Traiter les archives de la Toile. Une histoire d'un système d'information dans une communauté, WordPress (2003-2008)

Emmanuel Ruzé

TRAITER LES ARCHIVES DE LA TOILE. UNE HISTOIRE D'UN SYSTEME D'INFORMATION DANS UNE COMMUNAUTE, WORDPRESS (2003-2008).

par Emmanuel Ruzé ©
Centre de Recherche en Gestion (Ecole Polytechnique).

Nous proposons d'aborder, dans la lignée du numéro de 2006 d'*Entreprise et Histoire* sur le "déploiement du numérique", une thématique méthodologique essentielle, celle du traitement des archives "en ligne". Le cas de celles de la communauté WordPress est particulièrement instructif techniquement et représentatif des pratiques d'archivages et des données auxquelles les chercheurs de toutes les sciences sociales intéressés par l'analyse historique du « déploiement du numérique » peuvent être confrontés.

La question du *traitement* des matériaux spécifiquement numériques, sous forme d'archives ou non, n'a été soulevée que de façon marginale, en particulier chez les historiens (Rygiel, 2005), excepté la question de l'archivage des données de l'administration électronique (Dhérent, 2002). Pour le type d'archives que nous mentionnons ici, les références sont également rares (Langner, 1997), et les traitement de données des approches sociologiques, comme Conein et Latapy (2008) seraient jugés insuffisants par les historiens. La prise en compte de la numérisation existe pourtant. Le service « Web archives » est né en 1996[1], alors que la Toile naît en 1989. L'histoire des réseaux informatiques a plus de quarante ans, fait l'objet de débats, et l'histoire contemporaine ne saurait faire l'impasse sur un phénomène aussi majeur.

Le champ disciplinaire de ce travail relève cependant de la gestion, plus précisément du management des systèmes d'information. *Ce positionnement est pertinent car quelques recherches y montrent déjà l'intérêt d'une approche historique*. Ainsi, François-Xavier De Vaujany (2006) a étudié la notion de diachronie lors d'une enquête sur les systèmes d'information de la Curie romaine dans une perspective de « longue durée ». Bannister (2002) propose des explications de la rareté regrettable des approches historiques dans cette discipline.

[1] http://www.archive.org/about/about.php, consulté en novembre 2008.

Notre objectif, au-delà de la présentation d'archives excitant la curiosité, est donc ici d'ordre méthodologique et technique. Nous souhaitons montrer que la démarche historique est indispensable et son intérêt sur des archives nouvelles où un chercheur n'y aurait pas forcément pensé. Il s'agit d'une part de *l'aide* au traitement scientifique satisfaisant de ce type de données, d'autre part de *son utilité* dans un travail en gestion ou en économie[2].

De fait, on n'associe plus guère aujourd'hui la démarche historique à une visée expansionniste en sciences sociales, par comparaison avec l'âge d'or de l'Ecole des Annales. On observe en théories des organisations un retour cyclique et ponctuel du besoin d'analyse historique, comme le remarquait déjà Kieser (1994). Pourtant, Herbert Simon (1999) demandait à ce que la spécificité de la démarche historienne soit davantage reconnue pour aborder des problématisations économiques comme l'économie du savoir.

Ici, plusieurs problèmes essentiels de nature historique sont abordés: la nature des archives (listes de discussion, historiques de wiki...), le repérage des sources pertinentes, les problèmes associés aux traitements qualitatifs des "traces", l'articulation à des questionnements économiques et des démarches modélisatrices et quantitativistes. Nous donnons un aperçu de l'utilité de l'analyse historique à quatre questions de recherche en sciences de gestion qui abordent le phénomène économique de la « collaboration massivement distribuée » caractéristique des wikis : les formes d'auto-organisation sont un aspect essentiel du phénomène historique dit « déploiement du numérique »[3].

Notre objet d'étude, celui des wikis dans le contexte des communautés open-source est d'une portée historique certaine, car les wikis ont été employés d'abord dans ce type de communautés, mais les travaux font défauts (Wagner, 2007). Le premier wiki de l'histoire, le « C2 », a été créé en 1994.

QUELQUES ELEMENTS HISTORIQUES

Présentation du cas

Nous proposons d'aborder notre question à partir du cas de l'analyse de l'histoire du succès du produit open-source associé à la communauté WordPress (2003-2008). Elle est devenue un acteur incontournable en matière de conception de moteur de blog open-source. La communauté est née en juin 2003[4], la « jeune pousse » (« start-up ») Automattic est lancée en « symbiose » le 20 décembre 2005[5]. Elle propose plusieurs services payants (hébergement et garantie de maintenance, logiciel anti-spam…) associés au logiciel, qui reste gratuit. Le nombre de téléchargements du logiciel dépasse les quatre millions, et des organisations prestigieuses comme le New York Times font usage de WordPress. Ce succès justifie le choix de ce terrain pour l'étude de cas. L'abondance des données en fait également un objet d'étude privilégié (Yin, 2003).

Notre propos n'est pas l'aspect visible du succès de la « jeune pousse », *mais l'histoire humble de la communauté nécessitant un retour vers ses archives*. L'objet d'étude choisi, qu'on pourrait comparer à celui très différent de l'organisation et des usages d'une bibliothèque dans une communauté monastique, est son système d'information d'importance stratégique, organisé autour du wiki baptisé par la communauté « Codex ». Un wiki est un site dynamique où

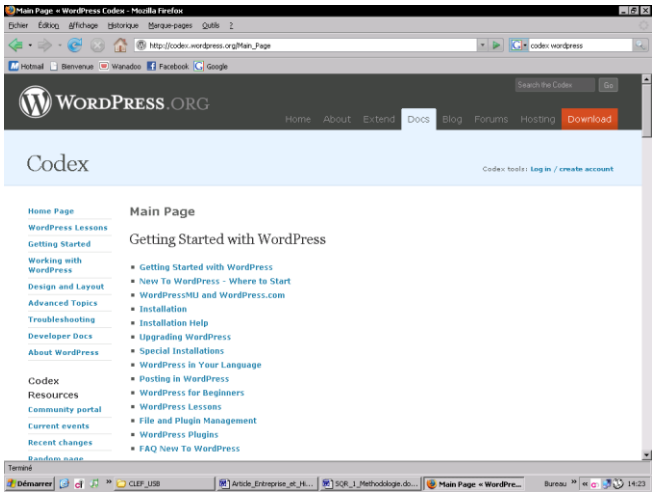
[2] Le cadre théorique en filigrane dans ce travail est l'économie cognitive au sens large. L'articulation entre l'économie hétérodoxe et l'histoire est *fondamentale*, mais n'a pas été développée ici (voir Lesourne et al., 2002 et Lorenz, 2001)

[3] Il existe toujours : http://c2.com/cgi/wiki?WelcomeVisitors (consulté le 02/03/2009)

[4] http://web.archive.org/web/*/http://wordpress.org/.

