ETAT DE L'ART EN MATIERE DE *CROWDSOURCING* DANS LES BIBLIOTHEQUES NUMERIQUES

Pauline Moirez
Jean Philippe Moreux
Isabelle Josse

Février 2013

Ce document a été réalisé dans le cadre du projet de R&D du FUI[1] 12 pour la conception d'une plateforme collaborative de correction et d'enrichissement des documents numérisés.

Le projet est porté par un consortium de 9 partenaires :

- Orange Labs : Coordinateur, architecture
- BnF : Expérimentation, Fonds documentaire, animation de réseaux sociaux
- Université Paris 8 : Ergonomie, sociologie et accessibilité
- Jamespot : Plateforme de réseau social
- Urbilog : Webservices, composant d'interface (transcription)
- I2S (Innovative Imaging Solutions) : scanner, valorisation du prototype
- ISEP (Institut Supérieur d'Electronique de Paris) : évaluation de la qualité
- INSA Lyon (équipe IMANGINE) : Traitement d'image
- Université Claude Bernard Lyon 1 (équipe SILEX du LIRIS) : Apprentissage, Analyse d'activité

Enjeux du projet

Après deux décennies de numérisation du patrimoine, il n'existe toujours pas de solution infaillible permettant de passer d'un document numérisé à une version en mode texte. Les systèmes de reconnaissance optique de caractères (OCR) permettent de détecter et transposer un mot à partir d'une image, mais ils laissent encore trop d'imperfections pour parvenir à une réédition du document. Le seul moyen est d'utiliser l'intelligence humaine pour analyser le contexte, les langues, la sémantique.

L'objectif du projet est de concevoir une plateforme de correction collaborative de documents numérisés pour en faire des documents rééditables, accessibles à l'ensemble des utilisateurs et adaptés aux nouveaux usages. Le projet envisage une approche par *crowdsourcing* pour permettre la correction et l'enrichissement collaborative des documents et va s'appuyer sur les réseaux sociaux pour organiser cette collaboration.

[1] Le fonds unique interministériel finance les projets de recherche et développement collaboratifs des pôles de compétitivité (http://competitivite.gouv.fr/accueil-3.html).

Table des matières

I. Définitions et typologie des projets de *crowdsourcing*

I.1. Les enjeux

Les bibliothèques numériques peuvent être extrêmement diverses, du point de vue du nombre de documents mis à disposition, des types de documents, des fonctionnalités, de la médiation des contenus, etc. La constitution d'une bibliothèque numérique repose ainsi sur trois volets interdépendants2 :

- Un volet documentaire : quel(s) contenu(s), quel(s) corpus, quelle typologie, quel volume…
- Un volet technique : architecture, formats, fonctionnalités offertes (zoom, recherche, espace personnel, téléchargement, etc.)
- Relationnel : identité des usagers, usages de la bibliothèque, lien avec l'extérieur…

La réussite d'une bibliothèque numérique dépend de la bonne articulation entre ces trois volets. Ainsi, lorsque la BnF a fait le choix de la numérisation de masse, l'augmentation des volumes dans Gallica a entraîné des changements techniques mais a également eu des répercussions sur les relations avec les utilisateurs.
C'est pourquoi l'intégration de fonctionnalités participatives au sein d'une bibliothèque numérique doit être pensée en termes d'usages et de fonctionnalités mais également en termes documentaires (en particulier choix du ou des corpus).

Les bibliothèques s'inscrivent en effet dans un écosystème du web où l'interaction est la norme : l'internaute s'attend à pouvoir intervenir sur les données et sur les contenus, que ce soit pour les commenter, les partager ou les enrichir. Même lorsqu'il n'utilise pas ces fonctionnalités[3], elles lui sont familières dans sa pratique courante du web, sur les réseaux sociaux ou les sites marchands. Elles constituent son cadre de référence, il se sentira enfermé et exclu s'il ne les a pas à sa disposition[4]. Et pourtant, l'expérience montre que l'intégration de fonctionnalités d'enrichissement collaboratif dans les catalogues ou bibliothèques numériques françaises rencontre rarement jusqu'à maintenant le succès escompté, et peine à atteindre la masse critique nécessaire pour améliorer notablement l'expérience de recherche des usagers[5], alors que des bibliothèques anglo-saxonnes ou d'autres institutions culturelles, en particulier les services d'archives[6], parviennent à mettre en place des projets particulièrement réussis.

---

[2] Lionel Maurel *Bibliothèques numériques : quels enjeux, quels modèles ?*, 2011.
[3] La règle du « 1-9-90 » veut que seul 1% des internautes participe activement à l'enrichissement de contenus en ligne, 9 % y contribuent occasionnellement, et 90 % soient des consommateurs passifs (http://fr.wikipedia.org/wiki/R%C3%A8gle_du_1_%25). On assiste toutefois à une remise en cause progressive de cette règle, vers une participation accrue des internautes (jusqu'aux ¾ de contributeurs au moins occasionnels au Royaume-Uni, par exemple), voir Aref Jdey, « La règle des 90/9/1 est désormais dépassée », *Demain la veille*, 2012, http://www.demainlaveille.fr/2012/07/02/la-regle-des-9091-est-desormais-depassee/.
[4] Etienne Cavalié, « Les tags dans les OPAC : ce n'est pas parce que personne ne s'en sert que ça ne sert à rien », *Bibliothèques (reloaded)*, 2010, http://bibliotheques.wordpress.com/2010/02/19/les-tags-dans-les-opac-ce-nest-pas-parce-que-personne-ne-sen-sert-que-ca-ne-sert-a-rien/.
[5] Lionel Dujol, « Le catalogue 2.0 ou le mythe de l'usager participatif ? », *La bibliothèque apprivoisée*, 2009, http://labibapprivoisee.wordpress.com/2009/10/14/le-catalogue-2-0-ou-le-mythe-de-lusager-participatif/.
Bertrand Calenge, « Des publics utilisateurs aux publics collaborateurs : une fausse bonne idée ? », *Bertrand Calenge : carnet de notes*, 2012, http://bccn.wordpress.com/2012/02/11/des-utilisateurs-aux-collaborateurs-une-fausse-bonne-idee/.
[6] Pauline Moirez, « Archives participatives », dans *Bibliothèques 2.0 à l'heure des médias sociaux*, dir. Muriel Amar et Véronique Mesguich, 2012, p. 187-197.

La participation des usagers, qui peut exister sur de simples données bibliographiques, est renforcée par la mise en ligne de documents numériques. En effet, la mise à disposition des usagers de documents numérisés, images voire textes OCRisés, permet des opérations de *crowdsourcing* plus ambitieuses qui enrichissent notablement la description des documents : indexation, identification de photographies, correction d'OCR ou encore transcription collaborative.

Il s'agit de s'insérer dans l'écosystème participatif du web pour contribuer à enrichir les catalogues de bibliothèques et à améliorer la description de leurs collections dans une mesure que l'on ne pouvait pas imaginer jusque-là, à la fois en termes de volumes de données produites et de nature même de ces données. En effet, les données produites par les internautes, qui peuvent être désignées globalement sous le terme de « **métadonnées sociales** », permettent de répondre à des besoins différents et d'offrir aux usagers et aux chercheurs des services différents et complémentaires à ceux ouverts par les métadonnées produites par les catalogueurs professionnels :

- Des informations qui correspondent davantage aux besoins et usages des internautes : besoins de recherche en plein texte, granularité de description plus fine, bases de données nominatives, géolocalisation des documents,
- Un volume extraordinaire de contributions,
- Le regroupement de compétences particulières, qui permet de faire appel aux compétences scientifiques et à l'expertise des chercheurs.

Toutefois, la coexistence dans les catalogues de bibliothèques de données produites par des professionnels et de données produites par les internautes nécessite d'apporter une grande vigilance à la qualité des données produites.

Il faut également rester vigilants à éviter l'écueil d'un collaboratif « cosmétique », réalisé pour se conformer aux codes d'un web par nature interactif, mais qui n'améliorerait pas véritablement les fonctionnalités offertes aux usagers, et tromperait finalement l'internaute qui croit contribuer à cette amélioration. Il est ainsi souhaitable de prévoir la réintégration des contenus enrichis dans les catalogues, sur les sites web des bibliothèques, pour qu'ils améliorent véritablement l'expérience de recherche de l'usager, que ces enrichissements collaboratifs aient été produits sur le site de la bibliothèque ou déportés sur des médias externes.

« *Sharing and Aggregating Social Metadata* », une étude de l'OCLC sur les métadonnées sociales[7]

« We believe it is riskier to do *nothing* and become irrelevant to our user communities than to *start using* social media features"

L'étude menée en 2011-2012 par l'OCLC[8] sur les métadonnées sociales dans bibliothèques, archives et musées montre bien l'intérêt soutenu des acteurs du web culturel pour le *crowdsourcing*, et la prise de conscience de ses implications aussi bien pour l'enrichissement des collections numériques que pour l'établissement de nouvelles interactions avec les usagers.

L'étude a produit plusieurs livrables :

- Une analyse de 76 sites de bibliothèques, archives ou musées proposant des

[7] http://www.oclc.org/research/activities/aggregating.html

[8] L'OCLC (*Online Computer Library Center*) est un organisme de recherche mondial, à but non lucratif, qui propose des produits et services aux bibliothèques dans le but d'accroître l'accès à l'information. http://www.oclc.org/fr/fr/default.htm

fonctionnalités participatives (*tagging*, indexation collaborative, commentaires, recommandations, utilisation de plateformes sociales et de réseaux sociaux, etc.)
- Une enquête auprès de responsables (42 réponses) de site d'institutions culturelles investies dans des pratiques participatives
- Une série de recommandations pour les institutions culturelles intéressées par ce type de programme

I.2. Terminologie

La fluctuation de la terminologie désignant ces opérations participatives souligne leur diversité, mais aussi les différences d'objectifs et de stratégies des institutions qui les mettent en œuvre.

On parlera de « participation[9] » des usagers lorsque l'on veut désigner la mise en œuvre de véritables compétences et connaissances des usagers, une interaction de haut niveau, de caractère scientifique, qui contribue à l'enrichissement de la description des collections numériques.

