Statistiques sur Wikisource

dans le cadre du partenariat

Bibliothèque nationale de France – Wikimédia France

version 1.0

24 octobre 2010

Introduction

Ce document présente une vue synthétique de l’avancement de la correction des livres téléversés sur Wikisource dans le cadre du partenariat Bibliothèque nationale de France – Wikimédia France. Il a vocation à reprendre périodiquement (tous les trois mois environ) les mêmes statistiques afin de visualiser l’évolution temporelle, et les futures versions de ce document pourront en conséquence être revues afin d’ajouter de nouveaux éléments.

Les données brutes ayant servi à l’élaboration de ce document accompagnent celui-ci et seront conservées au moins deux ans par Wikimédia France. Il s’agit des extractions de la base de données de Wikisource (*dumps* partiels relatifs au corpus du partenariat) ainsi que des fichiers générés à partir de ces derniers pour extraire des données statistiques. Du fait de cette conservation, il sera possible de regénérer d’autres statistiques même pour d’anciens *dumps* ; notons toutefois que, du fait du mécanisme d’historique du wiki, la conservation d’anciens *dumps* pourrait sembler superflue (si on met de côté les opérations de suppression et de renommage de page) mais cette conservation permettra de retrouver ultérieurement d’anciens résultats pour des données d’entrée identiques.

Ce document comprend six parties :

- description rapide de Wikisource et de l’évolution globale des Wikisources ;
- description des notations et notions préliminaires ;
- collecte de données sur les pages, livres et contributeurs ;
- statistiques : sur les livres, sur le travail demandé, sur les contributeurs ;
- commentaires sur chaque livraison ;
- notes de versions.

Description rapide de Wikisource et de l’évolution globale des Wikisources

Wikisource est né fin 2003 dans la troisième lignée des projets Wikimedia après Wikipédia (2001) et le Wiktionnaire (2002). Son objectif est de retranscrire sous forme numérique les livres publiés, et, pour des questions de droit, en se restreignant aux livres passés dans le domaine public ou aux livres sous licence compatible avec la licence libre CC-BY-SA. Wikisource est donc une bibliothèque numérique libre.

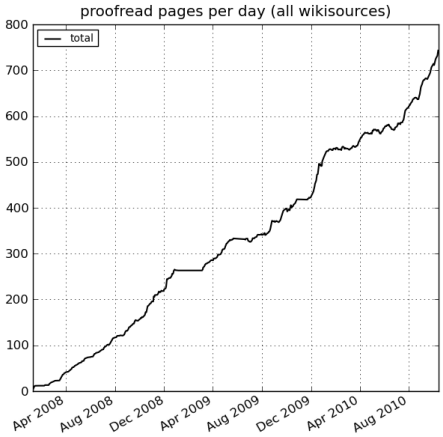
Son mode d’édition est le wiki, à l’image des projets Wikimedia, ce qui signifie que tout internaute peut contribuer à la relecture des livres sur la base du bénévolat ou encore à y téléverser des œuvres libres.

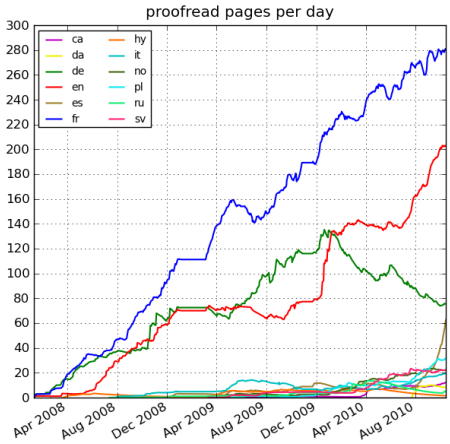
Initialement regroupé sur un seul site multilingue, Wikisource s’est divisé en communautés linguistiques en 2004, étant généralement admis qu’un livre dans une langue donnée doit aller sur la Wikisource dans cette langue, ou à défaut sur le site multilingue.

En tant que bibliothèque numérique, Wikisource a dû se doter d’outils permettant la gestion d’un nombre important de livres, même si ces quantités restent infimes au regard des volumes gérés par les bibliothèques nationales ou du monde de l’édition en général. Ainsi, le format DjVu a été adopté comme format courant (standard ?) de stockage des scans de livres, des métadonnées sont ajoutées sur la page de description du livre, et la relecture des livres est organisée page par page. Les pages comportent un indicateur d’avancement/de fiabilité : gris = page sans texte ; bleu = page ayant un problème ; rose = page non corrigée ; jaune = page corrigée par un contributeur ; vert = page relue et validée par un deuxième contributeur.

En suivant le nombre de pages relues et validées, on peut constater que le nombre de pages validées par jour et le nombre de pages relues par jour augmente linéairement depuis 2008 (date à laquelle le système de notation/statut des pages a été mis en place) pour arriver en septembre 2010 à 800 pages relues par jour (toutes Wikisources confondues) et 300 pages validées par jour. Une interprétation de cette linéarité peut être, outre que le système de notation s’est fait connaître progressivement, que le nombre de contributeurs augmente linéairement et/ou que les contributeurs passent de plus en plus de temps à relire les livres.

Pour suivre l’évolution des statistiques globales des Wikisources, voir
http://toolserver.org/~thomasv/





Description des notations et notions préliminaires

Notations

Les dates sont données en epochs, décomptant le nombre de secondes depuis 1970, ce qui permet des manipulations faciles des dates (additions, soustractions, moyennes, etc.), bien que cela complique fortement la lecture directe des dates.

