La correction participative de l'OCR par crowdsourcing au profit des bibliothèques numériques

Mathieu Andro (1,2), Imad Saleh (2)

(1) DV IST, Institut National de la Recherche Agronomique. mathieu.andro@versailles.inra.fr
(2) Paragraphe, Université Paris 8

Sommaire

Résumé français

Dans le cadre de leurs projets de numérisation, les bibliothèques génèrent une OCR brute comportant souvent des erreurs qui peuvent ensuite être corrigées par des prestataires faisant appel à de la main d'œuvre à bas coût. Mais les bibliothèques peuvent aussi faire appel à des internautes bénévoles (crowdsourcing explicite), à des internautes rémunérés (Amazon Mechanical Turk Marketplace), à des internautes qui corrigent l'OCR sous la forme de jeux (gamification) ou encore à des internautes qui corrigent de l'OCR sans le savoir (crowdsourcing implicite de reCAPTCHA). Dans cet article, nous étudions ces approches et comparons la rentabilité de ces expérimentations concrètes.

Résumé anglais

For their digitization projects, libraries produce often OCR with errors which can be corrected by providers employing low cost labor. But libraries may also appeal to web volunteers (explicit crowdsourcing) or to a paid crowd (like Amazon Mechanical Turk marketplace) or to users correcting OCR by playing games (gamification) or to internet users who don't know that they are correcting OCR (implicit crowdsourcing like reCAPTCHA). Profitability of these experiments is compared.

Introduction

Les documents numérisés par les bibliothèques font très souvent l'objet d'une océrisation, c'est à dire d'un traitement informatique de reconnaissance optique de caractères (OCR) qui va chercher à identifier à quel caractère correspond la photographie de tel caractère. La finalité de cette opération est généralement de permettre la production de fichiers pour liseuses, l'indexation par les moteurs, la recherche en texte intégral, la réutilisation, l'exploitation scientifique ou encore la fouille de textes (text mining). Malheureusement, ce type de traitement génère de nombreuses erreurs. Ainsi, une disparité, une déformation, une décoloration, une tâche, un trou dans le papier, des annotations manuscrites, des typographies anciennes, originales, irrégulières ou mal imprimées ou encore une numérisation de mauvaise qualité vont cacher ou déformer l'aspect d'un caractère et tromper le logiciel qui identifiera un autre caractère que celui réellement présent. Les multiples erreurs générées par le logiciel OCR pourront bien être partiellement corrigées avec l'aide d'une confrontation des textes avec des dictionnaires de mots, mais un contrôle humain demeurera nécessaire car, à l'issue du processus automatisé, jusqu'à 20 % d'erreurs demeureront et seule une correction non automatique sera susceptible de réduire ce pourcentage, dans la mesure où les solutions logicielles ne sont pas encore capables de rivaliser avec les capacités humaines. En ce qui concerne les écritures manuscrites en particulier, comme le rappelle (Brokfeld 2012), l'OCR n'existe encore qu'à l'état expérimental (« Intelligent Word Recognition ») et il est fort probable qu'il le demeure encore quelque temps.

Pour toutes ces raisons, les bibliothèques externalisent aujourd'hui ce travail de correction manuelle de l'OCR auprès de prestataires qui font appel à de la main d'œuvre à bas coût, à Madagascar, en Inde ou encore au Viêt Nam. Une alternative à ces coûteuses et parfois critiquables prestations est de faire appel au crowdsourcing, c'est à dire d'externaliser ces opérations auprès de la foule des internautes en les engageant à participer à corriger les textes numérisés volontairement (crowdsourcing explicite), contre rémunération, sous la forme de jeux (gamification) ou encore sans qu'ils en aient conscience (crowdsourcing implicite). (Andro et Saleh 2014).

A partir d'un panorama des principaux projets dans le cadre de bibliothèques numériques publiques ou privées, notre étude propose une taxonomie originale des grands types de projets et cherche à en évaluer le rendement en termes financiers.