Le terme de « crowdsourcing » désigne des projets collaboratifs de grande ampleur, mais l'accent sera davantage mis sur le nombre des participants, sur la notoriété du projet, sur la constitution de communautés de contributeurs, que sur la valeur scientifique de leurs contributions.

L'expression « métadonnées sociales » insiste quant à elle davantage sur l'enrichissement et l'amélioration de la description bibliographique.

I.3. Typologie

De nombreux projets collaboratifs sont d'ores et déjà mis en œuvre par des bibliothèques et plus largement des institutions culturelles un peu partout dans le monde, en s'appuyant souvent sur des collections patrimoniales numérisées. Il est possible d'établir une typologie de ces projets, qui ne peut toutefois être exhaustive tant l'imagination des professionnels en ce domaine est fertile.

- ***Tagging* et folksonomies**

L'usager peut être invité à enrichir l'indexation des ressources numériques par l'ajout de mot-clefs ou « *tags* ». Ce processus d'indexation et de classification collaborative, par des mots-clés librement choisis par chaque internaute, est appelé « folksonomie[10] ». Celle-ci n'apporte évidemment pas la qualité d'une indexation professionnelle normalisée et appuyée sur des référentiels contrôlés ; elle pose même des problèmes de polysémie, d'orthographe, d'absence de hiérarchie, ou encore de personnalisation des vocabulaires.

[9] L'archiviste américaine Kate Theimer définit ainsi les « archives participatives » : « un organisme, un site ou une collection auxquels des personnes qui ne sont pas des professionnels des archives apportent leur connaissance ou ajoutent des contenus, généralement dans un contexte numérique en ligne. Il en résulte une meilleure compréhension des documents d'archives. » Kate Theimer, « The participatory archives », *Archives Next*, 2011, http://www.archivesnext.com/?p=2319.

[10] Olivier Le Deuff, "Folksonomies. Les usagers indexent le web », *Bulletin des bibliothèques de France* (2006 - t. 51, n° 4), http://bbf.enssib.fr/consulter/bbf-2006-04-0066-002.

Mais le *tagging* social fournit une indexation simple, gratuite et rapide, appuyée sur une large communauté d'usagers, qui couvre potentiellement tous les domaines de la bibliothèque numérique et tous les types de documents. De plus, ces folksonomies sont conformes aux usages du web, elles s'expriment dans des vocabulaires simples et intuitifs qui correspondent aux modes de recherche en langage naturel des usagers[11].

Si le *tagging* des collections peut être intégré au sein de la bibliothèque numérique, malgré les risques liés à la faiblesse des interactions et à la difficulté d'obtenir une masse critique, les médias sociaux de partage de contenu (Flickr pour les photographies, Youtube ou Dailymotion pour les documents audiovisuels) restent le lieu privilégié pour ce type de service.

Par exemple, afin d'accroître la visibilité de ses collections sur le web, de s'intégrer dans des communautés d'usages collaboratifs, et d'étudier les impacts potentiels des folksonomies sur l'enrichissement du signalement et des modes de recherche des usagers, la Bibliothèque du Congrès diffuse depuis 2008 environ 4600 photographies anciennes sur Flickr [12], alliant ainsi la dissémination des contenus sur le web et l'ouverture à la participation des usagers.

En un peu moins d'une année, ces photographies ont été vues plus de 10 millions de fois, 7000 commentaires ont été saisis, et 67 000 tags ajoutés. La fréquentation de la bibliothèque numérique a augmenté de 20 % pendant cette période. La qualité des commentaires a permis la mise à jour et l'enrichissement de 500 notices bibliographiques, tandis que les tags apportent des compléments notables à l'indexation professionnelle (par exemple, des informations géographiques, des traductions, des relevés d'objets ou de couleurs présents sur les photos).

Au-delà des documents iconographiques, l'indexation peut également porter sur des documents audiovisuels, comme le montre le projet *Waisda?* de l'Institut néerlandais pour le Son et l'Image, qui propose, sous forme ludique, l'indexation collaborative des archives de la télévision, et qui a rencontré un excellent succès public (plus de 340 000 tags ajoutés pendant les 6 premiers mois)[13].

- **Indexation collaborative et constitution de bases de données**

A la différence des folksonomies où l'usager est laissé très libre de ses choix d'indexation, il est possible d'encadrer strictement les activités de dépouillement des usagers, pour permettre la constitution de bases de données structurées.

C'est le choix fait par de nombreux services d'archives, en France mais aussi aux Etats-Unis[14], pour le traitement de documents intéressant la généalogie. Une vingtaine de services d'archives français ont ainsi mis en place sur leurs sites web des modules d'indexation collaborative de documents nominatifs[15] (état-civil le plus souvent, mais aussi registres matricules militaires, recensements de population, etc.). La multiplication de ces services s'explique par leur succès, leur intérêt majeur pour l'amélioration de la recherche

[11] Olivier Ertzscheid, *Folksonomies et indexation sociale : le monde comme catalogue*, 2008, http://fr.slideshare.net/olivier/oe-abes-mai2008.

[12] *For the Common Good: The Library of Congress Flickr Pilot Project*, 2008, http://www.loc.gov/rr/print/flickr_report_final.pdf.

[13] Maarten Brinkerink, « Waisda? Video Labeling Game: Evaluation Report », *Images for the future*, 2010, http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/.

[14] Par exemple le projet 1940 US Census lancé en 2012 par les Archives nationales des Etats-Unis https://the1940census.com/, pour l'indexation des registres du recensement.

[15] Voir Edouard Bouyé, « Le web collaboratif dans les services d'archives publics : un pari sur l'intelligence et la motivation des publics », *La Gazette des Archives*, n°227 (2012-3).

dans les fonds d'archives concernés, mais aussi par l'intégration de ces modules dans les logiciels du marché.

Les chiffres sont parlants : aux Archives départementales de l'Ain, 500 000 pages ont été indexées en 2 ans ; aux Archives départementales du Cantal, 1000 micro-tâches d'indexation sont réalisées chaque jour…

- **Identification de documents iconographiques et catalogage collaboratif**

Les techniques de *crowdsourcing* sont utilisées tout particulièrement pour l'identification de documents iconographiques, auxquels il est impossible d'accéder par un moteur de recherche s'ils ne disposent pas d'un minimum de données descriptives. Plusieurs bibliothèques et services d'archives français ont ainsi mis en place d'efficaces outils collaboratifs d'identification de photographies, soit sur leur site institutionnel comme aux Archives de Haute-Garonne[16] ou encore sous forme ludique aux Archives de l'Ain avec les enquêtes « SOS détective[17] », soit sur un site spécifiquement dédié comme à la bibliothèque municipale de Lyon[18], soit encore sur des sites de partage comme Flickr, aux Archives des Alpes-Maritimes[19].

De même, les Archives nationales du Royaume-Uni ont lancé une très vaste opération d'identification de photographies intitulée « *Africa through a lens*[20] », diffusant sur leur site web et sur Flickr des milliers de photographies anciennes prises en Afrique, et appelant les internautes à mettre en commun leurs connaissances pour préciser l'identification des personnes et des lieux.

Projet d'initiative privée et individuelle, PhotosNormandie[21] constitue une opération originale en ce domaine, qui a pour objectif d'améliorer, via leur dissémination sur Flickr, les légendes des photographies de la Bataille de Normandie issues des collections des Archives nationales des Etat-Unis et du Canada, et de l'*Imperial War Museum* du Royaume-Uni, et publiées sur le site *Archives Normandie 1939-1945*[22]. Depuis 2007, près de 7000 descriptions d'un bon niveau scientifique ont ainsi été complétées et corrigées. Le succès du projet repose sur la constitution et l'animation dynamique d'un groupe de contributeurs qui ont largement utilisé les fonctionnalités sociales de Flickr pour échanger et commenter. Ici, le réseau social a eu un rôle de levier sur l'opération de *crowdsourcing*.

Outre une identification textuelle, le *crowdsourcing* peut permettre la géolocalisation de documents, comme l'interface ludique *Map Wraper* de la *New York Public Library*[23] qui propose de superposer des cartes anciennes de New York à des cartes actuelles, et calcule ensuite automatiquement les données géographiques.

Ce ne sont pas seulement les documents iconographiques que le *crowdsourcing* peut permettre d'identifier et de décrire, mais aussi des documents textuels. C'est ainsi un véritable catalogage collaboratif qui peut être mis en place sur des documents spécialisés complexes, pour lesquels la bibliothèque n'a pas forcément les compétences nécessaires en interne : par exemple, le catalogage de partitions musicales dans le projet « *What's the*

[16] http://www.archives.cg31.fr/archives_en_ligne/archives_identifier.html
[17] http://www.archives-numerisees.ain.fr/archives/enquete/enquetes_en_cours/n:77
[18] http://collections.bm-lyon.fr/photo-rhone-alpes/.
[19] http://www.flickr.com/photos/ad06
[20] http://www.nationalarchives.gov.uk/africa/
[21] http://www.flickr.com/people/photosnormandie/. Pour un bilan de l'opération : Patrick Pecatte, *PhotosNormandie a cinq ans – un bilan en forme de FAQ*, 2012, http://culturevisuelle.org/dejavu/1097.
[22] http://www.archivesnormandie39-45.org/
[23] http://maps.nypl.org/warper.

score at the Bodleian? » de la Bodleian Library[24], ou encore le catalogage de manuscrits en arabe à l'Université du Michigan[25].

- **Correction collaborative d'OCR et transcription collaborative**

Les utilisateurs peuvent également être invités à corriger un texte préalablement OCRisé, voire à le transcrire *ex nihilo*. Les techniques d'OCR automatique ne peuvent en effet pas obtenir des résultats complètement parfaits, seule une relecture humaine permet d'atteindre un taux de reconnaissance de 100 %. De plus, l'OCR n'est à ce jour efficace ni sur les écritures manuscrites anciennes ni sur les livres imprimés avant le XVII^{e} siècle ; là encore, seul l'œil humain permet de réaliser une transcription de ces documents, afin de disposer d'un mode texte nécessaire à la recherche plein texte, à la synthèse vocale pour les non-voyants ou encore à la réalisation de livres numériques.