Les statuts des pages sont les suivants : 0=page non créée ; 1=bleu ; 2=gris ; 3=rose ; 4=jaune ; 5=vert.

Notions préliminaires

Distance de Levenshtein : cette distance mesure la distance entre deux textes (dont les longueurs ne sont pas forcément égales) ; ou plus précisément le nombre minimal d'opérations (ajout d'un caractère, suppression d'un caractère, remplacement d'un caractère) qu'il faut effectuer pour transformer le premier texte en le deuxième.

La distance de Levenshtein entre deux textes sera égale à 0 si (et seulement si) les deux textes sont identiques, et sera égale inférieure ou égale au nombre de caractères du plus grand texte. Pour deux textes données, elle sera supérieure ou égale à la différence de taille des deux textes.

Pour une page de texte classique, comprenant environ 600 caractères pour les poèmes, 1300 caractères pour un texte aéré et 2600 caractères pour des romans, on peut estimer que une distance inférieure à 50 ou 100 peut être corrigée à partir de l'OCR, mais il peut être préférable de ne pas utiliser l'OCR lorsque la distance est supérieure à 30% ou 40% de la taille du texte.

Voici quelques exemples permettant de se rendre compte des quantités.

Très bon OCR	
http://fr.wikisource.org/wiki/Page:Allais_-_À_se_tordre_:_histoires_chatnoiresques.djvu/165	
Distance = 26 = 7%	
Texte 1a (OCR initial)	Texte 1b-ter (page corrigée et retrait des en-têtes et déplacement des références)
ji JEUNE FILLE ET LE VIEUX COCHON II y avait une fois une jeune fille d'une grande beauté qui était amoureuse d'un cochon. Éperdument! J Non pas un de ces petits cochons jolis, t roses, espiègles, de ces petits cochons qui fournissent au commerce de si exquis jambonneaux. Non. Mais un vieux cochon, dépenaillé, ayant perdu toutes ses soies, un cochon dont le charcutier le plus dévoyé de la contrée n'aurait pas donné un sou. LA	LA JEUNE FILLE ET LE VIEUX COCHON Il y avait une fois une jeune fille d'une grande beauté qui était amoureuse d'un cochon. Éperdument ! Non pas un de ces petits cochons jolis, roses, espiègles, de ces petits cochons qui fournissent au commerce de si exquis jambonneaux. Non. Mais un vieux cochon, dépenaillé, ayant perdu toutes ses soies, un cochon dont le charcutier le plus dévoyé de la contrée n'aurait pas donné un sou.

Bon OCR	
http://fr.wikisource.org/wiki/Page:Tamizey_de_Larroque_-_Note_sur_le_poète_lectourois_Lacarry.djvu/9	
Distance = 296 = 14%	
Texte 2a (OCR initial)	Texte 2b-ter (page corrigée et retrait des en-têtes et déplacement des références)
4 épigraphe ~M ignotis, n'a jamais rencontré la plaquette de Lacarry. Enfin, M. Léonce Couture, qui a mis tant de zë)e et de soin à réunir les matériaux de i.on Esquisse d'une his<<MreMMet-aM'c de la Gascogne, et qui. dans ces pages savantes et charmantes, a donné un si exact dénombrement des prosateurs et des poètes de la province ecclésiastique d'Auch (1), n'a pas été plus heureux que ses devanciers (2). Seuls, de notre temps, Brunet et Du Mége ont cité le nom de Lacarry. Mais je me demande s'ils ont eu son petit recueil entre les mains et s'ils ne l'ont pas plutôt signalé sur la foi u'autrui. Pour ce qui concerne l'auteur du ~aMM~ ~M Libraire, mon soupçon est à demi justifié par la manière dont il a écrit le nom du poète; car il a séparé la première syllabe de ce nom des deux suivantes (La Carry), tandis que, dans le frontispice de la plaquette, ce nom est imprimé ~carr< comme l'est partout celui du docte jésuite Gilles Lacarry (du diocèse de Castres), né en 1605, mort en 1684, auteur de divers travaux estimés, notamment de)V/M<orM c~M!<an<m. etc. Autre motif de doute Brunet cite incomplètement le titre de l'opuscute, se contentant de ces cinq mots Pour le triomphe du SM<cy, alors que le titre rée) estcetui-ci Clytiepoitr le <M~/te dit Soucy A ~oK~HeM?' ~'em~' pr~eM<(5).–Quant à (1) Tous les admirateurs de l'érudition et du talent. de M. L. Contare espèrent bien qu'il transformera cette esquisse en un tableau dëSnitif. (2) Clytie manquait à la collection toulousaine de (eu mon vénérable ami M. le docteur Desbarreaux-Bernard; elle manque aux coDeetions des grands amateurs d'aujourd'hui, ainsi qu'à nos plus considérantes dépôts publics, de sorte que je serais tenté de saluer dans l'exemplaire de M. Oément-Sjmon un exemplaire umque. s'il n'y avait toujours imprudence à dédarer qu'un exemptaire est unique et qu'un document est inédit. #(3) Brunet indique seulement !e iieu de puMicatipn, en substituant dans le nom de ce lieu la lettre Z à la lettre S Toloze. Voici les indications fonrnies par le livret: roiose. por.J'. BoM~. tmprtmetf)- ordinaire du Roy,, devant le college de Foix, à l'enseigne S. J~'t 1636 (in-8* de 1~ pages,. Brunet ajoute qu'a la vente Veinant tetivMt atteignit le prix de23 fr. Un en donnerMt aujourd'hui plus de dix francs par page. Deux <)..s plus fervents et des plus savants MMiophUes de notre époque. M..tûtes Dukas et M.iEm'te Picot m:aNrmentau'dsn'pntcennurextstencedeCi!;Meqa9 narta reyetatM~~ j)f<MMt<<it't.tt)'<ttft. r	4 épigraphe : diis ignotis, n'a jamais rencontré la plaquette de Lacarry. Enfin, M. Léonce Couture, qui a mis tant de zèle et de soin à réunir les matériaux de son Esquisse d'une histoire littéraire de la Gascogne, et qui, dans ces pages savantes et charmantes, a donné un si exact dénombrement des prosateurs et des poètes de la province ecclésiastique d'Auch, n'a pas été plus heureux que ses devanciers. Seuls, de notre temps, Brunet et Du Mége ont cité le nom de Lacarry. Mais je me demande s'ils ont eu son petit recueil entre les mains et s'ils ne l'ont pas plutôt signalé sur la foi d'autrui. Pour ce qui concerne l'auteur du Manuel du Libraire, mon soupçon est à demi justifié par la manière dont il a écrit le nom du poète ; car il a séparé la première syllabe de ce nom des deux suivantes (La Carry), tandis que, dans le frontispice de la plaquette, ce nom est imprimé Lacarry comme l'est partout celui du docte jésuite Gilles Lacarry (du diocèse de Castres), né en 1605, mort en 1684, auteur de divers travaux estimés, notamment de l'Historia colonarium, etc. Autre motif de doute : Brunet cite incomplètement le titre de l'opuscule, se contentant de ces cinq mots : Pour le triomphe du Soucy alors que le titre réel est cetui-ci : Clytie pour le triomphe du Soucy À Monseigneur le premier président . — Quant à (1) Tous les admirateurs de l'érudition et du talent de M. L. Couture espèrent bien qu'il transformera cette esquisse en un tableau définitif. (2) Clytie manquait à la collection toulousaine de feu mon vénérable ami M. le docteur Desbarreaux-Bernard ; elle manque aux collections des grands amateurs d'aujourd'hui, ainsi qu'à nos plus considérables dépôts publics, de sorte que je serais tenté de saluer dans l'exemplaire de M. Clément-Simon un exemplaire unique. s'il n'y avait toujours imprudence à déclarer qu'un exemplaire est unique et qu'un document est inédit. (3) Brunet indique seulement le lieu de publication, en substituant dans le nom de ce lieu la lettre Z à la lettre S : À Toloze. Voici les indications fonrnies par le livret : À Toloze, par I. Boude, imprimeur ordinaire du Roy, devant le college de Foix, à l'enseigne S. Jean 1636 (in-8° de 16 pages. Brunet ajoute qu'a la vente Veinant le livret atteignit le prix de 23 fr. On en donnerait aujourd'hui plus de dix francs par page. Deux des plus fervents et des plus savants bibliophiles de notre époque, M. Jules Dukas et M. Émile Picot m'affirment qu'ils n'ont connu l'existence de Clytie que par la révélation du Manuel du Libraire.