[5] http://web.archive.org/web/20051220125353/http://automattic.com/

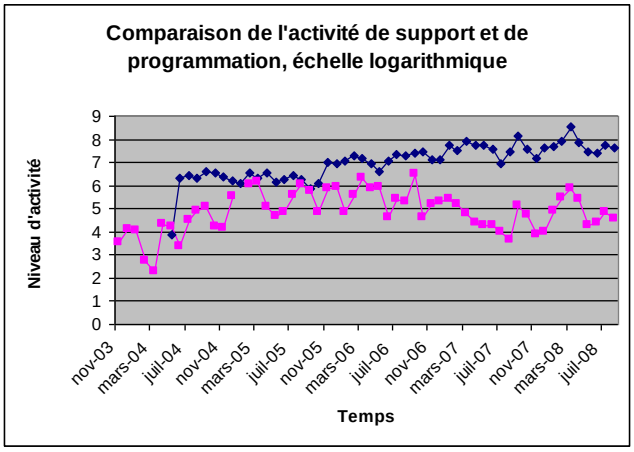
l'utilisateur peut modifier contenu et architecture des pages, ainsi qu'interagir avec d'autres contributeurs. L'objectif d'un tel outil est de répertorier de façon collaborative *l'ensemble du savoir concernant le produit offert par la communauté*, y compris les moyens et les façons de l'améliorer. Un premier wiki a été mis en place le 17 décembre 2003, il fut un échec. Le second a été lancé le 30 juillet 2004, et son histoire a été complexe.



Capture d'écran du système d'information « Codex » en août 2008.

Quelques éléments de l'histoire de la communauté.

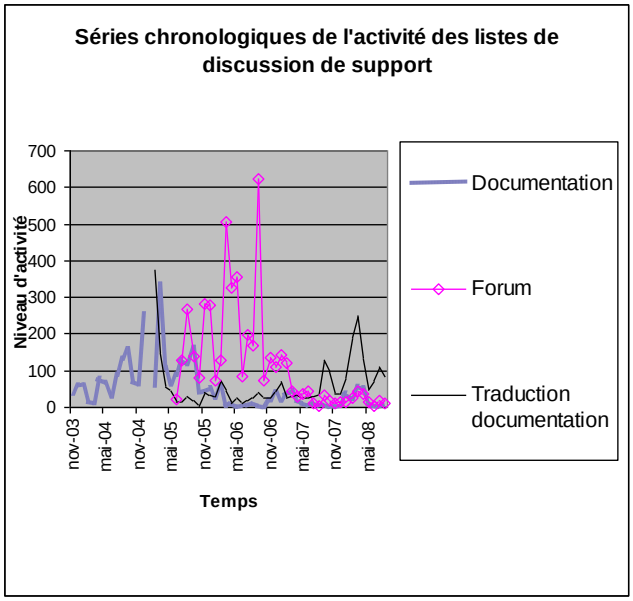
Le dépouillement des données des archives des listes de discussions et la quantification permet de dégager de grandes tendances de l'histoire de la communauté WordPress.



Comparaison de l'activité de support et de programmation, échelle logarithmique.

La courbe du bas correspond à l'addition de l'activité mensuelle de régulation des activités de support, qui inclut la documentation (2794 courriels). La courbe du haut correspond à l'addition de l'activité mensuelle des activités de programmation (73348 courriels). Les deux activités sont donc d'envergure différente : le pic de l'activité de programmation est de plus de 5000 courriels par mois, celui des activités de support 500.

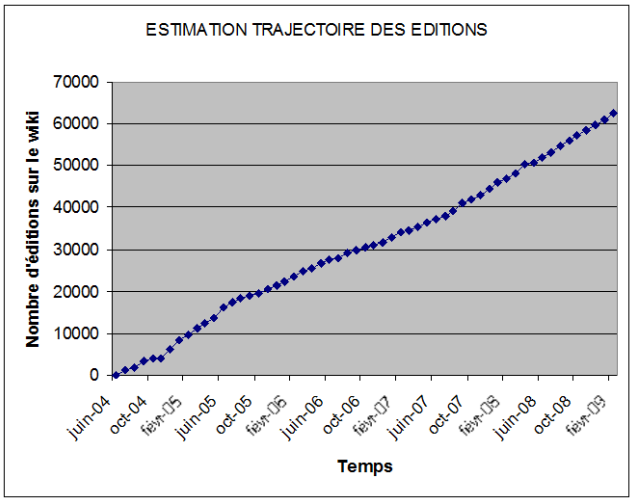
L'activité de programmation et l'activité de support n'obéissent pas aux mêmes rythmes. L'activité de support aux utilisateurs comporte trois phases de régulations successives de ses composantes : documentation, forum utilisateur, traduction de la documentation :



Séries chronologiques des listes de discussions des activités de support.

L'activité de programmation se caractérise au contraire par une période de stabilité (autour de 500 courriels par mois) et par l'émergence d'une activité croissante mais cyclique à partir de décembre 2006 liée au lancement de nouvelles versions du logiciel[6].

La reconstitution de l'activité totale sur le Codex suppose un travail d'ingénierie historique[7] sur Excel par interpolations à partir des traces statistiques disponibles sur archives.org (voir infra) et du dépouillement *manuel* par datation du nombre de nouveau articles créés chaque mois sur le wiki. On voit que la formation du bien collectif est un processus constant :



Estimation de la trajectoire des éditions sur le wiki.

PRESENTATION DES ARCHIVES EXAMINEES

Dans la communauté WordPress chaque sous-projet est coordonné par une liste de discussion (« mailing-list »), et leurs archives sont les principales disponibles ayant permis d'obtenir ces séries chronologiques. Le contenu est très technique et de longueur imposante.

La première capture d'écran en annexe présente un des entrepôts d'archives. Chaque dépôt peut être structuré, en cliquant sur un bouton, par fil de discussion, auteur, date, sujet. *Nous insistons, ces outils de structuration automatique d'archives sont très originaux pour un historien*. Un fichier sous format « txt » compressé, en partie droite, récapitule l'ensemble des courriels du mois, sous forme non structurée certes, ce qui est pourtant *très utile* pour rechercher des courriels par mots-clefs. Tous les dépôts d'archives comportent cette structure d'archivage automatique et la mise en ligne se fait automatiquement sous 24 heures d'après nos observations.

La deuxième capture d'écran présente l'activité mensuelle (quantifiable) d'une liste de discussion. Chaque ligne représente un lien vers un courriel.

[6] On peut deviner la présence de cycles dans le précédent graphique.