Ce type de projets fera l'objet d'une étude approfondie dans le présent document.

reCAPTCHA : un programme de correction collaborative d'OCR, non PAR les bibliothèques mais POUR les bibliothèques numériques

Le reCAPTCHA [26] est un service anti-spam qui demande à l'internaute de transcrire deux mots qui lui sont soumis ; l'un est un mot test, et l'autre un mot mal reconnu par un logiciel d'OCR ; en transcrivant les deux mots, l'internaute contribue à améliorer la qualité du plein texte. Racheté par Google en 2009, cet outil est notamment utilisé pour la numérisation des archives du New York Times, et pour les ouvrages de Google Books.

- **Co-création de contenus scientifiques**

Les bibliothèques et services d'archives peuvent aussi ouvrir à leurs usagers la possibilité d'apporter le résultat de leurs propres recherches pour enrichir les contenus numérisés mis en ligne. C'est ainsi que la bibliothèque municipale de Toulouse propose dans sa bibliothèque numérique Rosalis[27] une rubrique « Rosalipédie » où les chercheurs comme les bibliothécaires peuvent commenter et analyser les documents. Les Archives départementales de Vendée ont ouvert en 2011 un L@boratoire des internautes[28] qui ouvre des propositions de participation variées, dont la possibilité de travaux scientifiques en réseau (par exemple constitution d'un guide des sources sur la guerre de Vendée).

L'un des projets les plus remarquables dans ce domaine est celui du wiki *Your Archives*[29] des Archives nationales du Royaume-Uni, plateforme d'écriture collaborative de textes scientifiques sur le patrimoine et l'histoire britanniques, appuyés sur les documents conservés aux Archives nationales et dans les autres services d'archives du Royaume-Uni. Lancé en 2007, *Your Archives* regroupe plus de 21 000 articles, rédigés ou corrigés par

[24] Un projet du réseau Zooniverse http://www.whats-the-score.org/.
[25] http://www.lib.umich.edu/islamic/
[26] http://www.google.com/recaptcha
[27] http://rosalis.bibliotheque.toulouse.fr/
[28] http://laboratoire-archives.vendee.fr/.
[29] http://yourarchives.nationalarchives.gov.uk/index.php?title=Home_page

31 000 utilisateurs inscrits. Cet outil est toutefois en cours d'évolution, pour s'adapter aux usages en mutation des internautes : les Archives nationales du Royaume-Uni viennent d'annoncer la fermeture de *Your Archives*, pour une intégration progressive du service de *crowdsourcing* et des contenus générés au sein même du nouveau catalogue de l'institution.

II. Correction d'OCR et transcription collaboratives dans les bibliothèques numériques : exemples commentés

Pour être en accord avec la 1e phase du projet de recherche, nous avons privilégié dans ce document l'étude des projets de *crowdsourcing* portant sur la correction collaborative d'OCR ou sur la transcription de documents imprimés et manuscrits. Les projets d'enrichissement tel que l'indexation, l'annotation ou l'ajout de contenus scientifiques pourront être étudiés ultérieurement.

II.1. Trove : correction collaborative d'OCR des périodiques de la Bibliothèque nationale d'Australie

Description du projet

La bibliothèque numérique Trove propose une stratégie globale et cohérente de *crowdsourcing* (tagging et commentaires) sur l'ensemble des collections. Le programme de correction collaborative d'OCR sur les périodiques numérisés reste toutefois l'aspect le plus innovant de l'ensemble. Mis en place depuis 2008, il propose aux internautes de participer à l'amélioration de la transcription de plus de 8 millions de pages (chiffres de janvier 2013). 2 millions de lignes de texte sont ainsi corrigées chaque mois par environ 30 000 volontaires. L'intégration de ce service au cœur même de la bibliothèque numérique permet de rendre immédiatement disponibles aux internautes les enrichissements apportés.

L'interface propose de nombreuses fonctionnalités pour rechercher un document et pour naviguer à l'intérieur de celui-ci. La manipulation des outils de correction est facile et intuitive. Les instructions sont claires. La correction d'un document s'effectue de ligne à ligne. L'interface permet d'insérer des caractères spéciaux.
Il est possible de corriger un document sans être authentifié. Seul un système anti-spam de reCaptcha permet de sécuriser l'intervention des utilisateurs anonymes.
Lors de la mise en place de la version beta, l'équipe s'est interrogée sur les risques encourus en permettant aux utilisateurs de modifier directement les textes. Plusieurs arguments autour de l'idée que les utilisateurs voudraient participer au « bien commun » les ont convaincus :

- La qualité des données est améliorée pour tous les utilisateurs
- La recherche par mot clé est améliorée pour tous les utilisateurs
- La communauté se trouve impliquée et engagée dans l'amélioration et l'enrichissement des contenus

Les utilisateurs peuvent consulter l'historique des corrections sur un article. Ils peuvent également gérer leur profil ou visualiser leurs activités récentes, de correction, de commentaire ou de marquage.

Pour des raisons de confidentialité, l'interface ne permet pas de mettre les utilisateurs en contact directement les uns avec les autres. Cependant la plateforme TROVE comprend un forum de discussion.

Par ailleurs, le code source de la plateforme de correction d'OCR de la Bibliothèque nationale d'Australie a été repris et adapté dans le cadre du projet PlaIR[30], mené par l'Université de Rouen et les Archives départementales de Seine Maritime.

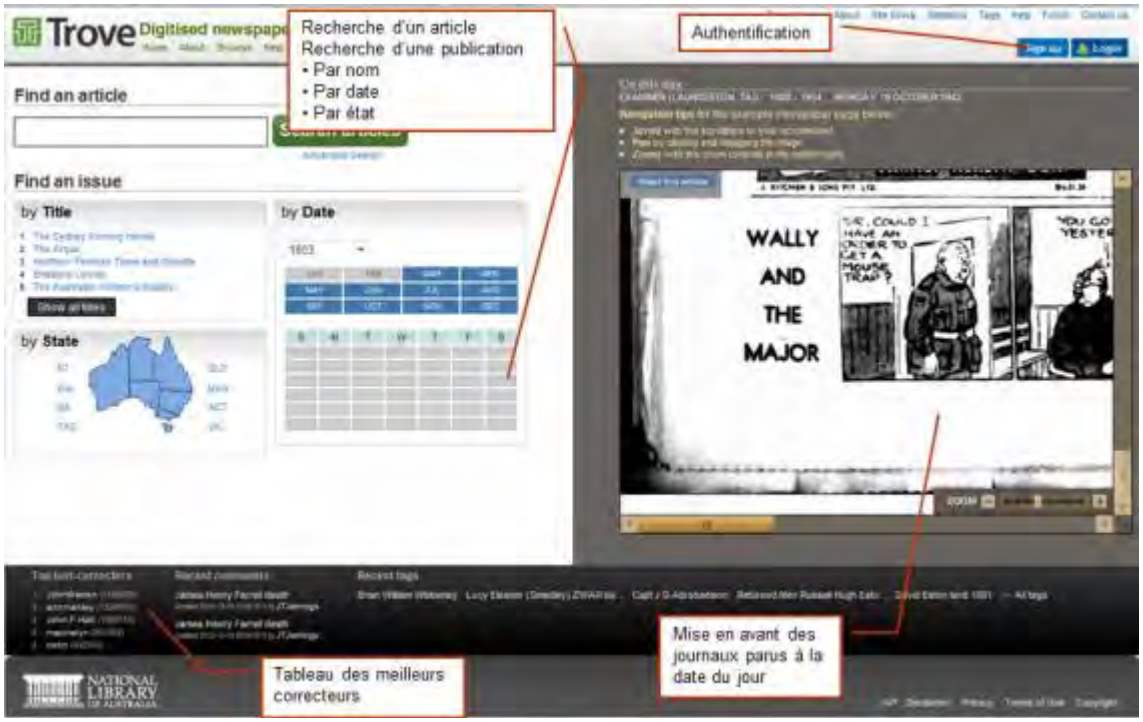
Facteurs de succès ou d'échec

Le succès de l'opération repose notamment sur une bonne animation de la communauté des contributeurs (valorisés par la mise en avant, chaque mois, des « *top correctors* »), sur une interface ergonomique et agréable, et sur l'intégration des contributions des internautes aux fonctionnalités de recherche, ce qui met en avant leur richesse et leurs apports et améliore notablement l'aisance de recherche dans les collections.

De plus, l'intérêt du sujet à traiter est un facteur de motivation. En effet, les journaux australiens de 1803 à 1954 sont uniques et internationalement recherchés. La section '*Shipping News*' des premiers journaux australiens est très importante pour les généalogistes car elle fournit des informations sur le mouvement des bagnards. La colonisation de l'Australie par les Britanniques est bien documentée ainsi que le traitement des peuples autochtones à cette époque. Ces journaux représentent ainsi une ressource précieuse pour les chercheurs et sont considérés comme un véritable patrimoine culturel et historique pour les Australiens.