Mauvais OCR			
http://fr.wikisource.org/wiki/Page:Abundance_-_Les_grans_et_merveilleux_faictz_du_seigneur_Nemo,_avec_les_privilèges_qu'il_a.djvu/5			
Distance = 2512 = 102%			
Texte 3a (OCR initial)		Texte 3b-ter (page corrigée et retrait des en-têtes et déplacement des références)	
I~MrceMe~urrd~OH~e tue cdtpourcequ~MMMurM '~o~nta~ toufto~e vtfoetMoumt 'nento ~t fcg VM~c.6ccf~;h<i.t. ~Oue vouk) voM6 q tevona oyc ~ctt~nc Moiccentctodtc Que oe nemo et ~efce fAt€t~ ~o~nen fcauroye Oire ~uy ma~ jEf~cnOtrayenco:cvMg Qm entre kc clerc3 eft commun ~c qui cit grant que fan 3 fauttc gcntcfinerMCtUc~~M~ ctt !?aukc ~< c!?o<e oont ~tcu efcondtc ~ce Ottctp~e~Mant il leur ofc Qnc a eutjcfcauotrnepp~rtcno~ Ouanr la fin ou monde feroit Ouon o~ou tH~etncntk tOMr ~ato tt tenr Me !o:o fane feiour Que ne nemo <y le fceuotc bien. ~e oie Mcent iHo vct t?o:a nemo fcic. ~~rd.4~. ~oïntne to::te ie croy et tien Que (c &un~ moye ic ~c cci~o~e ~cn parkr~pae ~c ne ofroye ~e nemo tout ce quon en trettue ~n chacun ttore Qui t~pp:eouc ~cfntcfamt~ot HnotteracopCC *nc!~o par vafn tangage et coptC ~eutc feduire kc gens en bïtef JËftcetn~crre envoie oc ïnefc~cf n~tïto vo feducaf thantb~vcrbtc. ~'HcmopcukbKn fane contredit lKcit)tcren f~icrecen OK a fce enncmye fang &OMbfer ~att~ qm~ le pu~nr rebouter. '~cmo un:mct9 audebu refittere. Xcturtc!.i<?. ~3rc faîne 3c!?a qMC nut ~Otne 'Me peult autTt bien bc~bn~ner ~c nuy~quo ootr p:~drc fon fo:~c Que nemo~m y veutf tonner. ~emrnoccmtcoo~anpof. 3o.!0 !Û~e(TctgncMrB poMrrjctCpdue ')~~r ce que tay 9tr cy betlue ~:tanr le h!~ oe pnceHc Umt nous ojtnc la vtc crernelle Quanr ton h$o:eu]ce]catncn ~era ccnH~otcreB Btnen. ~Xaus oco.	Hue nemo fut~c~fab~ jQua ~ng pourc qui auotf faut ~onna oc~ îmcfcc~ oc pain ~bcaH~ oc~tabtecnnct?e M ranr c~oH auarct ct~!c!?c ~mt ne tuy en voulotf oottncr ~.ne dt nctno abandonner Xu~ en voutur ptteutcmenr. ~r~t qutdc ~anpcrcuptcns recrea/ r oc tnïc~ que cadebanc ce mcnfa t)ùuti3:ct' nemo iUt oabac.~uc. nemo au~t paretHctncnc "t-- our !es mcnrcB et oeuoïr ~~ton tea o:OK~pcmt bien auou' ~ntj bigatnic ce me fetnbte Dcup fcnt'nce cfpoufcce cnfctnb~ 'nc.nùu hcenr &uaa v!:o:ee ~abcre ~rayemcf clercs c:c?<~ btc péfc nento~cr fon occeUcnce ~c cutde quit tcn efntcruetHe ~ar tuit ne fjicr c~ofc parctUe Ctnt peutr vendre et ahener Xe~ <:o:pa 'aùKf3~cn marc~ader X~ctncnrcra<a$Mtfc oc~ rrefo:~ oc f~tncfc cg~fc. '~c'no <!m ry:c~ Oittra~af.'nentO tncrcef&c facriftc~ ccdeiia~hctc. Hctno nctt pae d condencr ~eigneur~car ttpeutr fjnccncr Xca ~cna oonr kgh(e tye. *nc<no côfcnarvmcuta ecc~afttca. <~Mu<Yo:~nc!no nnt ne te n~c pcutr bien cccotnniumer ~ro.~oin !:nnu!te!!r ~cr ~~n~ cautc~c~ kcrrce aurcnnqs. 