[7] Trouver la relation manquante entre les deux données équivaudrait à estimer par tâtonnements une équation différentielle décrivant l'évolution croissante du ratio « nombre moyen d'éditions par article ». Cette estimation va plus loin que Den Besten et al. (2008)

Ceux se trouvant du côté gauche sont ceux qui ont initié un fil de discussion, ceux qui se décalent vers la droite sont des réponses successives aux premiers. Chaque courriel est associé à une adresse « url » permettant de le localiser précisément dans une archive[8].

La troisième capture d'écran représente un historique de contributions d'un administrateur sur les pages du wiki. Il peut tout à fait intervenir sur de nombreuses pages, ou, au contraire, se spécialiser sur un petit nombre. Toutes ses interventions sont datées à la minute près, et un lien nommé « diff » (pour « différence ») généré automatiquement à chaque intervention permet de discerner les apports incrémentaux du contributeur en question, et d'effectuer des comparaisons dans le temps. Tous les intervenants possèdent un historique retraçant leurs contributions, parfois sur plusieurs années.

La quatrième capture d'écran présente l'archivage du site de la communauté par archives.org, ici celui de l'activité du forum de discussions. Il permet de « remonter le temps » et de retrouver les anciennes versions des sites (par des liens derrière les dates), ce qui permet de dater certaines évolutions et d'obtenir certaines statistiques ponctuelles. On ne saurait trop insister sur *l'insuffisance* de ce type d'outil de collecte systématique par la communauté scientifique pour aborder des problématiques fines, qui montre donc la portée des analyses que nous proposons ici à partir des précédentes archives internes à la communauté.

LA NECESSITE D'UNE ANALYSE SPECIFIQUEMENT HISTORIQUE DES COMMUNAUTES

[8]Ainsi, « http://comox.textdrive.com/pipermail/wp-docs/2005-March/000309.html » désigne, dans l'entrepôt le courriel n° 309, en mars 2005, numéroté dans l'ordre de son émission.

Remarques sur les méthodologies existantes

La démarche souvent évoquée (voire dominante) mobilisée pour l'analyse qualitative des données sur la Toile serait « l'ethnographie virtuelle » (Hine, 2000). Notre thèse est qu'une telle approche, *dans une certaine mesure pertinente*, n'est pas suffisante. Il est impossible de toujours suivre en détail l'activité intense quotidienne dans la communauté, comme un ethnologue. L'approche historique qui suppose de sélectionner les sources est plus *efficiente*.

Nous avons tenté par ailleurs d'effectuer un travail de codage des données, dans une optique classique de type théorie ancrée (« grounded theory »), comme d'autres chercheurs. Là encore, les limites de cette approche dominante nous ont orienté en chemin vers une approche historique. Elle ne permet pas de prendre en compte la dimension dynamique des phénomènes, de tester des hypothèses. Un traitement informatique des données était indispensable ; cependant, les logiciels de codage [9] sont conçus pour cette approche usuelle, et nous y avons renoncé. Nous avons effectué entre autres des démarches de codage, avec pour objectif de pouvoir retrouver les fragments de textes pertinents. Sur ce point, le bricolage avec les outils bureautiques est satisfaisant, en attendant des outils prenant en compte les spécificités de l'approche historique[10].

Enfin, l'intérêt de l'histoire pour examiner les comportements réels et les processus est évident. De même que la fouille des poubelles est plus instructive que les entretiens pour comprendre les habitudes de consommation, de même l'examen

[9] Nous pensons à ceux qu'on appelle « CAQDAS » (Computer Aided Qualitative Data Analysis Software).

[10] Le récent colloque AnaLogiQual2008 (voir : http://analyses.ishs.ulg.ac.be/analogiqual/presentation.html) montre que beaucoup de choses restent à faire en la matière. *Nous avons d'ailleurs été confrontés à trois difficultés qu'il évoque* : l'analyse de données diachroniques, le détournement d'outils d'usage courant et le traitement de matériaux de type numérique.

des « dumps » des wikis ou des archives en dit certainement plus long que les affirmations orales ou les enquêtes par questionnaires électroniques. Enfin, nous ne saurions trop souligner l'intérêt d'une approche non-intrusive, notamment lors des périodes d'activité moindre de la communauté, où le risque de modifier les comportements des participants est présent. L'éthique de la recherche autorise l'analyse de ces archivages publics.

Eléments de critique des sources à portée générale

Ces archives ont un statut ambigu dans notre contexte, qui n'est certainement pas un cas particulier: *une pratique d'archivage n'est pas neutre,* même lorsqu'il est automatique. Les archives publiques sont considérées comme la mémoire *vivante* de la communauté, de ses choix et de ses erreurs, et les nouveaux venus peuvent être invités à parcourir le passé d'une communauté pour *apprendre*, ce qui aurait une influence sur les modes d'interactions et les contenus ajoutés. Le nombre d'interventions d'un contributeur sur la liste comporte une signification politique : plus un participant parle sur cette agora électronique, plus il est susceptible d'avoir du pouvoir. Ce processus demande examen, qu'il soit explicitement quantifié, comme sur la liste des « hackers », ou non. On observe aussi que certains courriels ne sont conservés que par morceaux, retrouvés plus loin dans l'archives. Les choses sont plus sérieuses pour d'autres formes d'interactions encore plus informelles, comme celles sous forme de dialogue par canal IRC (« chat ») qui ne sont archivées que partiellement : si les réunions hebdomadaires « en ligne » planifiées semblent complètes et sérieusement archivées, l'analyse de ces mêmes archives révèle l'existence d'un canal de discussion spécifique à l'organisation du Codex dont le contenu n'a hélas pas du tout été conservé[11]. La chose est aussi à signaler pour les interactions entre deux contributeurs non archivées par la liste de discussion, ce qui fait qu'il manque une partie des données concernant des décisions effectuées en privé. Lorsque les participants *anticipent* que leurs propos seront entreposés en public, ils peuvent tout à fait interagir par correspondance privée.

Si des biais d'énonciation sont observés, ce qui nécessite la comparaison des propos de chacun, en revanche le nombre important de participants permet de *multiplier les points de vue* et de limiter les biais. Par ailleurs, les listes de discussions « tamisent » les aspects significatifs d'une activité qu'elles coordonnent.

Un des avantages du support électronique est qu'il rend caduque la question de l'authenticité des sources, un des aspects de la critique externe. La critique interne conserve toute sa pertinence pour l'interprétation des textes. Sur ce point, comprendre ce qui se passe dans la communauté supposent une *connaissance minimale* en informatique, mais pas forcément la maîtrise d'un langage de programmation.

La sélection des sources dont la visibilité n'est pas immédiate et les risques de pertes de données.

L'analyse rigoureuse d'une telle communauté suppose un repérage de toutes les sources disponibles et intéressantes, ce qui n'est pas forcément évident, surtout sans guide des archives complet, guide qui est aussi un produit intéressant d'une recherche (nous fournissons quelques éléments en annexes).