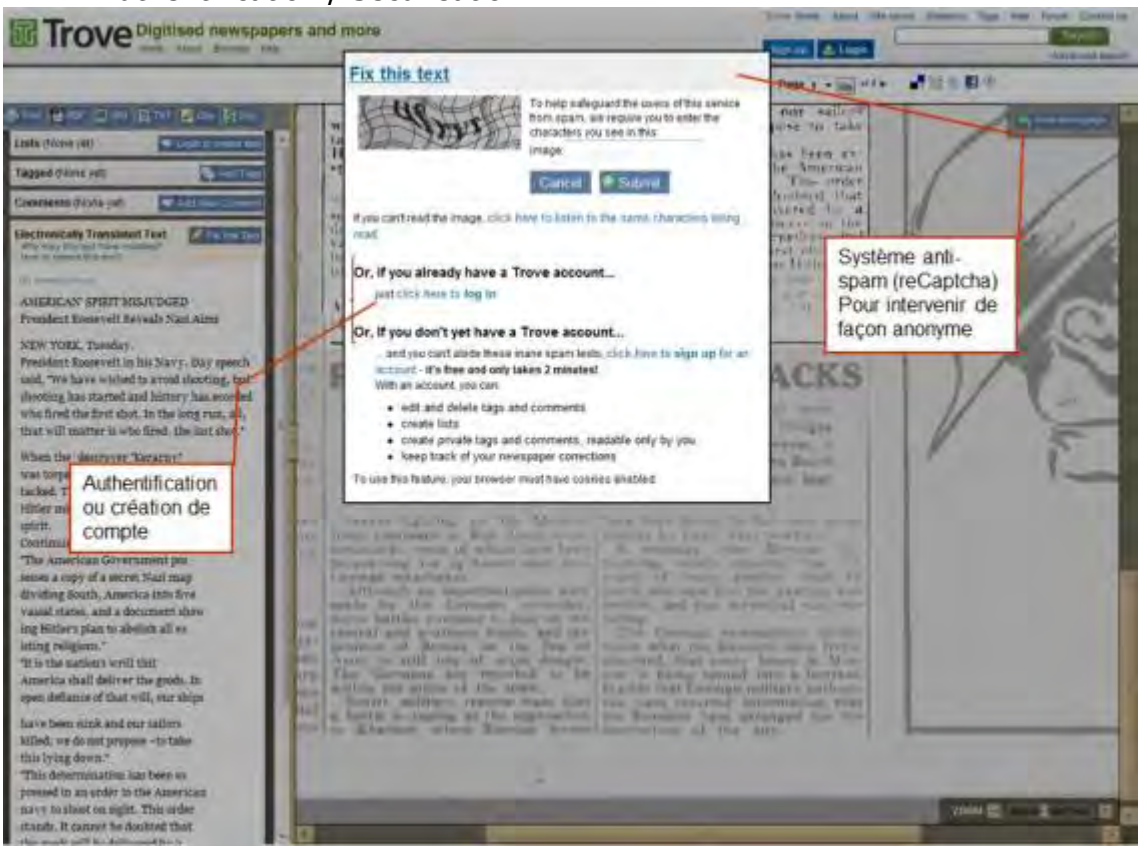
Copies d'écran

- Accueil



[30] http://plair.univ-rouen.fr
Présentation : https://ent.unr-runn.fr/filex/get?k=OzGVJUsedAyUPffwh8U

- Authentification / Sécurisation



- Consultation d'un document

- Correction d'une ligne de texte

- Gestion du profil de l'utilisateur

Bibliographie / webographie

Holley, Rose, Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers, Canberra: National Library of Australia. March 2009, 28 p. ISBN 9780642276940 [en ligne]
http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf

Holley, Rose, The Making of Our Digital Nation: Rose Holley at Mosman Library, [en ligne]
http://www.youtube.com/watch?v=a19icvJO_HE

Holley, Rose, "Crowdsourcing and social engagement in libraries: the state of play", 2010 [en ligne]
http://eprints.rclis.org/bitstream/10760/16385/1/Crowdsourcing%20State%20of%20Play%20June%202011.pdf

Holley, Rose, "Tagging Full Text Searchable Articles: An Overview of Social Tagging Activity in Historic Australian Newspapers August 2008–August 2009", dans : D-Lib Magazine., vol. 16, n°s 1/2, 2010, [en ligne] http://dlib.org/dlib/january10/holley/01holley.html.

Holley, Rose, "Crowdsourcing: How and Why Should Libraries Do it?", dans : D-Lib Magazine. Vol. 16, n°s 3/4, 2010, [en ligne] http://dlib.org/dlib/march10/holley/03holley.html

Holley, Rose, "How Good Can It Get? Analyzing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs", Dans : D-Lib Magazine, vol. 15, ns° 3/4, 2009, [en ligne] http://www.dlib.org/dlib/march09/holley/03holley.html

Holley, Rose, "A Success Story: Australian Newspapers Digitisation Program" Online Currents. Vol ; 23, n° 6, 2009, p. 283-295 http://www.nla.gov.au/ndp/project_details/documents/ANDP_StatsofANserviceusage_v1.pdf

II.2. Correction d'OCR et transcription collaborative sur Wikisource : l'exemple du partenariat avec la BnF

Description du projet

Wikisource, né fin 2003 dans la lignée des projets Wikimedia, a pour objectif de retranscrire sous forme numérique les livres publiés, tout particulièrement les ouvrages du domaine public. Son mode d'édition est le wiki, ce qui signifie que tout internaute peut contribuer à la relecture des livres sur la base du bénévolat ou encore y télécharger des œuvres libres.

La Bibliothèque nationale de France et Wikimédia France ont signé en 2009 un accord de partenariat pour l'ouverture à la transcription collaborative sur Wikisource de 1416 documents numérisés issus de la bibliothèque numérique Gallica. Il s'agit de pouvoir offrir finalement aux internautes ces œuvres transcrites, disponibles en plein texte, en s'appuyant sur l'interface de transcription proposée par Wikisource.

Les documents fournis par la BnF présentent des niveaux de difficulté variés, afin de tester les potentialités de la correction collaborative et d'étudier l'influence du taux de qualité sur l'activité des internautes : images numérisées seules (359 documents), ou accompagnées d'un OCR de qualité variable (1057 documents), soit 573 310 pages en tout.

La correction peut se faire en mode authentifié ou non authentifié (dans ce cas, comme pour tous les projets Wikimédia, c'est l'adresse IP qui identifie le contributeur). Chaque page doit être corrigée par un premier contributeur, puis relue par un second avant d'être considérée comme validée.

Un historique permet de suivre toutes les corrections effectuées sur une page, et par quel correcteur, et de retrouver une version précédente. Une page de discussion est ouverte pour chaque page de correction, afin que les contributeurs puissent échanger

entre eux sur d'éventuelles difficultés de correction ; cette fonctionnalité n'est toutefois presque jamais utilisée.

L'interface de correction permet de corriger le texte et d'apporter quelques éléments de structuration de la page (titres, notes de bas de page, etc.). Il s'agit d'un éditeur de texte dont la prise en main est simple pour des corrections ponctuelles ; pour des corrections plus lourdes en revanche (initialisation d'une page, indication de la segmentation d'un mot entre deux pages, etc.), des balises spécifiques doivent être ajoutées, qui sont expliquées dans un mode d'emploi complexe à appréhender. Un logiciel d'OCR intégré permet une automatisation partielle pour les textes non OCRisés.

Les analyses des statistiques de correction (non publiées) montrent que les correcteurs, peu nombreux mais très actifs, sont majoritairement des habitués de Wikisource.

Facteurs de succès ou d'échec

Le succès de l'opération s'avère mitigé, avec des volumes de corrections relativement faibles, qui ne permettent pas les analyses d'usages et de motivation initialement envisagées. Plusieurs causes en ont été identifiées :

- Faiblesse de la communication et de la médiation institutionnelle, qui n'a pas permis de « recruter » massivement les usagers de Gallica pour participer à la correction,
- Difficulté de prise en main d'une interface peu intuitive, qui nécessite une période de formation pour toute correction un peu complexe, ce qui exclut les potentiels contributeurs ponctuels,
- Difficulté de s'insérer dans un projet, Wikisource, et une communauté pré-existante, les Wikisourciens, externes à l'établissement.

Pour les documents qui ont été corrigés, on constate toutefois l'excellente qualité des corrections apportées : la double vérification permet d'obtenir une qualité presque parfaite.

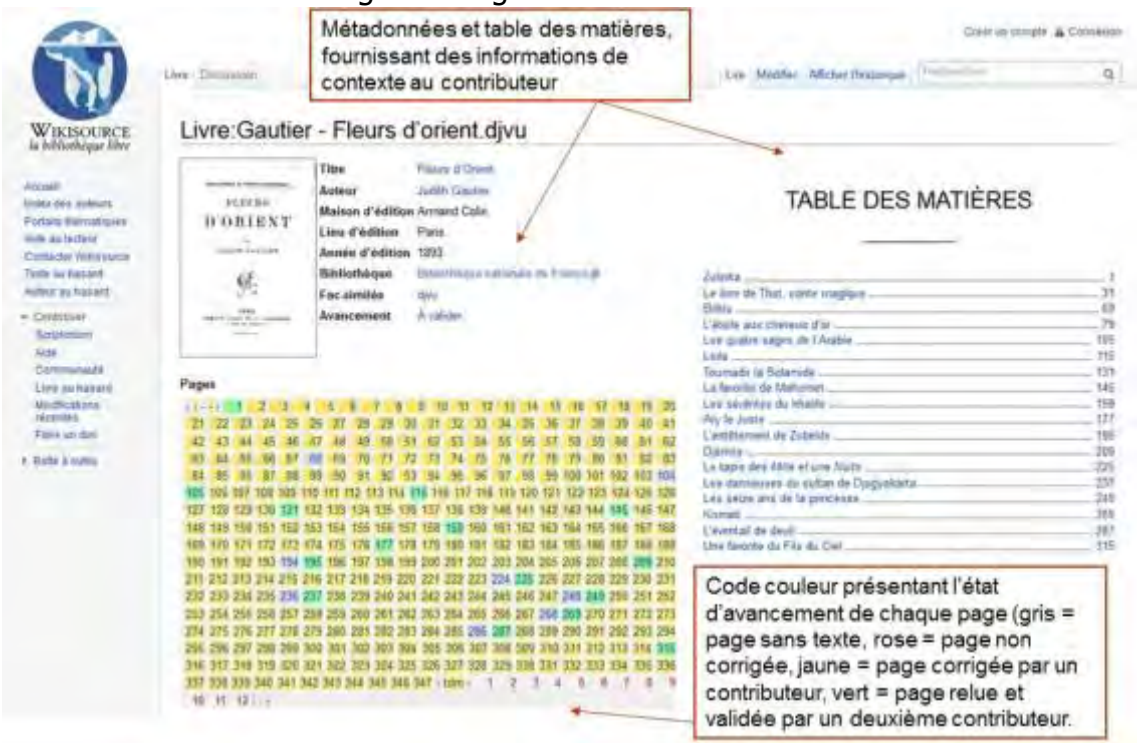
En revanche, le choix de l'exportation d'un tel projet sur un site extérieur à la bibliothèque, avec ses contraintes techniques propres, peut poser des problèmes de réintégration des données produites vers la bibliothèque numérique d'origine : le format DjVu utilisé par Wikisource ne contient pas d'informations de structure comme le format ALTO utilisé par Gallica (nécessaire à la mise en œuvre des fonctionnalités de recherche et de navigation dans les documents), et les fichiers corrigés ne peuvent donc pas être réintégrés automatiquement.