1- Taxonomie des formes de crowdsourcing utilisées

1.1- Le crowdsourcing explicite

Les projets de correction participative de l'OCR comme de transcription participative de manuscrits, peuvent être qualifiés de crowdsourcing explicite quand les internautes bénévoles qui y participent le font volontairement.

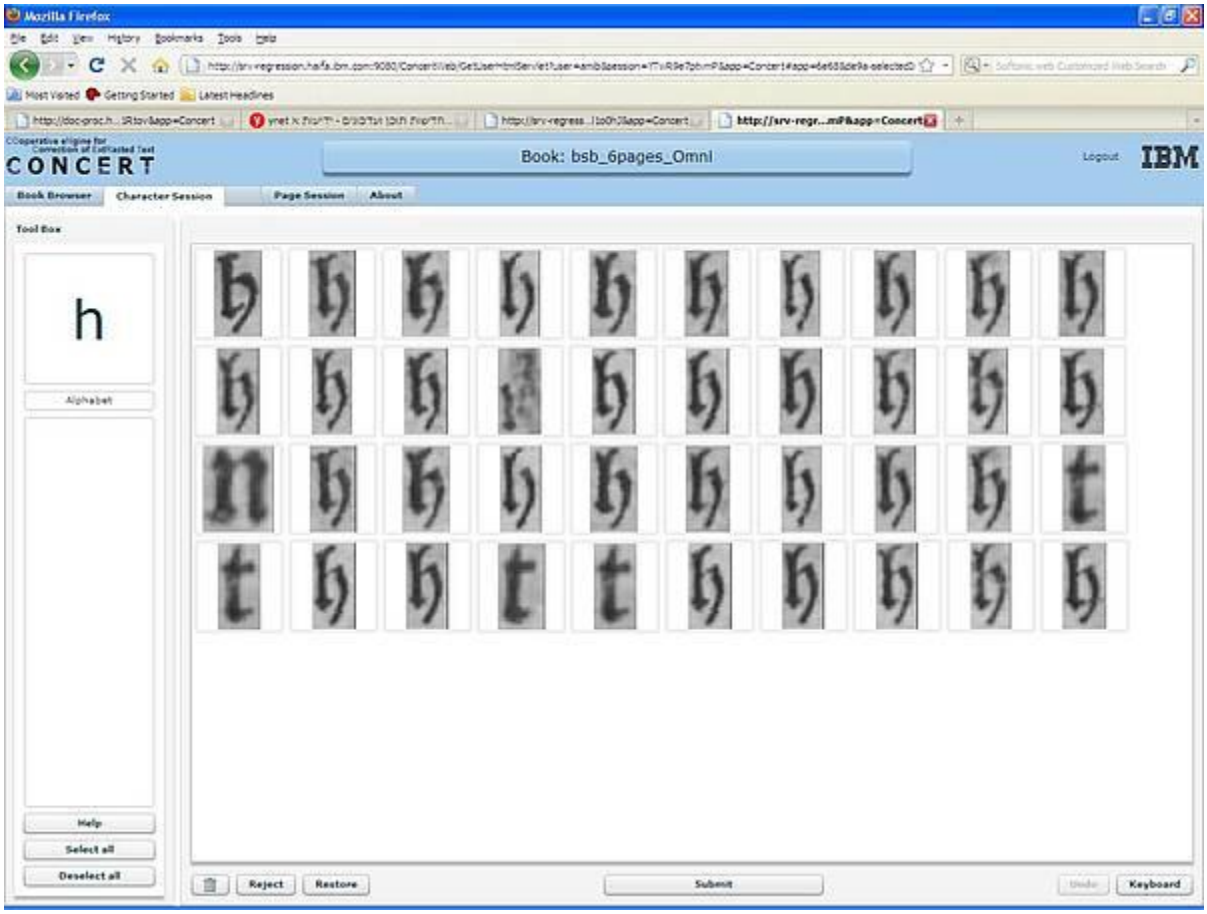
1.1.1- La correction participative de l'OCR

Concernant la correction participative et volontaire de l'OCR, en particulier, nous pouvons évoquer succinctement des projets comme Distributed Proofreaders, l'un des plus anciens projets de crowdsourcing dans le domaine de la numérisation et qui consiste à ce que les bénévoles produisent des ebooks pour le projet Gutenberg et pour Internet Archive, Wikisource utilisé en France dès 2008 par la Bibliothèque de l'Ecole Nationale Vétérinaire de Toulouse puis par la Bibliothèque nationale de France en avril 2010, l'Australian Newspapers Digitisation Program (TROVE), l'un des plus importants projets de correction participative de l'OCR avec près de 130 millions de lignes corrigées en mai 2014, le projet California Digital Newspaper Collection (CDNC) et le projet FUI12 Ozalid porté par la Bibliothèque nationale de France.