Pour prendre un exemple parlant, trois des premières archives (« docs », « hackers », « CVS ») n'apparaissent plus dans l'entrepôt officiel de la communauté et ont dû être découvertes par déduction et croisement de données à partir de la connaissance des archives

[11]Source:http://codex.wordpress.org/IRC_Meetups/2006/January/January04RawLog

précédentes. Si elles ne sont plus d'importance pour la communauté, elles sont indispensables en revanche pour le chercheur travaillant sur la genèse de la communauté. Cette nécessité de trouver les sources indispensables cachées doit être une priorité de chercheurs confrontés à une analyse du « déploiement du numérique ».

La nécessité de l'analyse de type historique *n'est jamais très loin* à cause du risque permanent de pertes de données : le premier wiki de la communauté, accessible au début de notre recherche, n'est plus accessible sur les serveurs actuellement, une désagréable surprise... La seule façon à présent de comprendre les débuts du déploiement d'une solution wiki comme système d'information est d'extraire et analyser les traces présentes dans les listes de discussions correspondant à cette période de l'histoire…dans les archives cachées mentionnées ci-dessus !

La taille imposante des archives oblige le chercheur à des *arbitrages*, certaines sources sont trop importantes pour être traitées « à la main ». Par exemple, le forum des utilisateurs comporte plus de 100000 fils de discussions et l'extraction de toutes les données pertinentes supposerait un travail de programmation.

QUELQUES QUESTIONNEMENTS ECONOMIQUES OU L'INTERVENTION D'UNE METHODOLOGIE HISTORIQUE EST INDISPENSABLE

Ces archives peuvent être traitées de plusieurs façons, qualitative, quantitative, ou au moyen d'une analyse de réseaux. Une *même source* peut subir au besoin plusieurs traitements différents de cette sorte. Une pratique tout aussi fructueuse est de *croiser les sources* : par exemple, il est possible de comparer ce que disent les informaticiens et les contributeurs à la documentation pour comprendre les débuts de celle-ci, en particulier pour confirmer l'idée que l'échec des débuts est *continuellement* dû à une problématique d'organisation et non d'incitation.

L'analyse économique des routines, fondement de la gouvernance de la communauté, oblige à une démarche historique

Une problématique d'économie évolutionniste par excellence est l'analyse de l'émergence, de la mutation et diffusion des *routines*. Elle est incontournable pour aborder les problèmes de régulation informelle de la communauté travaillant sur la constitution de la documentation. L'apport d'une *démarche historique* (analyse à partir de traces de la spécificité d'un phénomène de nature temporelle) est décisif pour observer et analyser les formes d'adaptations des pratiques, les formes d'exploration/exploitation individuelles et collectives, les solutions adoptées, celles qui auraient pu l'être.

La difficulté de ce type d'analyse est de faire la généalogie des routines significatives à partir d'une source différente des pratiques elles-mêmes : l'activité se fait sur le wiki, la liste de discussion archivée est un organe de régulation qui « filtre » les thématique abordées. C'est là la source indirecte principale, et l'analyse des traces de pratiques qui y sont évoquées est d'emblée de nature historique. Elle suit un protocole. Elle se fait par codage thématique des courriers électroniques, après repérage des passages décrivant la sélection de pratiques, des décisions et arbitrages, des réflexions de type « coût-avantage ». Nous cherchons alors quelles sont les heuristiques[12] employées alors par la communauté et les formes de légitimation sous-jacentes. Un travail de hiérarchisation de leur importance relative est évidemment nécessaire.

La communauté a en effet beaucoup hésité en matière de pratiques de documentation. En particulier, dans quelle mesure fallait-il adopter les pratiques de Wikipedia ?

[12] Une heuristique est, pour faire simple, une voie de résolution d'un problème.

"I agree that no two wikis are the same, and that it would be incredibly irresponsible of us to just copy someone else. But any decent wiki must at least consider what Wikipedia does, or risk repeating history. Please do not discount what Wikipedia does as a valuable model to learn from: turning a blind eye to the methods of such a successful site would be equally as irresponsible as blindly copying it."[13]

Le problème est donc d'examiner et caractériser la trajectoire *spécifique* du projet de documentation dans ce contexte, et en particulier les routines de formatage de l'information adoptées. L'analyse historique des débuts de la documentation montre effectivement des formes de tâtonnements, d'essais et d'erreurs avant routinisation, y compris l'idée d'utiliser un wiki, en fonction de contraintes assez complexes, qu'il s'agisse des questions de taille, de la tonalité, des thèmes à aborder, des modes de structuration (« task-based », « bush-like », ou « livresque », plus familière), afin de satisfaire des populations d'utilisateurs hétérogènes.

L'analyse des routines permet de mettre en évidence *deux types de principes structurels de justifications* qui rendent cohérentes les pratiques observées, deux « conventions » au sens de l'économie des conventions. Toutes les routines se rapportant à Wikipedia ou au monde de l'open-media relèveraient de la « cité de l'opinion », alors que celles qui proviendraient du monde des « hackers » relèveraient de la « cité inspirée » [14] . L'articulation entre deux voire plusieurs cités est l'un des axes d'analyse important des routines. Il est remarquable de constater que les routines justifiées par les principes de la « cité de l'opinion » se sont révélés économiquement viables (« edit at will »), *moyennant adaptations*. Ainsi, les premières contributions d'un utilisateur sur un nouvel article se font sur leur espace personnel (« user space »), sur les marges du wiki, et non plus sur l'espace public (« main space ») du système d'information. Les administrateurs décident plus tard de la parution (« release ») sur l'espace public, une pratique qui provient de l'open-source. Ce type de routine clairement *hybride* a été mis en place après l'échec du premier wiki, afin de limiter la désorganisation des pratiques éditoriales et garantir la qualité de l'information. Ce fragment du wiki n'aurait *jamais* pu être énoncé sur Wikipedia :

« ==Hey, Carthik!==
Why is my new article - DRAFT ARTICLE now sitting in the main stuff????? I'm in panic mode, folks. Please REMOVE IT AT ONCE and all redirects. It is NO WAY ready for prime time!!!!! [[User:Lorelle|Lorelle]] 19:16, 22 Mar 2005 (UTC) »[15]

Une approche historique est évidemment consubstantielle à la mise en évidence de phénomènes de « dépendance par rapport au chemin » (« path dependency »), où les éléments d'ordre technologiques (« technological interelatedness ») et organisationnels sont liés. C'est le cas par exemple de la diminution d'efficacité due à l'absence de mise à jour du logiciel. La décision n'aurait pas été prise pour des raisons complexes, à la fois erreur de paramétrage, manque de compétences, attentes de décision de l'équipe de documentation vis-à-vis du chef de projet qui lui-même s'attendait à ce que la décision soit prise de façon démocratique au niveau de l'équipe.