Afin de mettre en perspective les raisons du relativement faible nombre de contributions réalisées dans le cadre de ce partenariat, il est utile de le comparer à un autre partenariat mis en œuvre par les Archives départementales des Alpes-Maritimes. Cette institution a elle aussi fait le choix d'exporter sur Wikisource une activité de transcription collaborative, en ouvrant les fichiers numérisés de manuscrits de visites pastorales du XVIIe siècle.

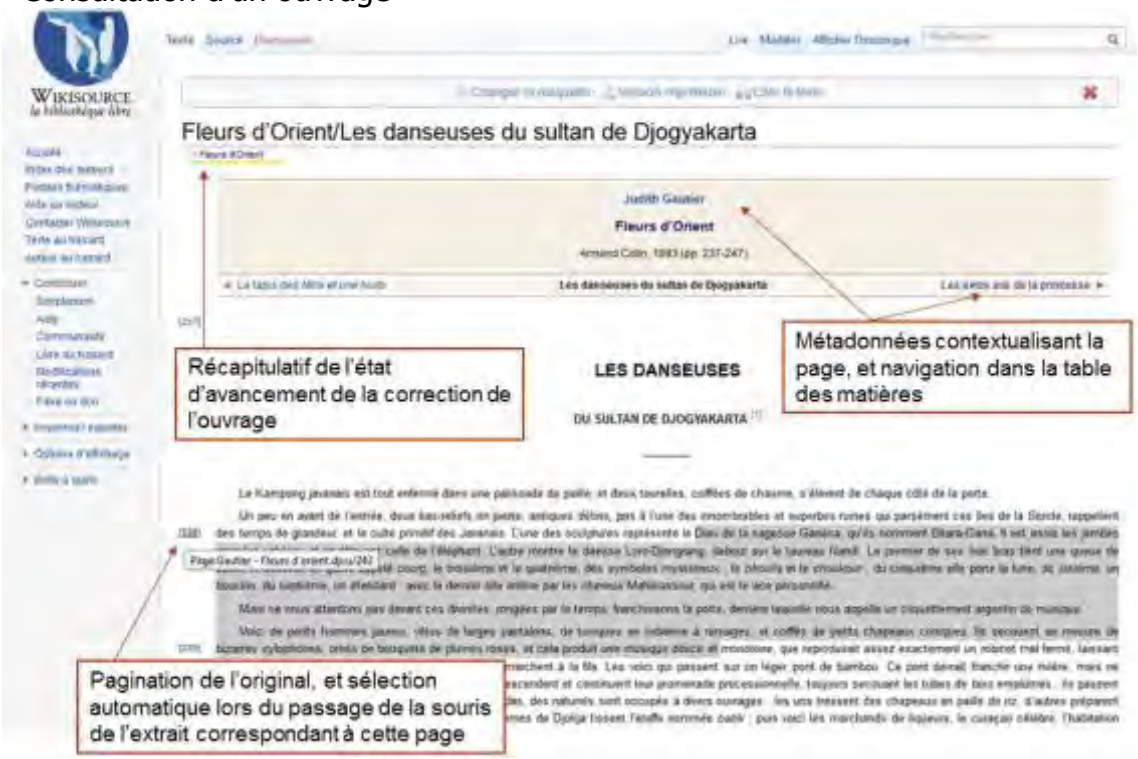
Le succès est au rendez-vous de ce projet modeste, grâce à un accompagnement très cadré des contributeurs, étudiants et amateurs coordonnés par une conservatrice des Archives départementales dans le cadre d'ateliers de paléographie, et soutenus par la communauté des wikisourciens. Dans ce cas, l'accompagnement et la formation des contributeurs a pu permettre de passer au-delà des difficultés de l'interface pour bénéficier des fonctionnalités efficaces de l'outil.

Copies d'écran

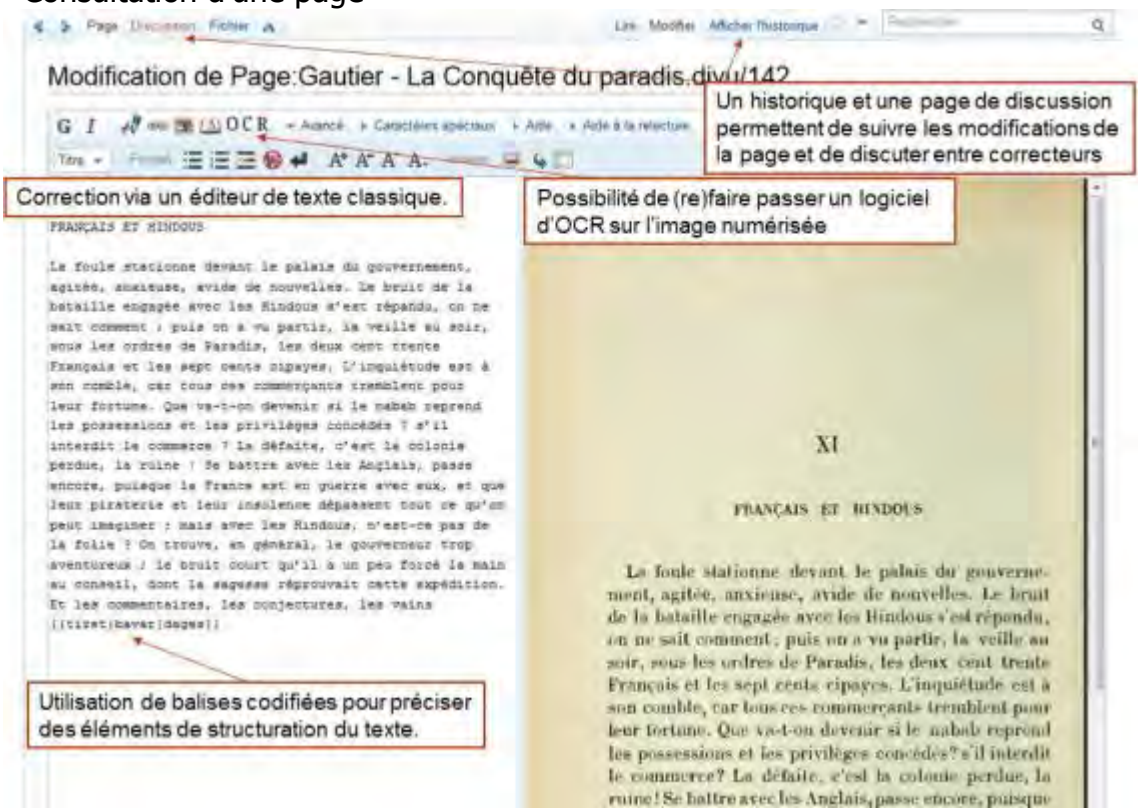
- Présentation d'un ouvrage à corriger



- Consultation d'un ouvrage



- Consultation d'une page



Bibliographie / webographie

« La BnF signe un partenairat avec Wikimédia France » (communiqué de presse, 2010), http://www.bnf.fr/documents/cp_wikimedia.pdf

Partenariat Wikisource / Gallica : page de présentation du projet http://fr.wikisource.org/wiki/Wikisource:Partenariats/Biblioth%C3%A8que_nationale_de_France

Partenariat Wikisource / Archives départementales des Alpes Maritime : page de présentation du projet http://fr.wikisource.org/wiki/Wikisource:Partenariats/Archives_D%C3%A9partementales_des_Alpes-Maritimes#On_en_parle

« Edition collaborative de manuscrits sur Wikisource », *La Tribune des archives*, 2012, http://latribunedesarchives.blogspot.fr/2012/02/edition-collaborative-de-manuscrits-sur.html

II.3. Correction collaborative d'OCR de la California Digital Newspaper Collection *(Etats-Unis)*

Description du projet

La *California Digital Newspaper Collection* contient plus de 400 000 pages de journaux californiens publiés entre 1846 et 1922, depuis le premier journal paru en Californie, le *Californian*. La collection inclut également des journaux contemporains. Ce projet fait partie du programme national « *USA's National Digital Newspaper Program* » (NDNP, 2003), piloté par la Bibliothèque du Congrès.
L'outil de correction de l'OCR a été fourni par DL Consulting sous la forme d'un module « *User Text Correction* » (UTC) ajouté à la solution logicielle Veridian, utilisée dans d'autres bibliothèques telles que Cornell, Princeton, Bibliothèque nationale de Singapour. Cet outil permet de corriger le texte ligne par ligne.

La numérisation est réalisée au niveau de l'article, au format METS/ALTO. Cependant, un document numérique de niveau page est également disponible, pour satisfaire les spécifications de la Bibliothèque du Congrès.

Depuis la fin de l'année 2011, la *California Digital Newspaper Collection* a pu évaluer l'ampleur des corrections réalisées :
…"In just 9 weeks, 96 users corrected nearly 50,000 lines of OCR text. The top text corrector alone improved over 10,000 lines. Furthermore, there was a 54% increase in the number of corrections made in month 2 compared with the first month. We have no reason but to expect further increases over time as the corrector community grows[31]"…
Un an après le lancement du module de correction, 309 utilisateurs avaient corrigé 400 000 lignes de texte. Remarquons qu'un petit pourcentage des utilisateurs réalise la majorité des corrections.

