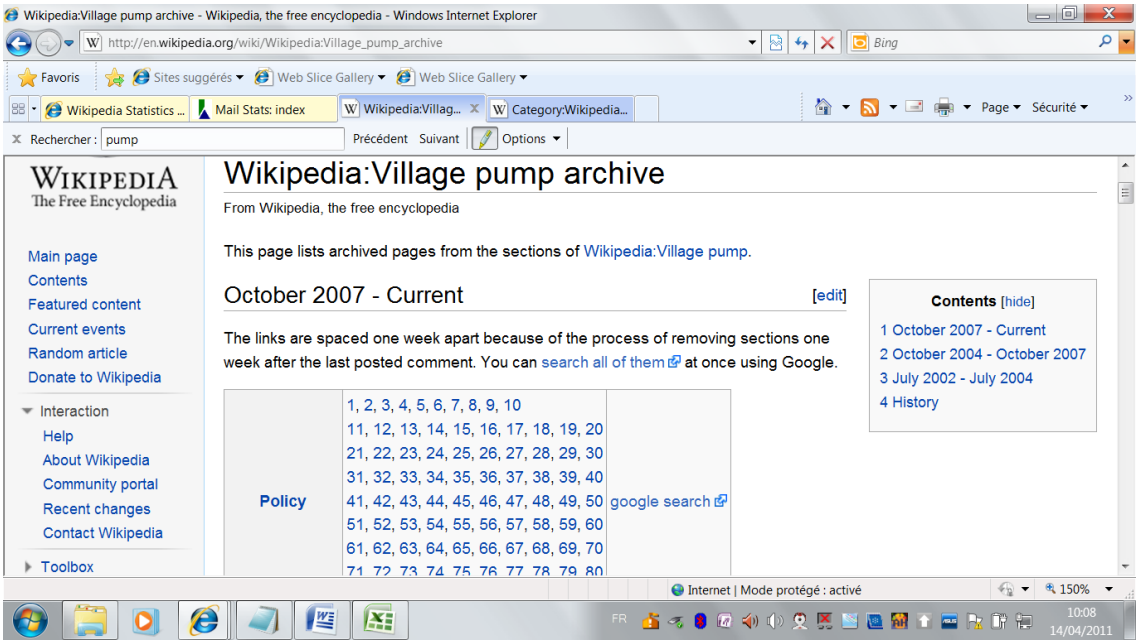
[30] Pour comparer avec les autres listes, on peut s'aider de cet outil: http://www.infodisiac.com/Wikipedia/ScanMail/_PowerPosters.html

Annexes : quelques captures d'écran de pratiques d'archivages et d'extraction de données

Annexe 1 : système d'archivage de la liste de discussions « Wikipedia-l »[31]



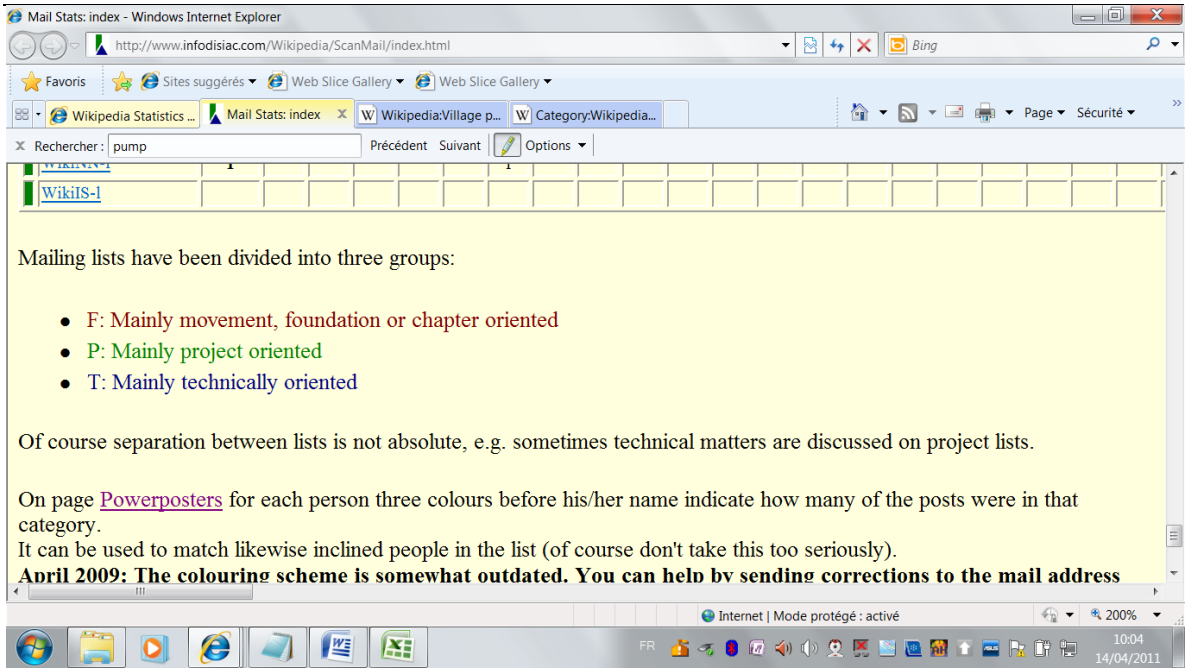
Annexe 2 : système d'archivage et de classement des discussions à la « pompe du village » de Wikipedia[32]



[31] Source : http://lists.wikimedia.org/pipermail/wikipedia-l/

[32] Source : http://en.wikipedia.org/wiki/Wikipedia:Village_pump_archive

Annexe 3 : système externe de classification des listes de discussions des projets et de statistiques élémentaires[33]



Annexe 4 : système automatique d'extraction de statistiques descriptives à partir des historiques d'articles[34]

[33] Source : http://www.infodisiac.com/Wikipedia/ScanMail/index.html

[34] Source : http://en.wikipedia.org/wiki/Wikipedia:History_of_Wikipedian_processes_and_people . Nous avons choisi justement les statistiques des contributeurs à un travail de mémoire du projet.