'nc'no porctt cjccôtcart ttnc cauta. ~j~t aux t~nUc~ apo!to!tquc9 Ttyctti!pa3parmor~]cp:M ~)chp~tic~ouatïc3p:C9 t~nc nemo pcutc cttre hctre 31c~ rotnp:c qui nctt pas pccirc. Ho!trc igK p:cienn6 conce~tont~ p~~tna!n nenum hccac tnfrmgcre. <j ~ee ro~s anirt cr <ïMrrcc pnncca jLr ~'3 p:ctidcno &c~ p:omnce~ ~n~nc i<3 rccomenc tce <cr~cn~ e~ contdUcr~ et ~urreo ~cn~ T! tc~ fonr pas mrer~h fonr ~ic a nemo t!3 reuckronc ~.es fccrcr~~c ïcnquïc tC~tove	Que Nemo fut tant charitable Qua vng poure qui auoit fain Donna des miettes de pain Cheans de la table du riche Qui tant estoit auar et chiche Quil ne luy en vouloit donner Mais cil nemo abandonner Lui en voulut piteussement Erat quidẽ pauper cupiens recreari de micis que cadebant de mensa diuitis:et nemo illi dabat. Luc.16. ¶ Nemo aussi pareillement Pour ses merites et deuoir Selon les droictz peuit bien auoir Sans bigamie ce me semble Deux femmmes espousees ensemble Nemini liceat duas vxores habere ¶ Urayemẽt cheres gẽs q biẽ pẽse A nemo/et son excellence Je cuide quil sen esmeruelle Car nul ne faict chose sans pareille Quil peult vendre et aliener Les corps sainctz/et en marchãnder Licitement et a sa guise Et des tresors de saincte eglise. Nemo martyres distrahat. Nemo mercet de sacrifiis ecclesiasticis. ¶ Nemo n est pas a condẽner Seigneurs/car il peult contẽner Les lyens dont l eglise lye. Nemo cõtẽnat vincula eccliastica. ¶ Plusfort/Nemo nul ne le nye Si peult bien excommunier Et d excommuniment lyer Sans cause/es lectres autentiqs. Nemo potest excõicari sine causa. ¶ Et aux bulles apostoliques Ny est il pas motz expres Escript/si est ou assez pres Que nom peult estre licite Les rompre qui n est pas petite. Nostre igit presentis concessionis paginam nemini liceat infringere. ¶ Les roys aussi et autres princes Et les presidens des prouinces Quant ilz recoiuent les sergens Des conseillers et autres gens Ne les font il pas iure / si font Que a nemo ilz reueleront	Les secrets/et senquis iestoye Pour cela/ie leur respondroye Que c est pour ce qu il ne mourra Poït/mais tousiors vif demourra Nemo est q sẽp viuat. Eccliastici.1. ¶ Que voulez vous q ie vous dye C est vne droicte melodie Que de nemo et de ses faictz Dont nen scauroye dire huy mais Et si en diray encore vng Qui entre les clercz est commun Et qui est si grant que sans faulte Je m esmerueille/aussi est haulte Le chose dont dieu escondit Ses disciples/quant il leur dit Que a eulx scauoir n appartenoit Quant la fin du monde seroit Qu on dit du iugement le iour Mais il leur dit lors sans seiour Que ne nemo sy le ssauoit bien. De die autem illo vel hora nemo scit. Marci.40. ¶ Somme toute ie croy et tien Que se dung moys ie ne cessoye D en parler/pas ie ne diroye De nemo tout ce qu on en trenne En chacun liure qui l aprenne Mesme saint Pol si nous recõpte Nemo par vain langaige et cõpte Peult seduire les gens en brief Et les mectre en vye de meschief Nemo vos seducat inanib9 verbis. ¶ Nemo peult bien sans contredit Resister en faict et en dit A ses ennemys sans doubter Sans qu ilz le puissent rebouter. Nemo inimicis audebit resistere. Levitici.26. ¶ Itẽ saint Jehã dit que nul hõme Ne peult aussi bien besongner De nuyt/quõ doit prẽdre son sõme Que nom/sil y veult soigner. Uenit nor c~u nẽo opari põt. Jo.10 ¶ Messeigneurs pourrt ie 9clus Par ce que iay dit cy dessus Priant le filz de la pucelle Qu iul nous doint la vie eternelle Quant son rigoreux examen Sera tenu/dictes Amen. ¶ Laus oco.