Une éventuelle corrélation entre l'ajout de la fonctionnalité de correction et la fréquentation du site est difficile à prouver, mais on note une augmentation du temps moyen passé sur le site et une diminution du nombre de pages vues, ce qui tendrait à démontrer le développement de l'activité de correction. Par contre, l'interaction entre le public et les conservateurs de la bibliothèque s'est considérablement accrue :
…"Many of those users emailed us directly with questions about or praise for the UTC, building direct, personal connections between our staff and users that hadn't existed before."

Facteurs de succès ou d'échec

+ L'interface de correction est totalement intégrée à l'interface de consultation de la bibliothèque numérique (pas de plateforme ou d'outil externe).
+ Accès au contenu au niveau article
+ Format interne standard (METS/ALTO)
+ Traces et statistiques (pour les utilisateurs et les administrateurs)
– Peu ou pas de fonctions sociales

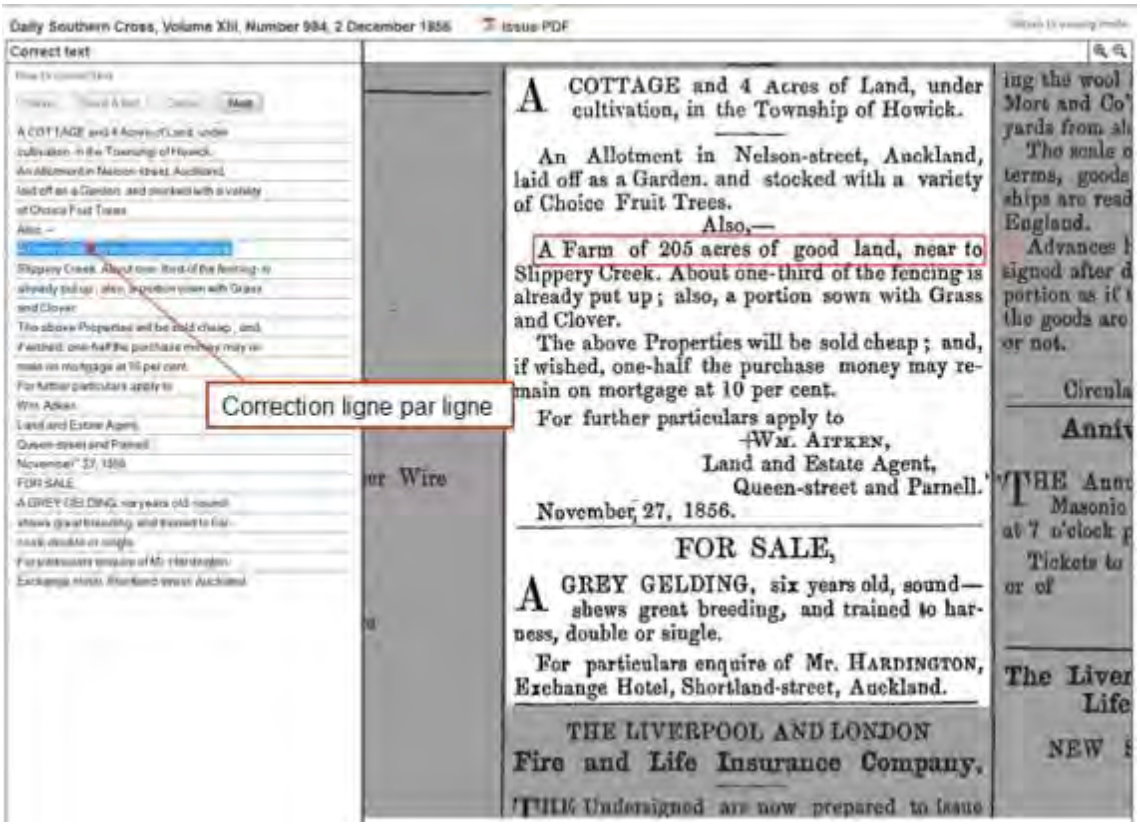
[31] http://www.dlconsulting.com/crowdsourcing/user-text-correction-results-at-cdnc

Copies d'écran

- Accueil

- Recherche d'une publication par date

- Consultation d'un document



Bibliographie / webographie

California Digital Newspaper Collection http://www.cdnc.ucr.edu/cdnc

Zarndt, Frederick & Geiger, Brian. *Experiences with User Text Correction at the California Digital Newspaper Collection,* dans : LITA National Forum in St. Louis, 2011. [résumé en ligne]

DL Consulting, *User Text Correction Results at CDNC*, November 9, 2011 [en ligne] http://www.dlconsulting.com/crowdsourcing/user-text-correction-results-at-cdnc/

Dewsnip, Michael. *Veridian User Text Correction demo site*. http://www.dlconsulting.com/crowdsourcing/veridian-user-text-correction-demo-site/

II.4. Digitalkoot : correction collaborative d'OCR à la Bibliothèque nationale de Finlande

Description du projet

Digitalkoot est un programme collaboratif pour l'amélioration de l'OCR et de la lisibilité des pages numérisées des collections patrimoniales de la Bibliothèque nationale de Finlande. Ce programme, ouvert au public en février 2011, repose sur le *crowdsourcing*. Les tâches de correction sont réalisées par « gamification », via les jeux Mole Bridge et Mole Hunt. Mole Hunt permet de faire valider par les joueurs les résultats de l'OCR alors que Mole Bridge, un peu plus complexe, permet de réaliser de la saisie de mots.

Le découpage des documents en microtâches élémentaires est fait à l'aide de MicroTask, une plateforme collaborative développée par IBM dans le programme de recherche Impact.
Les mots sont soumis à une validation visuelle ou bien à une correction manuelle ou une transcription totale par les internautes.

L'évaluation de l'efficacité des participants, et notamment la détection des comportements « déviants » (« trolls » et autres hooligans numériques), est menée en soumettant des mots issus de documents avec vérité terrain, ce lors des toutes premières minutes de jeu d'un nouvel utilisateur. Le système distribue des tâches dont la réponse est connue. Une fois que le joueur démontre qu'il joue correctement, la proportion des tâches de vérification diminue progressivement.

DigiTalkoot a été un grand succès : près de 110 000 participants ont répondu à plus de 8 millions de tâches de correction de mots (la population de la Finlande étant de 5.3 millions d'habitants). Durant les sept premières semaines, 5 000 utilisateurs ont réalisé 2 740 heures de correction.

La bibliothèque va poursuivre ses efforts avec Kuvatalkoot, un nouveau service permettant d'annoter des articles de journaux. Kuvatalkoot sera lancé au public d'ici la fin de l'année 2013.

Facteurs de succès ou d'échec

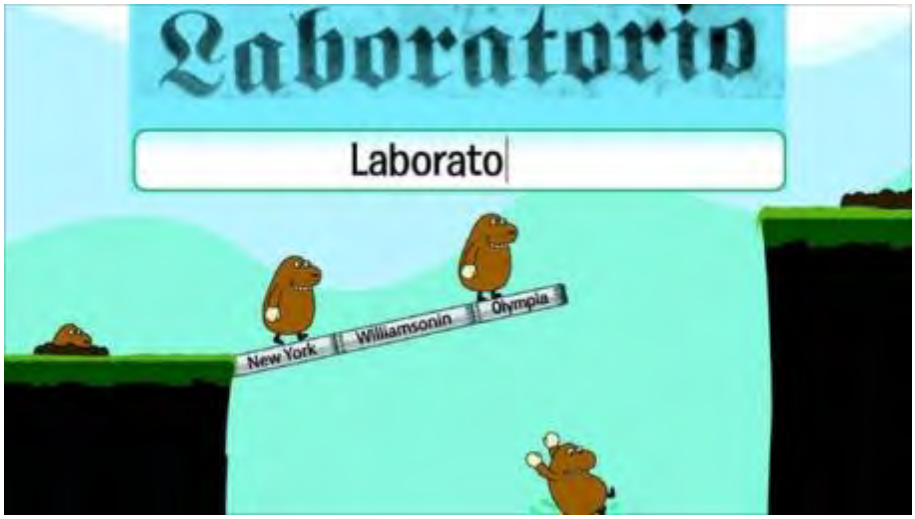
+ Gamification
+ Système de vérification de l'efficacité des correcteurs

+ Mise en œuvre d'une large redondance des corrections pour obtenir un taux OCR de 99 %, malgré la difficulté (police Fraktur)
+ Format interne standard (METS/ALTO)
+ Authentification via Facebook
+ Beaucoup d'écho dans les médias
– Peu ou pas de fonctions sociales, mais envisagé pour le futur

Copies d'écran

- Accueil

- Modules de jeu



Mole Bridge

Mole Hunt

Bibliographie / webographie

Digitalkoot : http://www.digitalkoot.fi/index_en.html

Blog Microtask sur YouTube http://www.youtube.com/user/microtaskblog

Benzinga. *OCR NL of Finland launches Europe's first national e-program for Crowdsourced archive digitization with microtasks* [blog en ligne] www.benzinga.com/press-releases/11/02/p845555/national-library-of-finland-launches-europes-first-national-e-program-f

Chrons, Otto and Sundell, Sami, "Digitalkoot: Making Old Archives Accessible Using Crowdsourcing", dans : *Association for the Advancement of ArtificialIntelligence*

(www.aaai.org), 2001, 6 p. et dans [en ligne]
http://cdn2.microtask.com/assets/download/chrons-sundell.pdf

De Benetti, Tommaso, *The secrets of Digitalkoot: Lessons learned crowdsourcing data entry to 50,000 people (for free)* : 16 juin 2011 [blog en ligne]
http://blog.microtask.com/2011/06/the-secrets-of-digitalkoot-lessons-learned-crowdsourcing-data-entry-to-50000-people-for-free/

Miettinen ,Ville. "Digitalkoot: electrifying the finnish cultural heritage", In : Proceedings of the 4th *ACM workshop on Online books*, complementary social media and crowdsourcing , N.Y. ACM, 2011, ISBN 978-1-4503-0961-5 [Résumé en ligne]
http://dl.acm.org/citation.cfm?id=2064071&dl=ACM&coll=DL&CFID=67658596&CFTOKEN=15897325

PRWeb Newswire. *Microtask Nominated as a Red Herring TOP 100 Europe Winner*. May 27, 2011, [blog en ligne] http://www.prweb.com/releases/Microtask/Red-Herring-TOP-100/prweb8497682.htm

Sterling, Bruce, *Digitalkoot, a game-ified social Finnish cultural endeavor*
http://www.wired.com/beyond_the_beyond/2011/03/digitalkoot-a-game-ified-crowdsourced-finnish-cultural-endeavor/

National Library of Finland Turns to Crowdsourcing, Games to HelpDigitize Its Archives
[blog] RedwriteWeb, Feb 2011

II.5. CONCERT (IBM Israël) *:* une plateforme de correction d'OCR développée dans le cadre du programme européen *IMPACT*

Description du projet

CONCERT (COoperative eNgine for Correction of ExtRacted Text) est un système de correction collaborative permettant de valider et corriger les résultats d'OCR. Cette plateforme de crowdsourcing fait partie des outils de post-correction d'OCR développés dans le cadre du Programme européen IMPACT[32].
Le système rationalise, simplifie et accélère le processus de validation du texte OCRisé. Il permet une validation rapide des résultats de l'OCR et repose sur une superposition de contextes de correction (Caractère => Mot => Page) qui décomposent les processus de vérification en plusieurs tâches complémentaires.