1.1.2- La transcription participative de manuscrits

S'agissant de la transcription participative et volontaire de manuscrits, nous pouvons citer le projet Transcribe Bentham qui consiste à transcrire les manuscrits du philosophe utilitariste afin de pouvoir les publier, le projet What's on the menu ? (WOTM) qui propose une transcription participative de 45 000 menus de restaurant depuis 1840, le projet Ancient Lives qui consiste en une transcription de papyrus égyptiens, le projet ArcHIVE de transcription de catalogues d'archives, le projet What's the score (WTS) pour des partitions de musique, le projet Monasterium Collaborative Archive (MOM-CA), pour les manuscrits médiévaux ou encore le projet Citizen Archivist Dashboard pour les archives nationales des USA.

La correction classique dans le contexte répond bien aux besoins d'internautes qui souhaitent profiter de leur travail bénévole pour mieux prendre connaissance de textes qui les intéressent. Par contre, elle serait moins performante que la correction hors contexte qui permettrait d'obtenir des résultats optimisés comme proposé, par exemple, par le projet COoperative eNgine for Correction of ExtRacted Text (CONCERT) développé par IBM Israël :



Capture d'écran de CONCERT (étape du "tapis") d'après https://www.digitisation.eu
Certaines lettres ne correspondent pas à la lettre h et peuvent être facilement identifiées

Le "tapis" ci-dessus affiche toutes les occurrences d'un même caractère trouvé dans le livre et identifiés comme suspects par le logiciel OCR. Au lieu de corriger chacun de ces caractères indépendamment dans leur contexte, le système permet à l'internaute d'identifier très rapidement les caractères qui ne correspondent pas. Dans l'exemple ci-dessus, un caractère est illisible, on lit 4 caractères "t" et 1 caractère "n", tous les autres sont bien des caractères "h". A partir des résultats obtenus au cours de cette étape, le système va apprendre, grâce à l'usager, à mieux effectuer son OCR.

1.2- Le crowdsourcing rémunéré

Au lieu de faire appel au travail gratuit et bénévole des internautes, il est également possible de rémunérer leur travail. Des plateformes de crowdsourcing rémunéré comme l'Amazon mechanical turk marketplace (https://www.mturk.com), leader sur ce marché, mais aussi Guru (http://www.guru.com), crowdflower (http://crowdflower.com) et, pour la France, FouleFactory (http://www.foulefactory.com), permettent ainsi à des institutions et à des sociétés de proposer des microtâches rémunérées à accomplir à des internautes. Généralement, il s'agit de tâches difficiles à automatiser avec des programmes informatiques comme l'indexation d'images, la classification, la transcription audio, la rédaction de résumés, l'identification d'images obscènes,

l'ajout de « likes », de relations, d'avis ou de commentaires sur les réseaux sociaux, et, parfois même, la correction de l'OCR.
Concernant l'utilisation de l'Amazon Mechanical Turk Marketplace au bénéfice de projets de numérisation du patrimoine des bibliothèques, une expérimentation de transcription de manuscrits aurait permis d'obtenir des coût de 60 $ pour la transcription de 200 pages de manuscrits quand cette prestation aurait coûté 400 $ avec un prestataire traditionnel (Lang et Rio-Ross 2011).