Dans les statistiques, il y a toujours deux distances de Levenshtein indiquées :

- la première mesure la distance « brute » entre l'OCR de la BnF et le texte considéré de la page
- la deuxième mesure la distance « fine » entre l'OCR de la BnF et le texte considéré de la page, ce dernier étant modifié pour prendre en compte les principales différences conceptuelles dûes à la syntaxe wiki :
  - les marqueurs de gras et d'italique sont retirés « '''mot en gras''' » et « ''mot en italique'' »
  - les notes de bas de page étant indiquées dans le texte lui-même « le texte<ref>et la note de bas de page</ref>, ainsi que la suite du texte », celle-ci sont déplacées à la fin du texte afin de comparer les deux textes de façon plus correcte

La deuxième distance est généralement plus faible que la première (et souvent beaucoup plus avec un rapport pouvant être de 1 à 5) et est au pire du même ordre de grandeur que la première.

Par exemple, pour le texte 2, trois versions sont déclinées :

- texte 2a : texte OCR d'origine ;
- texte 2b : texte après une première correction (statut rose) ;
- texte 2b-bis : texte 2b après retrait des sections d'en-tête et de pied-de-page et des marquages d'italique et de gras ;
- texte 2b-ter : texte 2b-bis après déplacement des références.

	T1a	T2b	T2b-bis	T2b-ter
T2a	0	1640	1473	296
T2b	1640	0	182	1450
T2b-bis	1473	182	0	1273
T2b-ter	296	1450	1273	0

Editcounts : les « *editcounts* » (compteurs d'édition) sont une métrique très couramment utilisée par les contributeurs aux projets Wikimedia pour évaluer l'âge d'un contributeur, métrique qui est parfois critiquée principalement sur Wikipédia arguant le fait qu'il est possible de faire beaucoup d'éditions de faible qualité ou peu d'éditions de très grande qualité.

Un ordre de grandeur de l'*editcount* peut être (sujet à appréciation personnelle) :

- 0 contributions : personne qui s'est seulement inscrite sans contribuer (environ 90% des comptes généralement) ;
- de 0 à 10 : contributeur ayant généralement testé et qui est reparti ;
- de 10 à 100 : nouveau contributeur ;
- de 100 à 500 : contributeur passant ou ayant passé un peu de temps sur le projet ;
- de 500 à 2000 : contributeur régulier (ou ayant été régulier), généralement ayant passé au moins 1 an à contribuer ;
- de 2000 à 10000 : contributeur habitué et faisant entièrement partie de la communauté, souvent présent depuis plusieurs années et/ou ayant passé beaucoup de temps sur le projet ;
- plus de 10000 : « pilier » de la communauté, présent depuis plusieurs années.

Il est à distinguer aussi les robots, comportant généralement le mot « bot » dans leur pseudo, qui sont des programmes automatiques effectuant des éditions répétitives et qui sont actionnés par un contributeur. Ces comptes ont généralement des editcounts de l'ordre du millier très rapidement et peuvent atteindre les dizaines de milliers, voire les centaines de milliers pour les plus anciens.

Collecte de données sur les pages, livres et contributeurs

À partir des dumps de l'espace de noms « Page » sont générés plusieurs fichiers : le premier contenant des informations sur la page en question, le deuxième étant une agrégation de toutes les pages d'un livre, et le troisième se focalisant sur l'activité des contributeurs. Ces fichiers permettent d'avoir une granularité fine pour retrouver des informations spécifiques à une page, un livre ou un contributeur lorsqu'elles seront noyées dans les moyennes diverses, et par exemple expliquer une valeur anormalement forte ou faible de certaines moyennes.

Informations sur les pages :

- Identifiant BnF et nom du livre ;
- Nombre de pages du livre et nombre de pages ;
- Numéro de la page ;
- Nombre de révisions de la page ;
- Date de création ;
- Statuts lors de la création et dernier statut de la page ;
- Nombre de contributeurs enregistrés et non enregistrés accompagné des listes ;
- Dates de passage au statut supérieur (gris, rose, jaune, vert), ainsi qu'un indicateur de non-linéarité du processus ;
- Indicateur de présence initiale de l'OCR BnF ;
- Distance de Levenshtein brute et fine entre l'OCR BnF et la dernière version comportant respectivement les statuts rose, jaune, vert (dans l'ordre brut-rose, brut-jaune, brut-vert, fin-rose, fin-jaune, fin-vert).