Dans un premier temps les utilisateurs ne travaillent qu'au niveau du caractère et déterminent, sur la base d'un échantillon de formes similaires, quelles sont les erreurs potentielles de l'OCR. La validation s'effectue ainsi en un seul écran. Les caractères rejetés sont alors vérifiés dans le contexte du « mot ».

[32] IMPACT est un projet d'intégration à grande échelle financé par la Commission européenne dans le cadre du septième programme-cadre (7e PC). L'un des objectifs du projet est de développer des outils qui aident à améliorer les résultats de l'OCR pour les textes imprimés historiques, en particulier ceux des ouvrages publiés avant la production industrielle de livres à partir du milieu du 19ème siècle.

Ainsi, *« Au lieu d'afficher une page entière numérisée, les examinateurs ne voient que les vraies lettres ou des mots en question. Par exemple, la combinaison des lettres "r" et "n" ("rn") peut apparaître impossible à distinguer de la lettre «m». Dans ces cas, le système recueille de nombreux cas, de la lettre «m» et met ces échantillons à côté des lettres en question, ce qui rend beaucoup plus facile de déterminer la véritable identité de la lettre.* »

Dans les cas où un mot entier est suspect, il est ajouté à une collection d'autres termes douteux, qui sont ensuite classés par ordre alphabétique. La vue « Mot » rassemble des termes considérés comme non fiables. Les utilisateurs doivent accepter ou de rejeter des substituts proposés. En outre, le système utilise une base de donnée partagée et évolutive, une méthode dans laquelle de nouveaux mots sont ajoutés à un dictionnaire central basé sur le recoupement d'identification et de correction par d'autres utilisateurs.
La dernière étape permet, dans un contexte de « page », d'identifier les faux positifs et de corriger les erreurs de segmentation.

Deux interfaces de jeu (web et smartphone sous Android) ont été développées autour des tâches à réaliser. Le programme européenn IMPACT se poursuit dans le cadre d'un centre de compétence qui va développer d'autres fonctionnalités.

Le système développé par IBM envisage une approche qui rémunère les utilisateurs. L'outil intègre un suivi de la performance des utilisateurs. Le contrôle de la qualité est mesuré notamment par l'insertion de pseudo-erreurs à partir desquelles est établi un pourcentage de caractères manqués. De même, pour former les nouveaux utilisateurs, il est envisagé de présenter des écrans déjà corrigés par d'autres.

Facteurs de succès ou d'échec

La séparation des processus de vérification en plusieurs tâches complémentaires est intéressante. Chaque outil proposé est adapté à la tâche demandée. L'ensemble est cohérent et relativement intuitif.

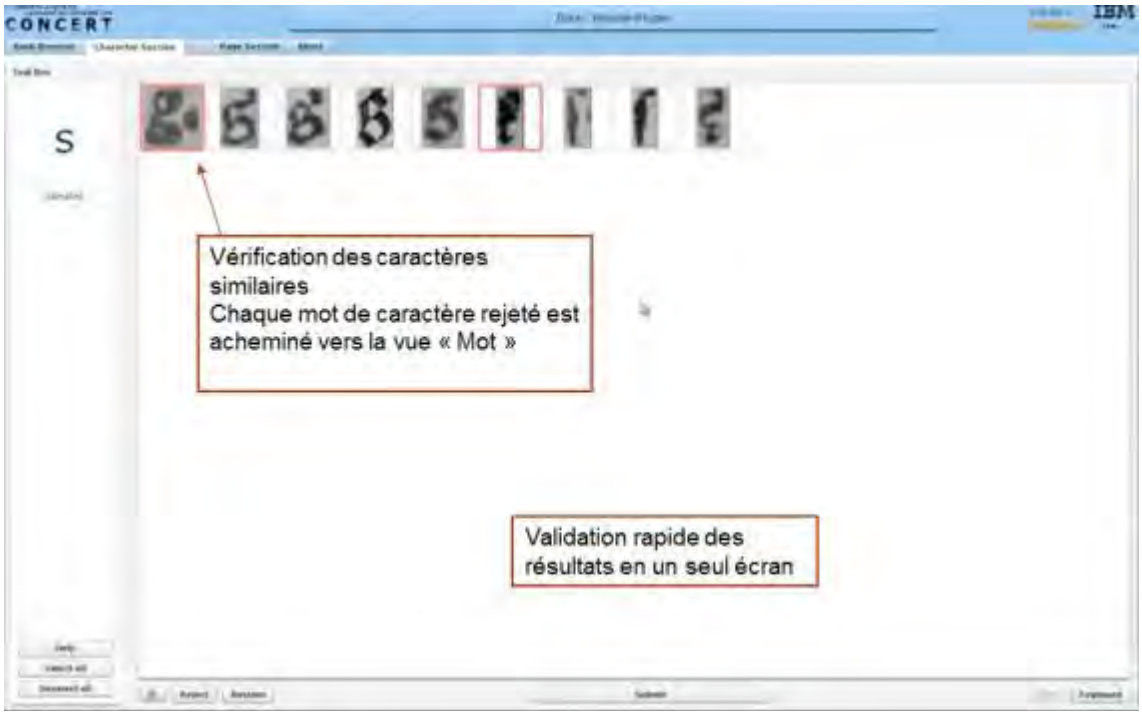
L'organisation en plusieurs étapes contextualisées donne un sentiment de progression dans les tâches à réaliser.
Un autre avantage de cette méthode de décomposition des tâches, est que la charge de travail peut être attribuée aux utilisateurs en fonction de leurs compétences*. « Par exemple, dans le traitement de remboursement de frais médicaux, de la simple reconnaissance de chiffres sera effectuée par les employés de niveau de base, tandis que la validation des noms de maladies sera confiée à des personnes ayant une certaine expérience dans le domaine médical. »*

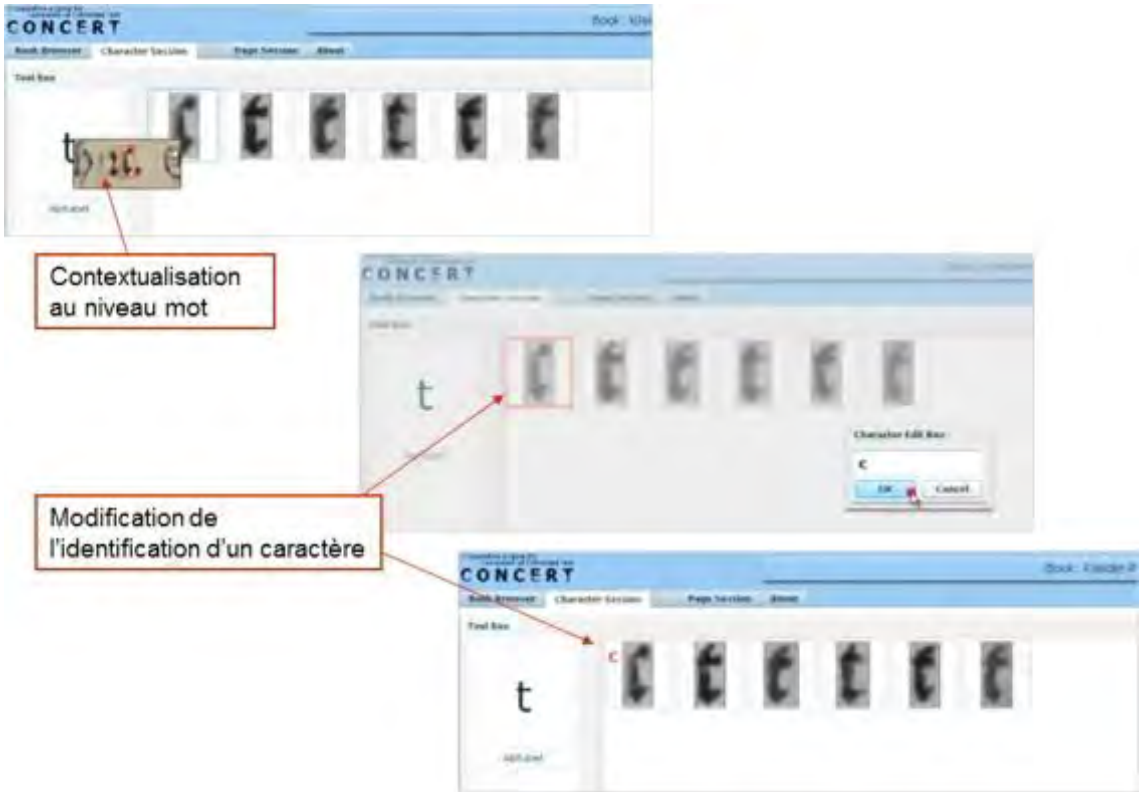
Mais il n'y a pas d'organisation de la collaboration en ligne. Les tâches sont réalisées en parallèle. Le projet appartenant à un projet pilote, le retour d'expérience des utilisateurs n'est pas encore forcément significatif. De même l'interface n'est pas très ergonomique.