1.3- La gamification

Il est également possible de faire jouer les internautes au bénéfice des projets de numérisation en leur faisant corriger l'OCR brute grâce à des jeux avec une finalité ("games with a purpose"). Ainsi, la Bibliothèque Nationale de Finlande, à travers le projet Digitalkoot, propose un jeu, "Mole Hunt" (ou chasse aux taupes) qui consiste à ce que des taupes sortent de leurs trous pour exposer à l'internaute une image du mot et une proposition de transcription. L'internaute doit déterminer le plus rapidement possible si ce sont bien les mêmes mots en validant ou en la refusant. Ce faisant, il corrige les résultats de l'OCR brut.

Figure 1. Capture d'écran du jeu Mole Hunt

Le deuxième jeu ("Mole Bridge" ou pont des taupes) consiste à transcrire les mots images pour construire un pont et faire traverser une armée de taupes. A chaque bonne réponse, une brique du pont à construire s'ajoute aux précédentes. En cas d'erreur, une brique du pont explosera et les taupes risqueront de tomber dans l'eau. A l'instar de n'importe quel jeu d'arcades, les décors, les vitesses, les distances… changent en fonction des changements de niveaux, ce qui stimule les joueurs à poursuivre leur progression et à continuer de jouer.

Figure 2. Capture d'écran du jeu Mole Bridge

Nous pourrions également évoquer TypeAttack, un jeu Facebook permettant de faire corriger l'OCR de textes numérisés par des internautes, le jeu Word Soup Game ou encore le projet COoperative eNgine for Correction of ExtRacted Text (CONCERT) précédemment mentionné, car il fait également largement appel aux ressorts de la gamification.

1.4- Le crowdsourcing implicite

Enfin, il est également possible de bénéficier des contributions involontaires et inconscientes des internautes sous la forme de crowdsourcing implicite.

Ainsi, reCAPTCHA dont le slogan est "Stop spam, read books" et dont la vocation première est de vérifier qu'il s'agit bien d'un humain et non d'un robot malveillant qui souhaite ouvrir un compte sur un site web a été astucieusement utilisé par le New York Times puis par Google Books et Google Street View afin de faire corriger les textes océrisés ou les numéros de rues par les internautes (Von Ahn, Maurer, Mcmillen, Abraham, Blum 2008). Ces derniers ignorent bien souvent qu'ils participent à corriger les textes de Google Books et qu'ils effectuent leurs saisies exclusivement pour des raisons de sécurité et non pour participer à un projet culturel, C'est la raison pour laquelle on parle de crowdsourcing implicite. Les mots océrisés dans le cadre de ces projets de numérisation et non reconnus par des dictionnaires sont alors soumis aux internautes pour correction. Ainsi, reCAPTCHA affiche aux internautes systématiquement 2 mots sous forme d'images distordues : un mot en OCR à corriger par l'internaute et un mot en OCR déjà corrigé afin de vérifier que l'internaute n'est pas un robot.

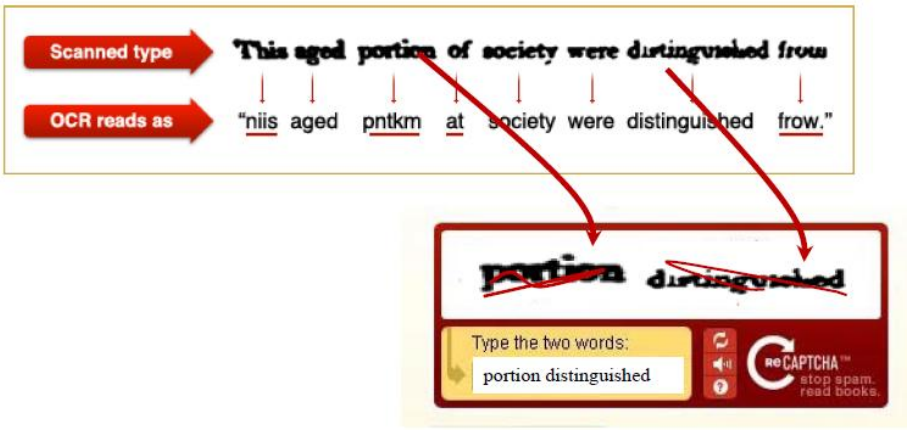


Figure 3. Schémas explicatifs du fonctionnement de reCAPTCHA, d'après (Ipeirotis 2011)