L'analyse historique permet enfin de mettre en évidence « en creux » les routines qui n'ont pas émergé, mais qui auraient pu apparaître ou se révéler nécessaires. L'examen des archives montre par exemple qu'il n'existe pas de pratiques incitatives notables et d'infirmer l'idée

[13] Source : http://comox.textdrive.com/pipermail/wp-docs/2005-March/000309.html

[14] « Code is poetry » est le slogan de WordPress. Les programmeurs du libre conçoivent leur activité comme un art et un amusement, aussi exotique que cela puisse paraître.

[15]Source :http://codex.wordpress.org/index.php?title=User_talk:Carthik&diff=next&oldid=9862

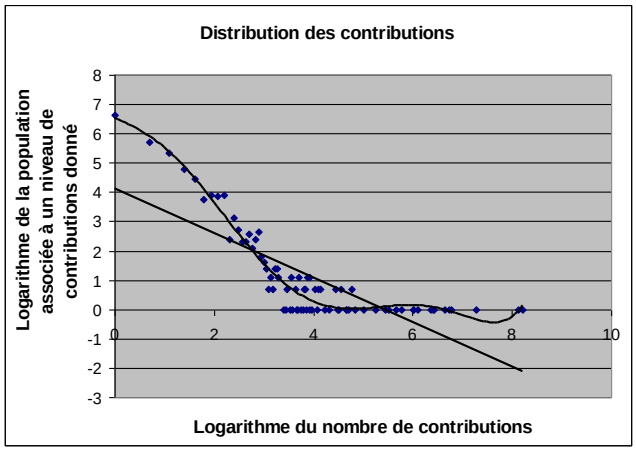
préconçue qu'il aurait existé un problème d'incitation non surmontable à la constitution du bien collectif.

Enfin, l'histoire impose de dire ce que les archives n'ont permis de comprendre : certaines évolutions de routines, observées dans les archives, comme celle de placer les fonctionnalités additionnelles (comme les « plugins ») sur un autre système d'information que le wiki, n'ont pu être comprises clairement à partir de nos sources malgré un travail de croisement.

Analyser les comportements réels de contribution des participants à ce bien collectif suppose une démarche « archéologique ».

A notre sens, *les sources les plus difficiles à traiter* sont les historiques des wikis (« dumps »), montrant *toute modification* du système d'information par les contributeurs. Comprendre l'émergence d'un bien collectif (par addition de contributions) dans le cadre d'une démarche qu'on pourrait qualifier de « *génétique* » (bien différente de l'approche sociologique par entretiens) est particulièrement ardu « à la main ». Autant il est possible d'analyser, voire de coder (Jones, 2008) des historiques d'articles, autant il est ingrat d'analyser les comportements individuels et les contributions de chacun « clic par clic ».

Une approche quantitative des contributions comporte *paradoxalement* moins de difficultés dans une certaine mesure, des outils internes du logiciel de partage de connaissance permettant de compter les contributions par catégories. Etre patient permet de tester certaines hypothèses, par exemple l'existence d'une loi de puissance[16] :



L'analyse des motivations des contributeurs peut commencer à partir de là : à partir du repérage des contributeurs les plus significatifs, une mise en relation avec des données qualitatives des autres parties de la communauté permet de mieux comprendre les raisons de leurs contributions : dans quelle mesure le volontariat et l'altruisme sont-ils les causes essentielles des contributions ? Voici, extraite du début de l'historique du wiki la motivation originale du futur administrateur et contributeur majeur MichaelH:

"I'm new to the world of WordPress and MediaWiki and am just awe-struck by the contributions made by all the talented people. If I can contribute just one iota of what they've contributed..."[17]

Par ailleurs, la gouvernance de la communauté, au-delà des routines, suppose en effet une forme de management et de travail d'équipe permettant *d'aboutir à la distribution de contributions observée ci-dessus*. La gouvernance, au sens étymologique de « navigation », consiste ici à maintenir le cap entre deux écueils :

"I think that for now, and perhaps forever, the documentation needs to have some tighter control. Development of code is in the hands of a small group of developers; this is necessary to keep focus and to get the job done. I think

[16] La population des contributeurs se caractérise par un grand nombre de contributeurs éphémères et très peu de contributeurs significatifs, avec un certain nombre de situations intermédiaires entre ces deux populations.

[17] Source:http://codex.wordpress.org/index.php?title=User:MichaelH&diff=prev&oldid=11280

that a parallel should be drawn to documentation. Many hands makes small work, but too many hands makes chaos […]I'm concerned that "restriction" on access to working on documentation may be perceived by some as "elitist" or that the docs crew are somehow "special" or part of a "clique" or something."[18]

Ce fragment datait des débuts de la documentation, décembre 2003. Un autre d'un des administrateurs de la documentation datant d'octobre 2006 et énoncé dans un autre dépôt d'archives est un « incident critique » intéressant : il suggère *une discontinuité*, la régulation de la documentation serait tombée dans le deuxième travers :

"Soon we had lots of good folks jumping in to help out, and they were indeed wiki savvy. However, what happened was that soon these people were basically doing everything, while people like me wound up being relegated to the sidelines.

Right now the support and Codex are dominated by a small group. Again, this is not an attack or judgement, it's simply the reality of a lot of work to be done, and only a core group of folks with the occasional help from a lot of others.

I believe that WP has grown to the point where there needs to be some "middle management" and those folks who would so choose could become involved towards managing support and managing documentation almost as distinct open-source projects of their own."[19]

Que s'est-il exactement passé en matière de gouvernance des comportements effectifs de contributions entre ces deux points du temps ?

Problématiser les modes de collaboration suppose un croisement des méthodes d'analyses de nos sources.

Répondre à cette dernière question suppose d'analyser les interactions entre les contributeurs qui apparaissent dans les listes de discussions : *l'analyse des réseaux est une pratique (nouvelle) d'historien* [20] . Le schéma suivant montre le résultat d'une extraction des données des 2575 courriels de la liste de coordination de l'activité de documentation de novembre 2003 à juin 2007 [21] . Elle montre clairement dans quelle mesure on peut observer la présence d'une forme de collaboration distribuée entre plusieurs contributeurs significatifs au niveau du « noyau central » de la population concernée (les points rouges entourés d'un noir indiquent la densité des interactions)[22].