Exemple :

```
5085 Anonyme_-_Raoul_de_Cambrai.djvu 497 419 7 1284268262 3 4 3 1 Zyephyrus|Shaihulud|Joele.godard 130.208.138.241 0 1278371925 1282206858 0 False True 341 793 -1 313 366 -1
```

Informations sur les livres :

- Identifiant BnF et nom du livre ;
- Nombre de pages du livre ;
- Nombre de révisions ;
- Nombre de pages créées et de pages non créées ;
- Nombre de pages ayant le statut respectivement bleu, gris, rose, jaune et vert ;
- Nombre de contributeurs enregistrés et non enregistrés accompagné des listes ;
- Indicateur de présence initiale de l'OCR de la BnF ;
- Distances de Levenshtein moyennes et écarts-types sur les pages créées du livres entre l'OCR BnF et la dernière version comportant respectivement les statuts rose, jaune, vert : dans l'ordre 1) moyennes brutes rose, jaune, vert, 2) écarts-types bruts rose, jaune, vert, 3) moyennes fines rose, jaune, vert, 4) écarts-types fins rose, jaunes, vert ;
- Nombre de contributions non effectuées par des robots et pourcentage de telles contributions pour le livre.

Exemple :

```
74783 Gautier_-_les_noces_de_Cana_de_Paul_Véronèse.djvu 24 6 5 19 0 0 2 3 0 1 1 VIGNERON 82.216.237.91 True 243 432 -1 113 200 -1 163 208 -1 324 593 -1 121 309 -1 113 204 -1 41 82 -1 202 480 -1 6 100.0
```

Informations sur les contributeurs :

- Nom du contributeur
- Nombre total de contributions (« *editcount* »)
- Nombre de contributions sur le corpus
- Dates des contributions sur le corpus
- Date de la première contribution sur le corpus
- Date de la dernière contribution sur le corpus
- Premier, deuxième et troisième quartile des dates de contributions
- Indicateur vrai si le contributeur est un robot (« bot »)

Exemple :

```
Seb35 576 85 1278846960|1278847446|1278862117|1278862193|1279003238|
1279441325|1279441820|1279442059|1279442139|1279442481|1279442925|
1279446722|1279446743|1279447098|1279447336|1279450625|1279450647|
1279450675|1279450693|1279450708|1279450723|1279458136|1279458196|
1279458672|1279477757|1279478101|1279478912|1279479359|1279479724|
1279526220|1279527195|1279527524|1279527969|1279528461|1279633852|
1279635175|1279643863|1279644148|1279644224|1279644256|1279644283|
1279644306|1279644323|1279644338|1279644359|1279644374|1279644387|
1279644401|1279644417|1279644428|1279644439|1279644453|1279644465|
1279644515|1279644533|1279644544|1279644558|1279644569|1279644583|
1279644594|1279644607|1279644646|1279644657|1279644680|1279644694|
1279644706|1279644720|1279644733|1279644749|1279644760|1279644772|
1279644785|1279644798|1279644813|1279644825|1279644838|1279644849|
1279644863|1279644876|1279644892|1279644906|1279644927|1279644970|
1280146421|1280146458 1278846960 1280146458 1279458136 1279644323 1279644680
False
```

Statistiques

Sur les livres

Au départ :

- 1416 livres ont été donnés dans le cadre du partenariat
- 1057 avaient un OCR (et 359 sans OCR)
- 573 310 pages cumulées
- 404 pages par livre en moyenne

Nombre de...	*Quantité étudiée*	*03/10/ 2010*	*01/01/ 2011*	*01/04/ 2011*	*01/07/ 2011*	*01/10/ 2011*	*01/01/ 2012*	*01/04/ 2012*	*01/07/ 2012*	*01/10/ 2012*
livres...	...traités	169								
	...avec OCR initial	133								
pages...	...au total	60473								
	...créées	12393								
	...non créées	48080								
	...grises	260								
	...bleues	65								
	...roses	7751								
	...jaunes	3266								
	...vertes	1019								
	Moyenne par livre	358								
contributions...	...au total	19456								
	...au total ne venant pas de bots	10290								
	Moyenne par livre	115								
	Pourcentage ne venant pas de bots	52,889								
contributeurs par livre...	Moyenne des contributeurs enregistrés	1,811								
	Moyennne de contributeurs non-enregistrés	0,249								

Sur le travail demandé

				03/10/ 2010	*01/01/ 2011*	*01/04/ 2011*	*01/07/ 2011*	*01/10/ 2011*	*01/01/ 2012*	*01/04/ 2012*	*01/07/ 1012*	*01/10/ 2012*
Distance de Levenshtein OCR-texte	fine	moyenne	rose	323								
			jaune	370								
			vert	376								
		écart-type	rose	112								
			jaune	172								
			vert	155								
		minimum	rose	83								
			jaune	125								
			vert	139								
		maximum	rose	1962								
			jaune	2161								
			vert	1448								
	brute	moyenne	rose	239								
			jaune	239								
			vert	229								
		écart-type	rose	111								
			jaune	131								
			vert	115								
		minimum	rose	14								
			jaune	7								
			vert	88								
		maximum	rose	1855								
			jaune	2185								
			vert	536								

Sur les contributeurs

		03/10/ 2010	*01/01/ 2011*	*01/04/ 2011*	*01/07/ 2011*	*01/10/ 2011*	*01/01/ 2012*	*01/04/ 2012*	*01/07/ 2012*	*01/10/ 2012*
Contributeurs	Total	114								
	Plus de 5 contributions sur le corpus	68								
	Editcount >= 100	53								
Editcounts	Moyenne	7333								
	Écart-type	14561								
	Médiane	807								
Contributions par livre	Moyenne	85								
	Écart-type	206								
	Médiane	11								
Activité temporelle des contributeurs sur ces livres	Date minimum (moyenne)	13/07/ 2010								
	Date maximum (moyenne)	11/08/ 2010								
	Premier quartile (moyenne)	22/07/ 2010								
	Deuxième quartile (moyenne)	26/07/ 2010								
	Troisième quartile (moyenne)	01/08/ 2010								
	Premier quartile (écart-type)	49								
	Deuxième quartile (écart-type)	47								
	Troisième quartile (moyenne)	44								