Le même mot à corriger est soumis à 3 internautes différents dans le monde et avec des distorsions différentes à chaque fois afin d'éviter qu'elles génèrent les mêmes erreurs. Leurs saisies sont ensuite confrontées. Si elles sont identiques, le mot est alors considéré comme corrigé. Sinon, le mot est soumis à d'autres internautes, chaque humain comptant pour 1 vote et chaque identification d'OCR pour 0,5 vote, jusqu'à ce qu'une hypothèse obtienne le total de 2,5 voix. Il est possible, pour les internautes de demander à reCAPTCHA d'afficher un autre mot au cas où ils ne parviendraient pas à le lire. Si le même mot est rejeté 6 fois, il est considéré comme illisible.

2- Calculs, bénéfices et coûts

D'après les statistiques de reCAPTCHA, 68 % des mots sont corrigés avec l'aide de seulement 2 internautes, 18 % en nécessitent 3, 7 % en nécessitent 4, 3 % en nécessitent 5, 4 % en nécessitent 6 ou plus. Le résultat obtenu est un OCR corrigé à 99,1 %. En 2012, chaque jour, près de **200 millions de mots** reCAPTCHA auraient ainsi été saisis par les internautes qui y auraient consacré 12 000 heures de travail par jour et ce rythme, annoncé comme croissant, pourrait être bien supérieur aujourd'hui.

Si nous souhaitons calculer les quantités que reCAPTCHA évite à Google Books de dépenser en correction de l'OCR, nous pouvons proposer les calculs suivants :

- Si 68 % des mots nécessitent 2 internautes, chaque jour 200 millions de mots x 0,68 = 136 millions de mots saisis permettent d'obtenir 136 millions de mots / 2 internautes = 68 millions de mots validés.
- Si 18 % des mots nécessitent 3 internautes, chaque jour 200 x 0,18 = 36 millions de mots saisis permettent d'obtenir 36 / 3 = 12 millions de mots validés.
- Si 7 % des mots nécessitent 4 internautes, chaque jour 200 x 0,07 = 14 millions de mots saisis permettent d'obtenir 14 / 4 = 3,5 millions de mots validés.
- Si 3 % des mots nécessitent 5 internautes, chaque jour 200 x 0,03 = 6 millions de mots saisis permettent d'obtenir 6 / 5 = 1,2 millions de mots validés.
- Si 4 % des mots nécessitent au moins 6 internautes, chaque jour 200 x 0,04 = 8 millions de mots saisis permettent d'obtenir 8 / 6 = 1,33 millions de mots validés que nous

pouvons arrondir à 1,3 millions dans la mesure où ils nécessitent 6 internautes ou parfois plus.

Au total, nous aurions donc 68+12+3,5+1,2+1,3 = **86 millions de mots validés par jour**.

Si nous considérons qu'il y a 75 000 mots en moyenne dans un livre de 300 pages comportant 250 mots par page, chaque jour, l'équivalent de 86 millions / 75 000 = **1147 livres seraient ainsi corrigés chaque jour à un taux OCR de 99,1 %**.

Si Google Books semble désormais ralentir le rythme de son programme de numérisation, son programme n'est pas encore achevé et il dépasse déjà de beaucoup les objectifs de 15 millions de livres annoncés début 2005 et qui semblaient, à l'époque, irréalistes. Le nombre de livres numérisés par Google Books dépasserait aujourd'hui les 30 millions de livres. Afin de corriger l'OCR de ces 30 millions de livres, il faudrait donc 30 millions / 1147 livres corrigés par jour = un peu plus de 70 ans au rythme de 2008. Leonid Taycher, ingénieur chez Google, évaluait sur son blog, le 5 août 2010, à 129 864 880 le nombre total de livres imprimés depuis le début de l'imprimerie. La correction de l'OCR de tous ces ouvrages imprimés prendrait quant à elle 310 ans toujours au rythme de 2008.