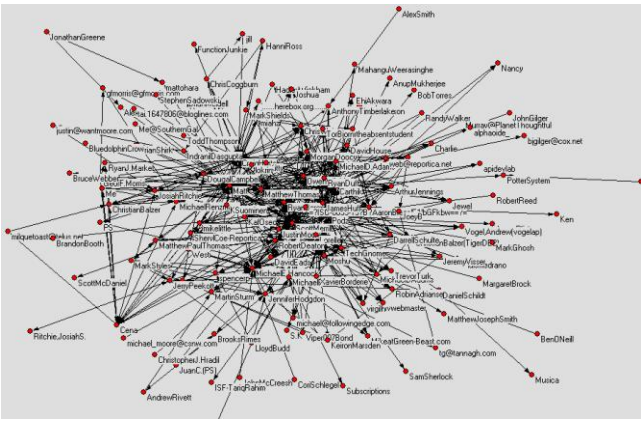
[18] Source :http://comox.textdrive.com/pipermail/docs/2003-December/000082.html

[19] Source : http://comox.textdrive.com/pipermail/wp-forums/2006-October/003788.html

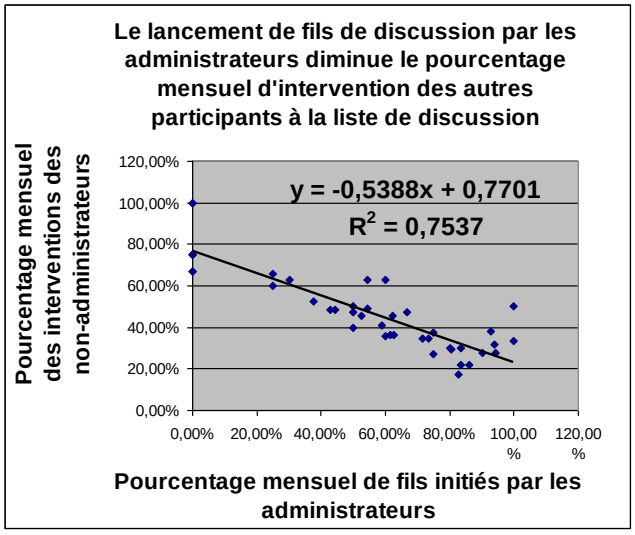
[20] Nous renvoyons par exemple au Pôle méthodologique "Analyse des données relationnelles" http://www.ehess.fr/ldh/Atelier_ADR/atelier_ADR.html

[21] Une telle démarche de traitement des archives suppose un petit protocole d'extraction des données à partir des fichiers « .txt » (« parsing ») écrit en Python qui a été mis en oeuvre par Jean-Philippe Cointet, doctorant en sciences cognitives au CREA. Par ailleurs, le traitement historique des données mentionnées suppose un nettoyage des sources à partir de la connaissance du contexte, en particulier pour éliminer les « doublons » : certains participants apparaissent sous deux voire plusieurs dénominations différentes, *un problème récurrent* d'analyse des communautés en ligne que tout chercheur doit prendre en compte.

[22] Des analyses plus poussées seraient possibles, par exemple examiner avec le logiciel Pajek plusieurs indicateurs de centralité possibles.



Des traitements *statistiques* complémentaires des archives sont indispensables pour dépasser ce point de vue statique, pour par exemple comprendre par une régression dans quelle mesure la structuration de la liste de discussion obéit à un mécanisme *d'attachement préférentiel* : plus les administrateurs commencent à intervenir sur la liste de discussion en lançant un fil, plus il y a de chance qu'un autre administrateur lui réponde et moins de chance qu'un participant non administrateur intervienne.



Ces analyses permettent de mieux comprendre le phénomène de « clique » ressenti par certains membres de la communauté. De telles constatations *nuancent* le tableau plus démocratique de la régulation de la communauté qui apparaît souvent à la lecture qualitative de la prise de décision dans les archives. On peut cependant s'interroger sur l'*effet réel* de ce type de phénomène sur la viabilité de la communauté qui a créé cependant un outil relativement efficace, ce à quoi tente de répondre la dernière sous-question que nous aborderons ici.

L'analyse de l'efficacité du système d'information nécessite une analyse des traces, une méthodologie historique.

Nous avons eu l'intuition qu'en cherchant dans les archives des listes de discussions servant à la régulation des autres partie de la communauté, par exemple celles de l'activité d'informatique (programmation, test..) et celles du forum d'aide aux utilisateurs, nous obtiendrions d'une façon inductive des indices sur l'efficacité, l'appropriation, la qualité, les usages, etc. de la documentation. Le résultat *comparatif* auquel nous avons abouti par ce travail qualitatif et exploratoire est que l'efficacité est nuancée en fonction du contexte, en particulier du type d'usagers concerné.

Le protocole adopté a été de filtrer les courriels pertinents en les sélectionnant à partir de plusieurs mots-clefs, ce qui a permis une « saturation » du corpus de courriers électroniques pertinent[23]. Cette démarche de tri dans les archives a permis de constituer des bases de données comportant en tout près de 1400 éléments pertinents (extraits à partir de plus de 30000 courriels). Enfin, un codage thématique a permis de *catégoriser* les fragments pertinents mais aussi de repérer des passages importants. Une telle procédure a permis d'établir quelques régularités économiques intéressantes.

Tout d'abord, pour les informaticiens, elle ne permet pas de confirmer que l'information proposée par le système est de mauvaise qualité, alors que paradoxalement on attendrait le contraire, étant réticents à documenter leur propre production, ce qui est moins amusant et

[23] L'extraction s'effectue à partir des fichiers « .txt » mensuels.

prestigieux que la programmation. Cela n'exclut pas des formes significatives de mécontentement mais dont il est difficile de prouver le caractère décisif (la communauté pouvant réagir très vite à des disfonctionnements ponctuels). Une approche plus quantitative de la présence mensuelle des fragments étudiés montre en tout cas la *permanence* et la régularité de l'usage du système d'information durant toute la période d'étude dans ces populations de contributeurs.

En revanche, en ce qui concerne le forum d'aide aux novices, l'usage du wiki n'est pas évident, n'a pas atteint certains de ses objectifs, et suppose des formes non triviales de régulation. Plusieurs fragments montrent clairement que l'usage du système d'information ne fait pas consensus :

"I mostly agree with Podz on this. Codex is a huge and valuable resource, but for new users in particular, it's not the easiest thing to use."[24]

On observe plutôt une complémentarité entre les deux systèmes d'information, entre savoir explicite, codifié, sur le wiki, et savoir tacite sur le forum, ce que montre ces fragments de ces contributeurs critiques :

"What I actually do not want are codex links all over the place - it's rare that you can home directly in on a codex topic that helps directly and Codex is linked from many places (like every WP blog). What is not linked is all the gold that is in the forums."[25]

"This exercise is not about replacing Codex, nor is it about removing anything from the forums. It is about using knowledge better."[26]

La façon d'orienter les utilisateurs vers les bons endroits de la documentation est fondamentale, suppose une *régulation des usages* (informatisée ou non), l'outil ne se suffit pas à lui-même, ce qui a demandé des formes d'apprentissage, d'essais et d'erreurs, et des débats. Nous citons ce passage ironique :

"Much better. Only problem I encountered was searching for something I KNEW was in the forum, but got sent to a no results page because the default choice for search is docs. That's great from the aspect of "real people search" though! Nice. Thank you VERY much.
V »[27]

Puis, les fragments tirés de ces archives montrent que le système d'information est davantage une affaire de connaisseur et son usage demande une expertise (on trouve beaucoup plus de fragments codés « expert », quelles que soient les archives), voire des formes d'intermédiation de connaisseurs, observées même chez les hackers. Si la chose ne pose pas de problèmes notables pour les informaticiens, elle est source de frustrations pour les utilisateurs novices.