		03/10/ 2010	*01/01/ 2011*	*01/04/ 2011*	*01/07/ 2011*	*01/10/ 2011*	*01/01/ 2012*	*01/04/ 2012*	*01/07/ 2012*	*01/10/ 2012*
Activité temporelle des contributeurs sur ces livres (5+ modifications sur le corpus)	Date minimum (moyenne)	05/07/ 2010								
	Date maximum (moyenne)	19/08/ 2010								
	Premier quartile (moyenne)	20/07/ 2010								
	Deuxième quartile (moyenne)	24/07/ 2010								
	Troisième quartile (moyenne)	02/08/ 2010								
	Premier quartile (écart-type)	51								
	Deuxième quartile (écart-type)	51								
	Troisième quartile (écart-type)	45								

Commentaires

1re livraison – 3 octobre 2010

Livres présents auparavant

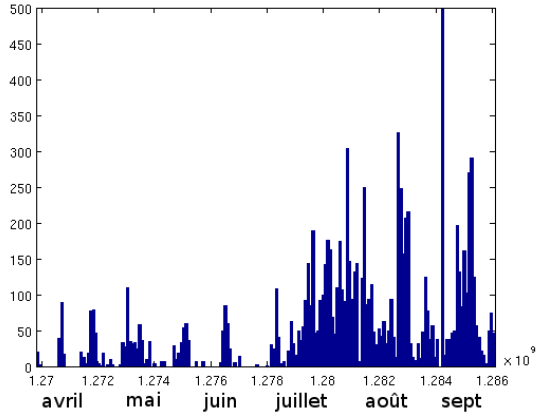
Dans les premières versions de ces statistiques, les résultats indiquaient une date moyenne de contribution vers le 18 mai 2010, ce qui était étrange étant donné que le téléversement des fichiers, mis à part trois fichiers de test en avril, s'est déroulé le 8 juillet 2010.

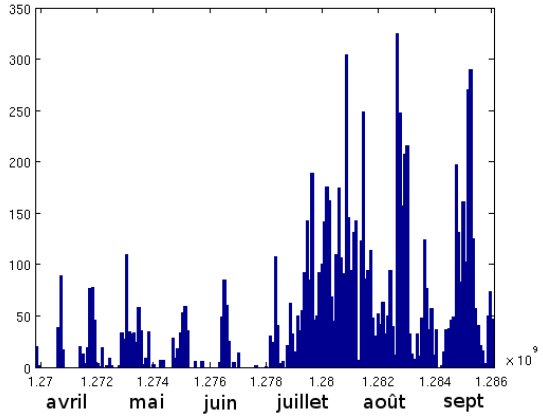
Il s'est trouvé que trois livres étaient déjà sur Wikisource avec le même titre avant ce partenariat. Il s'agit de « About - La Grèce contemporaine.djvu », « Zola - La Débâcle.djvu » et « Zola - Nana.djvu ». Ces trois livres ont été exclus des statistiques.

Un autre livre, « Schopenhauer - Le Monde comme volonté et comme représentation, Burdeau, tome 2, 1913.djvu », existait auparavant sous forme d'images indépendantes. Du fait de l'arrivée d'une version DjVu, les anciennes pages corrigées ont été déplacées vers ce DjVu, ce qui fait que le début de la correction est antérieure au partenariat. Ce livre a été ensuite exclus des statistiques.

Activité temporelle

La répartition des contributions au cours du temps est la suivante (premier graphique). La résolution est de une barre par jour, du 28 mars 2010 au 3 octobre 2010 (188 jours).





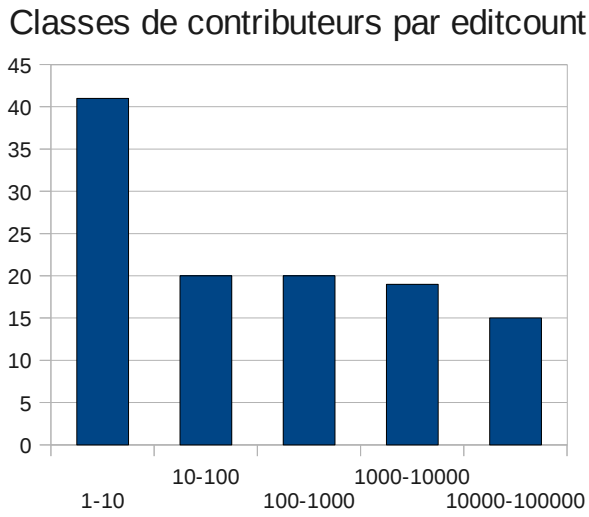
L'étendue de 500 contributions par jour est imposée par le pic du 12 septembre 2010 à 07:10 CEST, lors du déplacement d'un livre de 400 pages (« Anonyme - Raoul de Cambrai.djvu »).