Néanmoins, ce rythme est sans commune comparaison avec celui de correction de l'OCR pratiqué par les bibliothèques et qui est effectué à la main par des prestataires faisant appel à des pays en voie de développement comme Madagascar, le Viêt-Nam ou l'Inde. Si, à l'instar des bibliothèques, Google faisait appel à ce type de main d'oeuvre au lieu d'utiliser le crowdsourcing implicite des internautes via reCAPTCHA, et s'il finançait la correction de l'OCR pour un prix à la page compris entre 1 € et 1,5 € soit entre 300 € et 450 € pour un livre de 300 pages, il lui faudrait payer entre 300 € x 1147 livres = 344 100 € et 450 € x 1147 livres = 516 150 €. On peut donc raisonnablement considérer que reCAPTCHA évite à Google de dépenser plus de 400 000 € par jour, soit **146 millions d'euros par an**.

En complétant ces estimations avec celles de (Ipeirotis 2011), de (Geiger et Zarndt 2012) et de (Zarndt 2014), nous avons évalué les coûts que les projets suivants auraient eu à dépenser pour la correction humaine de l'OCR s'ils avaient fait appel à un prestataire classique :

Projet	**Type de crowdsourcing**	**Coût non dépensé**
California Digital Newspaper Collection (fin 2011-)	Crowdsourcing explicite	53 130 $ cumulés en juin 2014
TROVE (août 2008-)	Crowdsourcing explicite	2 580 926 $ cumulés en mai 2014
Digitalkoot (février 2011-)	Gamification	Entre 31 000 et 55 000 € cumulés en octobre 2012
Google Books et reCAPTCHA (2008-)	Crowdsourcing implicite	146 millions d'euros par an (au rythme de 2008)

Tableau 1. Estimation du coût non dépensé en prestations de correction de l'OCR par quelques projets de crowdsourcing

Ces coûts non dépensés doivent toutefois être relativisés et être rapportés aux coûts qui ont eux été dépensés pour développer les systèmes de crowdsourcing, les administrer, assurer la

communication des projets, le community management des volontaires et parfois, pour réintégrer les données produites par les internautes dans la bibliothèque numérique.

3- Communitysourcing plutôt que crowdsourcing ?

Force est de constater que si les projets de crowdsourcing explicite s'adressent en théorie à des foules importantes d'internautes indifférenciés et anonymes, en pratique, la majeure partie du travail est généralement l'œuvre d'une minorité de bénévoles motivés qui contribuent régulièrement aux projets, qu'il s'agit finalement d'une communauté de fidèles volontaires qui s'assistent mutuellement (Owens 2013) (Carletti, Giannachi, Price, McAuley 2013).
Une manière originale et intéressante d'illustrer ce phénomène est proposée par le projet de transcription d'observation météorologiques Old Weather porté par la Citizen Science Alliance. Dans ce graphique, la taille de chaque carré est proportionnelle au nombre de notices transcrites. Sur le 1,6 millions de notices transcrites par 16 000 volontaires, un dixième l'a été par seulement 20 contributeurs :

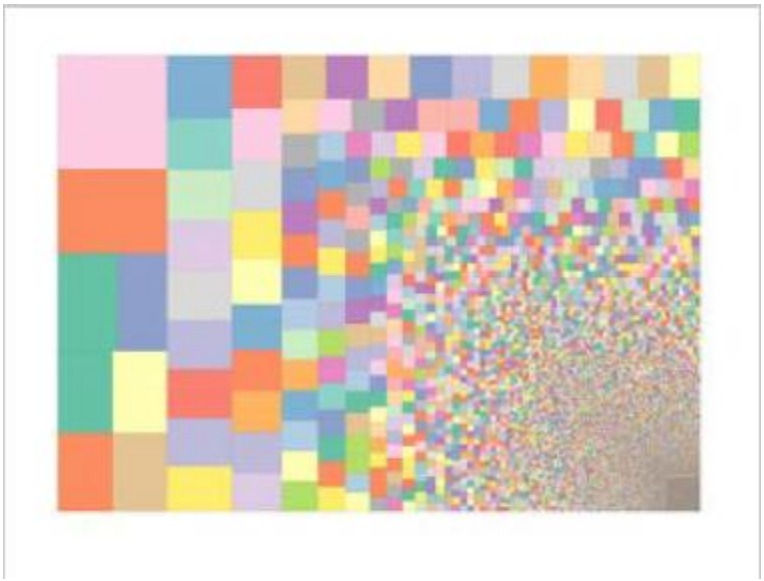


Figure 4. Illustration graphique du fait que quelques internautes produisent la majeure partie des contributions (Brumfield 2012)