CONCLUSIONS

Nous avons donc montré qu'une analyse sérieuse des dynamiques et de l'efficacité d'une communauté suppose en fait nécessairement une approche historique par comparaison avec les autres approches existantes. Connaissance des événements sur la moyenne durée, mise en évidence de périodes, de ruptures, ou au contraire de permanences structurelles, ainsi que des phénomènes de « dépendance par rapport au chemin », connaissance des archives, importance de la critique des sources, spécificité de l'analyse des traces et des comportements effectifs (motivations, routines), mise en évidence des spécificités des phénomènes étudiés, et de leur complexité, discussion sérieuse sur ce qu'on peut savoir ou ne pas savoir à partir des données, et enfin attention aux « humbles » au rôle économique essentiel, tout cela fait partie de la tradition et du savoir-faire de l'historien.

[24] Source : http://comox.textdrive.com/pipermail/wp-forums/2005-December/000947.html.

[25] Source : http://comox.textdrive.com/pipermail/wp-forums/2005-December/000943.html

[26] Source : http://comox.textdrive.com/pipermail/wp-forums/2005-December/000956.html

[27] Source : http://comox.textdrive.com/pipermail/wp-forums/2006-March/001825.html

L'analyse historique comporte un avantage absolu pour traiter de questions nuancées et complexes.

Notre contribution a pour but d'inciter par l'exemple d'autres à poursuivre dans cette voie pour analyser le « déploiement du numérique » et à prendre au sérieux l'analyse historique qui va au-delà du concept « d'étude longitudinale » usité en sciences de gestion. *Le risque de pertes d'archives rappelle aussi qu'il y a urgence.*

Bibliographie

Banister Frank, 2002, "The Dimension of Time: Historiography in Information System Research", *Electronic Journal of Business Research Methods*, 1-1:1-10.

Conein Bernard et Latapy Matthieu, 2008, « Les usages des réseaux de communication électroniques » : le cas de l'Open-Source, *Sociologie du Travail*, 50 :331-352.

Dhérent Catherine, 2002, *Les archives électroniques, manuel pratique*, Paris :La Documentation Française 104.

Den Besten Matthijs L., Rossi Alessandro, Gaio Loris, Loubser Max et Dalle, Jean-Michel, 2008, « Mining for Practices in Community Collections: Finds from Simple Wikipedia", *Open Source Development, Communities and Quality*, pp. 105-120

Kieser Alfred, 1994, "Why Organization Theory Needs Historical Analysis – And How This Should Be Performed", *Organization Science*, 5-4:608-620.

Hine Christine, *Virtual Ethnography*, London:Sage Publishers, 2000.

Flinn Andrew, 2007, "Community Histories, Community Archives: some opportunities and Challenges", *Journal of the Society of Archivists*, 28-2:51-176.

Jones John, 2008, « Patterns of Revision in Online Writing: A Study of Wikipedia's Featured Articles », *Written Communication*, 25:262 - 289.

Langner Irene, 1997, "An introduction to Internet Mailing-list Research", Document de travail.

Lesourne Jaques et al., 2002, *Leçons de microéconomie évolutionniste*, Paris :Odile Jacob.

Lorenz Edward, 2001, "Models of Cognition, the Contextualisation of Knowledge and Organisational Theory", *Journal of Management and Governance*, 5:307-330.

Rygiel Philippe, 2005, « Des archives numériques sans historiens ? Un point de vue », *Matériaux*, 79 :11-13.

Simon Herbert, 1999, "The many shapes of knowledge", *Revue d'économie industrielle*, 88:23-39

De Vaujany François-Xavier, 2006, "Between eternity and actualization: the co-evolution of the fields of communication in the Vatican," MPRA Paper 4082.

Wagner Christian, Majchrzak Anne, 2007, Enabling Customer-Centricity Using Wikis and the Wiki Way, *Journal of Management Information Systems*, 23-3:17-43.

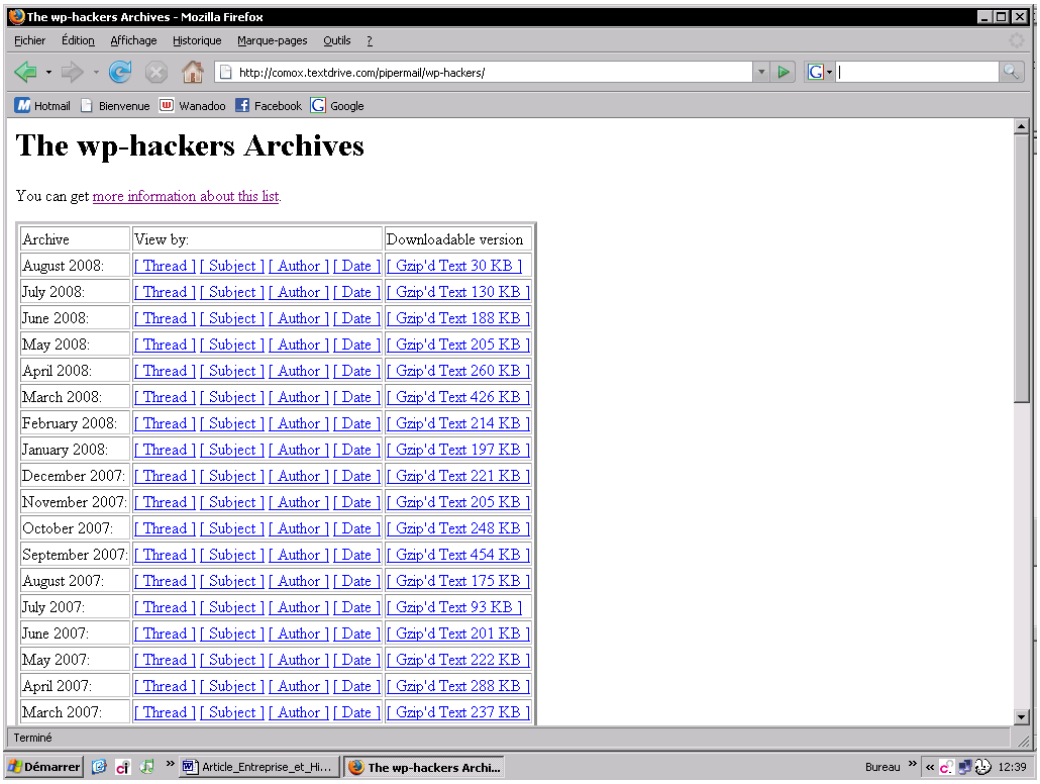
Yin Robert K., *Case Study Research, Design and Methods*, Sage Publications, Thousand Oaks, 2003, 181 p.