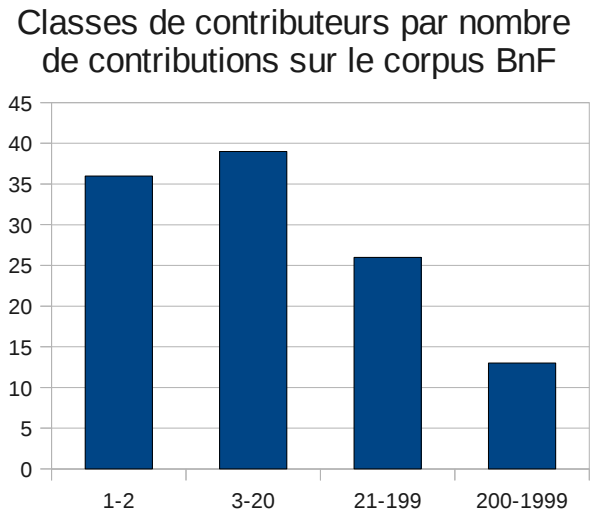
En retirant ce pic (deuxième graphique), nous obtenons une étendue de 330 contributions maximum par jour. En recherchant le deuxième pic (25 août 2010), on trouve que celui-ci correspond l'activité cumulée de deux contributeurs : deux initialisations avec corrections et passage en statut jaune de plusieurs pages de deux livres par deux contributeurs expérimentés (« Allais - À se tordre : histoires chatnoiresques.djvu » et « Boutroux - Pascal.djvu »).

On peut supposer que les autres pics correspondent à des opérations de maintenance (telles que l'initialisation d'un livre en statut rose pratiquée par certains contributeurs avant de commencer réellement la correction) ou à des corrections par des contributeurs expérimentés rapides ou à l'activité cumulée de plusieurs de ces opérations.

Activité générale des contributeurs

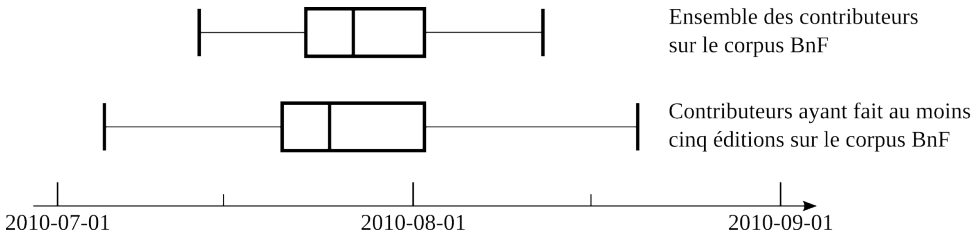
Sur les 116 contributeurs ayant participé à au moins une page du corpus, 71 ont fait plus de 5 contributions sur le corpus et 54 ont un *editcount* supérieur à 100. Cela signifie qu'une grosse partie des contributeurs qui ont participé au corpus sont des habitués de Wikisource, quelques nouveaux (une vingtaine) ont fait quelques éditions sur le corpus et une quarantaine (dont une majorité de contributeurs non-enregistrés) ont seulement fqit quelques passages (moins de 5 éditions).



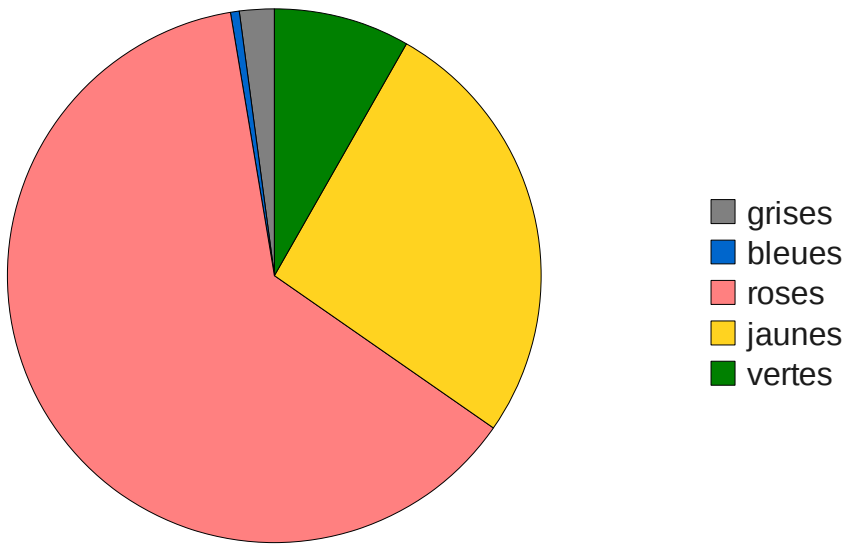


Activité temporelle des contributeurs

Les deux graphiques suivants indiquent les dates, toujours moyennées sur les contributeurs, de première et dernière contributions sur le corpus, ainsi que les premier, deuxième et troisième écarts-types, c'est-à-dire les dates du quart, de la moitié et des trois-quart de contributions des contributeurs.



Répartition du corpus pages



Notes de version

--- 0.1 – 10 octobre 2010 ---
- version initiale, il manque les distances de Levenshtein, certains paragraphes d'introduction ne sont pas finalisés
- première proposition de relecture par d'autres wikimédiens de Wikimédia France

--- 0.2 – 14 octobre 2010 ---
- ajout des exemples de distances de Levenshtein
- correction des résultats pour les distances de Levenshtein
- vérification des dates pour le tableau contributeurs (voir commentaires)

--- 0.3 – 17 octobre 2010 ---
- ajout de commentaires

--- 1.0RC1 – 17 octobre 2010 ---
- dernière proposition de relecture par d'autres wikimédiens et wikisourciers (pas de retours depuis la première proposition)

--- 1.0RC2 – 19 octobre 2010 ---
- prise en compte des remarques de Thierry Coudray (Wikimédia France), Zyephyrus (Wikisource), Pmx (Wikisource)

--- 1.0 – 24 octobre 2010 ---
- prise en compte du diagramme sur la répartition des pages de Jean-Frédéric Berthelot
- ajout des graphiques « Classes de contributeurs » et « activité temporelle des contributeurs »