Les participants aux projets Trove, California Digital Newspaper Collection ou Cambridge Public Library sont, pour le majeure partie d'entre eux, des généalogistes, des retraités aisés s'intéressant à l'histoire locale (Hagon 2013). Dans ces conditions, il est donc préférable de parler, pour tous ces projets, de communitysourcing ou encore de nichesourcing plutôt que de crowdsourcing (Causer et Wallace 2012).

Conclusion

Crowdsourcing explicite, rémunéré, gamification ou crowdsourcing implicite, la correction et la transcription participatives peuvent prendre des formes variées.

Les limites du temps de travail bénévole susceptible d'être mobilisé par le crowdsourcing explicite semblent se rapprocher et, avec la disparition prévisible des générations de retraités qui étaient passionnées de généalogie et d'histoire locale, et avec l'amélioration des capacités des logiciels de reconnaissance de caractères, la correction participative de l'OCR pourrait avoir atteint un seuil critique à partir duquel les possibilités offertes par le crowdsourcing explicite seraient en train de se resserrer. Ainsi, comme le constate (Ayres 2013), malgré une croissance continue du nombre de textes dont la correction est proposée par l'un des plus importants projets de correction participative de l'OCR, le projet TROVE, le nombre de corrections a finalement cessé de croître depuis 2011.

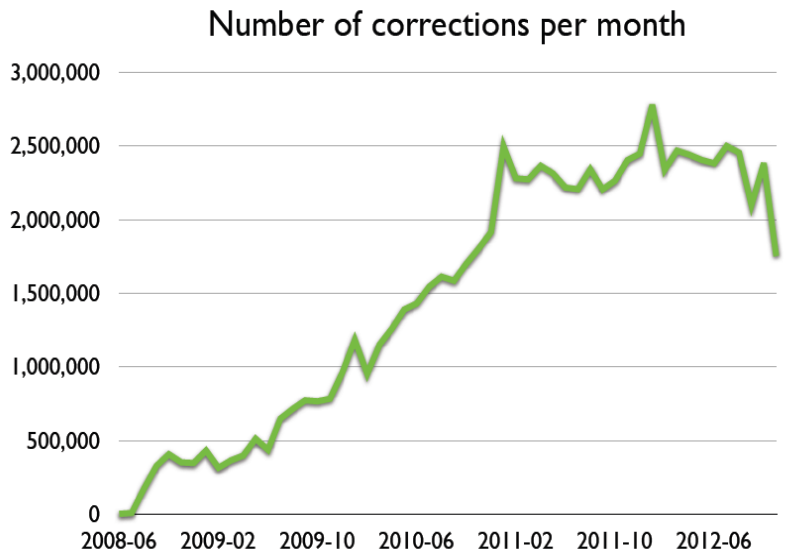


Figure 5. L'évolution du nombre de corrections sur TROVE, d'après (Hagon 2013)

Bien que moins enrichissant intellectuellement pour les internautes, le crowdsourcing implicite sur le modèle de celui mis en œuvre par reCAPTCHA semble avoir de beaux jours devant lui. Néanmoins, le 3 décembre 2015 Google a annoncé[1] son remplacement par "No CAPTCHA reCAPTCHA", un système de sécurité faisant désormais plutôt appel à la logique humaine.
Le crowdsourcing implicite tient compte du fait que seule une infime minorité d'internautes participe à Wikipedia ou à d'autres projets de crowdsourcing explicite et philanthropique et qu'il est, par conséquent, plus astucieux et beaucoup plus efficace, de recycler l'énergie des internautes dans leurs activités courantes sur le web. L'utilisation de ce système de Captcha par d'autres grandes bibliothèques numériques comme Internet Archive (Gary 2013) pourrait donc être prometteur et apporter des corrections complémentaires à celles obtenues via la voie plus classique du crowdsourcing explicite et volontaire. Mais il ne pourra s'adresser qu'aux programmes de numérisation de masse. Pour les bibliothèques numériques plus modestes à la recherche de solutions simples et pragmatiques, en dehors de l'étude précédemment mentionnée (Lang et Rio-Ross 2011), le crowdsourcing rémunéré avec l'aide de l'Amazon