Liens vers les archives en ligne

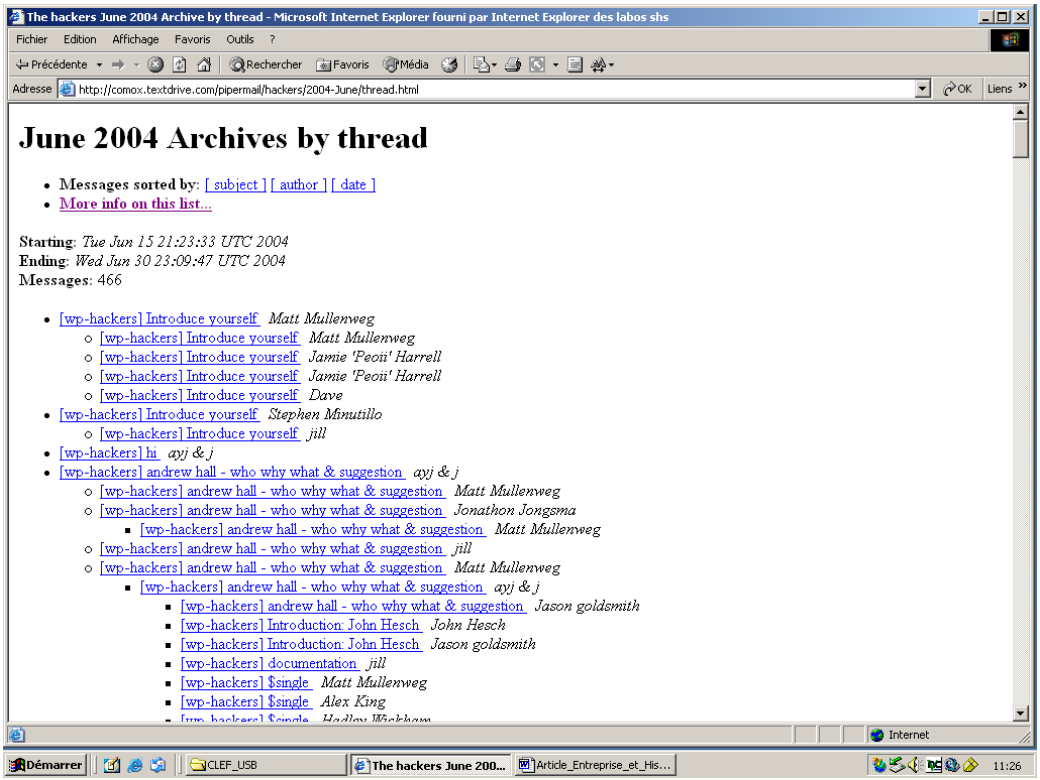
- Le Codex : http://codex.wordpress.org/Main_Page
- Entrepôt des listes de discussions : http://lists.automattic.com/mailman/listinfo
- Les anciennes listes de discussions: http://comox.textdrive.com/pipermail/hackers/ http://comox.textdrive.com/pipermail/docs/ http://comox.textdrive.com/pipermail/cvs/
- WebArchives : http://web.archive.org/web/*/http://wordpress.org/
- Les « minutes » de dialogue sur IRC : http://codex.wordpress.org/IRC_Meetups

Captures d’écran d’archives

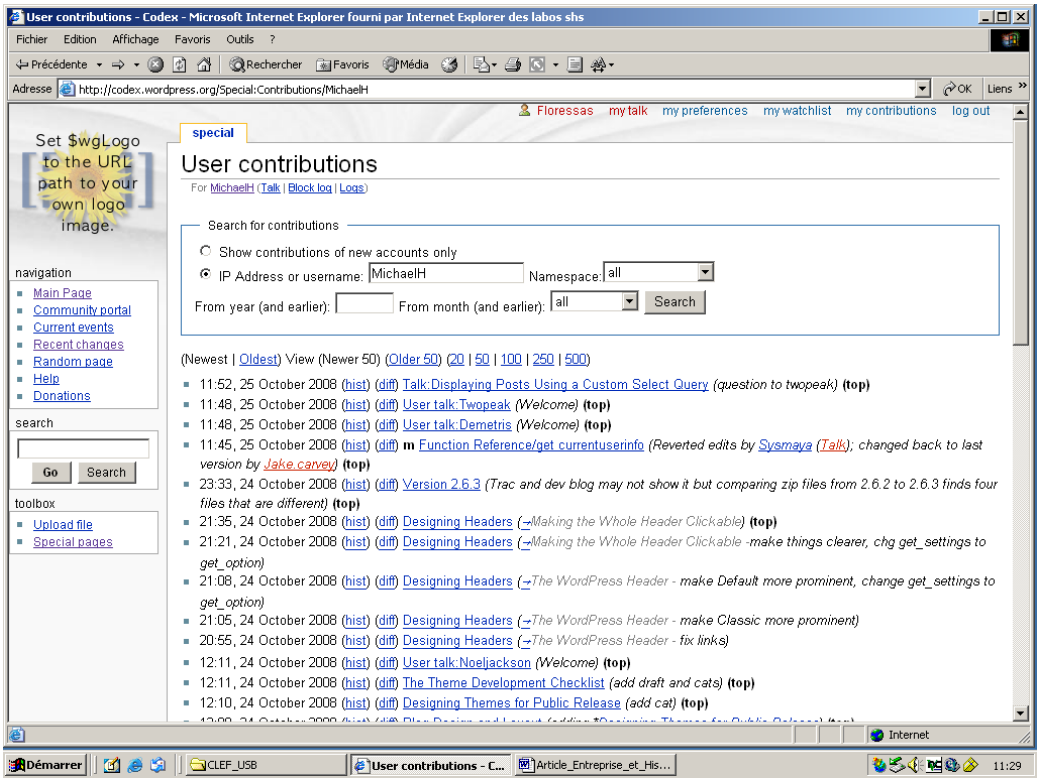
Système d’archivage des listes de discussion de « wp-hackers » :



Archives de « hackers », mois de juin 2004, présentées par fils de discussion



Historique des contributions de l'administrateur Michael Hancock



Archives du forum de la communauté sur Internet Archives (http://web.archive.org/web/*/http://wordpress.org/support/ , consulté le 01/03/09)