[1] http://googleonlinesecurity.blogspot.fr/2014/12/are-you-robot-introducing-no-captcha.html

Mechanical Turk Marketplace a encore assez peu été expérimenté. Il devrait faire l'objet de nouvelles expérimentations très prochainement.

Bibliograhie

- Andro, Mathieu, Saleh, Imad. 2014. « Bibliothèques numériques et crowdsourcing : une synthèse de la littérature académique et professionnelle internationale sur le sujet ». Colloque International sur le Document Numérique, CIDE 17, 10 p.
- Ayres, Marie-Louise. 2013. « 'Singing for their supper': Trove, Australian newspapers, and the crowd ». IFLA World Library and Information Congress. Singapore. 9 p.
- Brokfeld, Jens. 2012. « Die digitale Edition der „preußischen Zeitungsberichte": Evaluation von Editionswerkzeugen zur nutzergenerierten Transkription handschriftlicher Quellen ». Master Informationswissenschaften, 148 p.
- Brumfield Ben. 2012. « Bilateral Digitization at Digital Frontiers 2012 ». http://manuscripttranscription.blogspot.fr/2012/09/bilateral-digitization-at-digital.html
- Carletti, Laura, Giannachi, Gabriella, Price, Dominic, McAuley, Derek. 2013. « Digital Humanities and Crowdsourcing: An Exploration ». MW2013: Museums and the Web 2013: The annual conference of Museums and the Web, April 17-20, 2013, Portland, USA. 18 p.
- Causer, Tim, Wallace Valerie. 2012. « Building A Volunteer Community: Results and Findings from Transcribe Bentham ». Digital Humanities Queterly, vol. 6, no 2, 26 p.
- Gary, Nicolas. 2013. « Exclusif : Un Captcha pour Internet Archive, concurrent de Google Books ». ActuaLitté. https://www.actualitte.com/usages/exclusif-un-captcha-pour-internet-archive-concurrent-de-google-books-45939.htm
- Geiger, Brian, Zarndt, Frederick. 2012. « No tempest in my teapot: analysis of crowdsourced data and user experience at the California Digital Newspaper Collection ». http://fr.slideshare.net/cowboyMontana/20121105-no-tempest-in-my-teapot-dlf-forum-denver
- Hagon, Paul. 2013. « Trove crowdsourcing behaviour ». http://www.information-online.com.au/pdf/Tuesday_Concurrent_2_1125_Hagon.pdf
- Ipeirotis, Panagiotis G. 2011. « Managing Crowdsourced Human Computation ». WWW2011 tutorial, 29 March 2011, 196 p.
- Lang, Andrew S. I. D., Rio-Ross, Joshua. 2011. « Using Amazon Mechanical Turk to Transcribe Historical Handwritten Documents ». Code4lib Journal, vol. 15, 10-31.
- Owens, Trevor. 2013. « Digital Cultural Heritage and the Crowd ». Curator: The Museum Journal, vol. 56, 121–130. doi:10.1111/cura.12012.
- Von Ahn, Luis, Maurer, Benjamin, Mcmillen, Colin, Abraham, Davis, Blum, Manuel. 2008. « reCAPTCHA: Human-Based Character Recognition via Web Security Measures ». Science no 321, 1465-1468. doi:10.1126/science.1160379.
- Zarndt, Frederick. 2014. « Crowdsourcing family history, and long tails for libraries ». http://fr.slideshare.net/cowboyMontana/20140628-crowdsourcing-family-history-and-long-tails-for-libraries-ala-annual-las-vegas