Copies d'écran

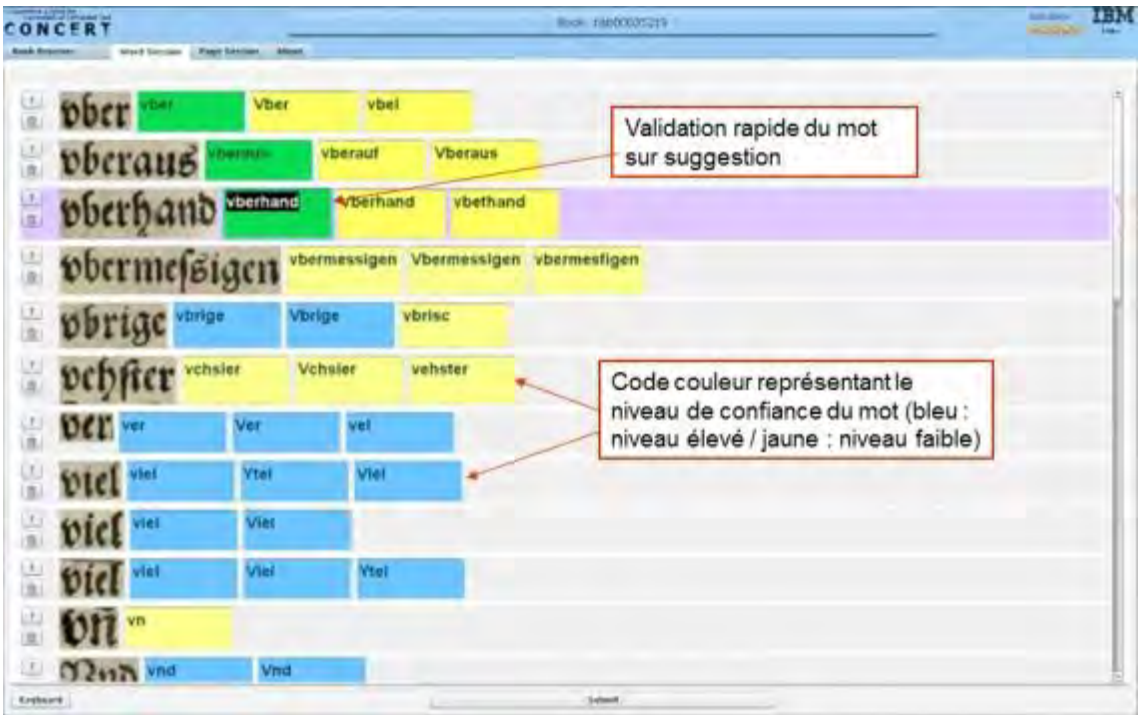
- Accueil

- Vue « Caractère »

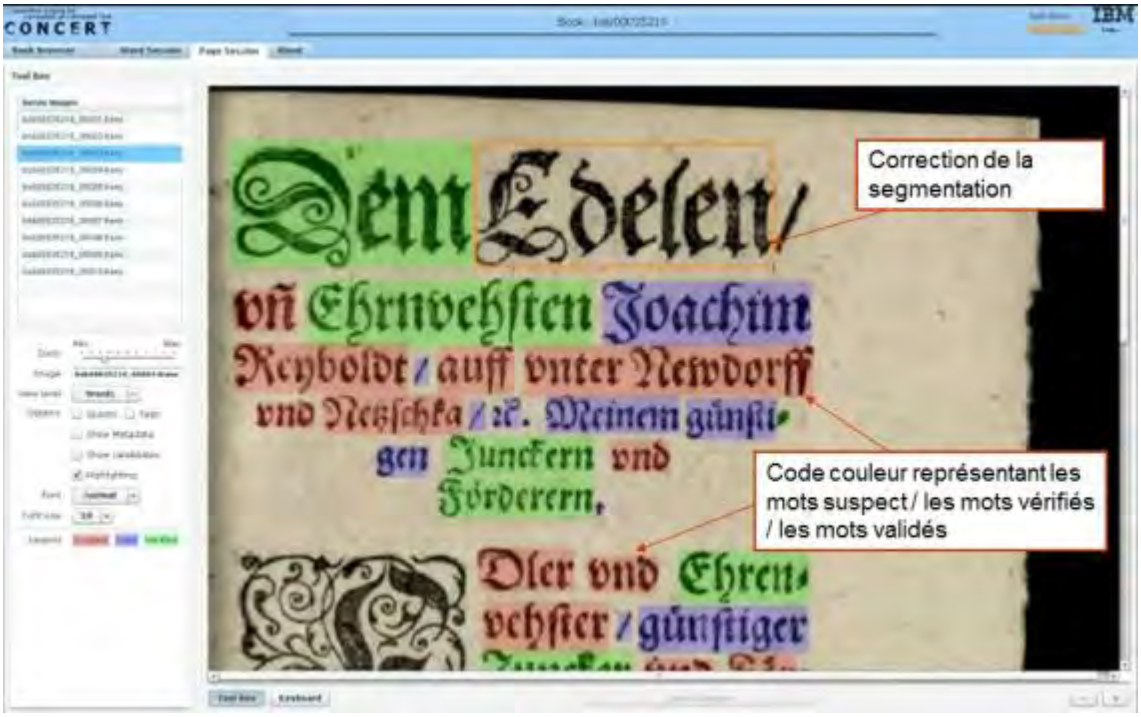




- Vue « Mot »



- Vue « Page »



- Interfaces de jeu

Bibliographie / webographie

Karnin, Ehud D. ; Walach, Eugene ; Drory, Tal. "Crowdsourcing in the Document Processing Practice A Short Practitioner/Visionary Paper", dans : Computer Science, vol. 6385, 2010, Current Trends in Web Engineering, p. 408-411 [en ligne pdf] http://www.springerlink.com/content/l54q178rm08j6120/fulltext.pdf

http://www.youtube.com/watch?v=JCAzn0tcvQM [10mn48] / Mark-Oliver Fischer, Bibliothèque d'état de Bavière. (pr le Munich Digitisation Centre)

Neudecker, C. and A. Tzadok, User Collaboration for Improving Access to Historical Text LIBER2010 Annual Conference, 29 June - 1 July 2010, Arhus, Denmark. Also published as a paper in LIBER Quarterly, vol. 20 (2010) no.1.

CONCERT - COoperative eNgine for Correction of ExtRacted Text. IMPACT Final Conference 2011, 24-25 October, London, UK
Presentation: IMPACT Final Conference - Asaf Tzadok

Projet IMPACT, http://www.impact-project.eu

Communiqué de presse : http://www-03.ibm.com/press/us/en/pressrelease/32380.wss

II.6. **Transcribe Bentham** *: transcription collaborative des œuvres de Jeremy Bentham*

Description du projet

Transcribe Bentham est un projet de transcription massive des 60 000 manuscrits non édités du philosophe anglais Jeremy Bentham. Ce programme a été initié par l'*University College of London* (UCL) dans le cadre du *Bentham Project* dont l'ambition est d'aboutir à l'édition des œuvres intégrales du philosophe. 4 760 manuscrits ont été transcrits ou partiellement transcrit à ce jour, dont 94 % sont complétement terminés.

Les utilisateurs doivent s'inscrire préalablement pour participer à la transcription des documents. Le tableau de bord du projet (*Transcription Desk*) permet de choisir un manuscrit selon différents critères (thématique, chronologique, document non encore retranscrit ou partiellement retranscrit, voire en fonction du niveau de difficulté). Il offre aussi la possibilité d'obtenir une page au hasard.

L'interface de transcription est intuitive. Une barre d'outils permet d'apposer simplement des balises XML pour signaler les exergues, ratures, passages à la ligne, paragraphes, etc… Un guide d'utilisation et des vidéos de démonstration accompagnent les premiers pas des nouveaux arrivants.
Les utilisateurs peuvent échanger des idées, se poser des questions ou contacter les administrateurs du projet via un forum de discussion. Le "*Benthamometer*" affiche les progrès de la transcription, tandis que le tableau de classement valorise les utilisateurs les plus assidus. Dans l'optique de motiver les utilisateurs, un système de notation a été conçu sur la base des modifications apportées aboutissant à un classement allant de "stagiaire" à "prodige".

La validation des corrections passe par une vérification de l'équipe de chercheurs de UCL, qui peuvent modifier a posteriori le texte ainsi que le code XML. Le manuscrit doit avoir été étudié au préalable par un nombre suffisant d'utilisateurs pour qu'une comparaison des sources aboutisse à un résultat fiable. La transcription validée est alors verrouillée dans la base de données du Transcription Desk, puis l'équipe de chercheurs

décide si un manuscrit peut être transmis au service d'édition (numérique et papier) des œuvres complètes de Bentham.

L'objectif étant de créer une communauté soudée autour du projet, plusieurs outils ont été mis en place pour la mobiliser et l'animer : une page Facebook, un compte twitter, un blog publiant les progressions mensuelles du projet. L'équipe projet a même organisé une série d'événements de sensibilisation du public. Mais cette dernière stratégie a eu un succès limité.

Le projet Transcribe Bentham a reçu en mai 2011 le prix Ars Electronica dans la catégorie Digital Communities (tout comme Wikipedia en 2004 et Wikileaks en 2009).

Facteurs de succès ou d'échec

Ce projet est particulièrement intéressant pour la réflexion qui a été menée et les analyses qui ont pu être faites en termes de recrutement des volontaires, d'approche communautaire et d'animation.

Pour toucher le plus grand nombre, l'équipe projet a fait le choix d'investir dans la communication et la publicité (communiqué de presse, dépliant remis lors des conférences, vidéo, mailing…). Un compte Google AdWords a également été créé en vue de générer du trafic, mais s'est avéré un échec comme stratégie de recrutement.
En termes de sensibilisation, la campagne de communication a été un succès. Le projet a reçu une couverture médiatique dans 12 pays et a été mentionné dans environ 70 blogs, 13 articles de presse et 2 émissions de radio.

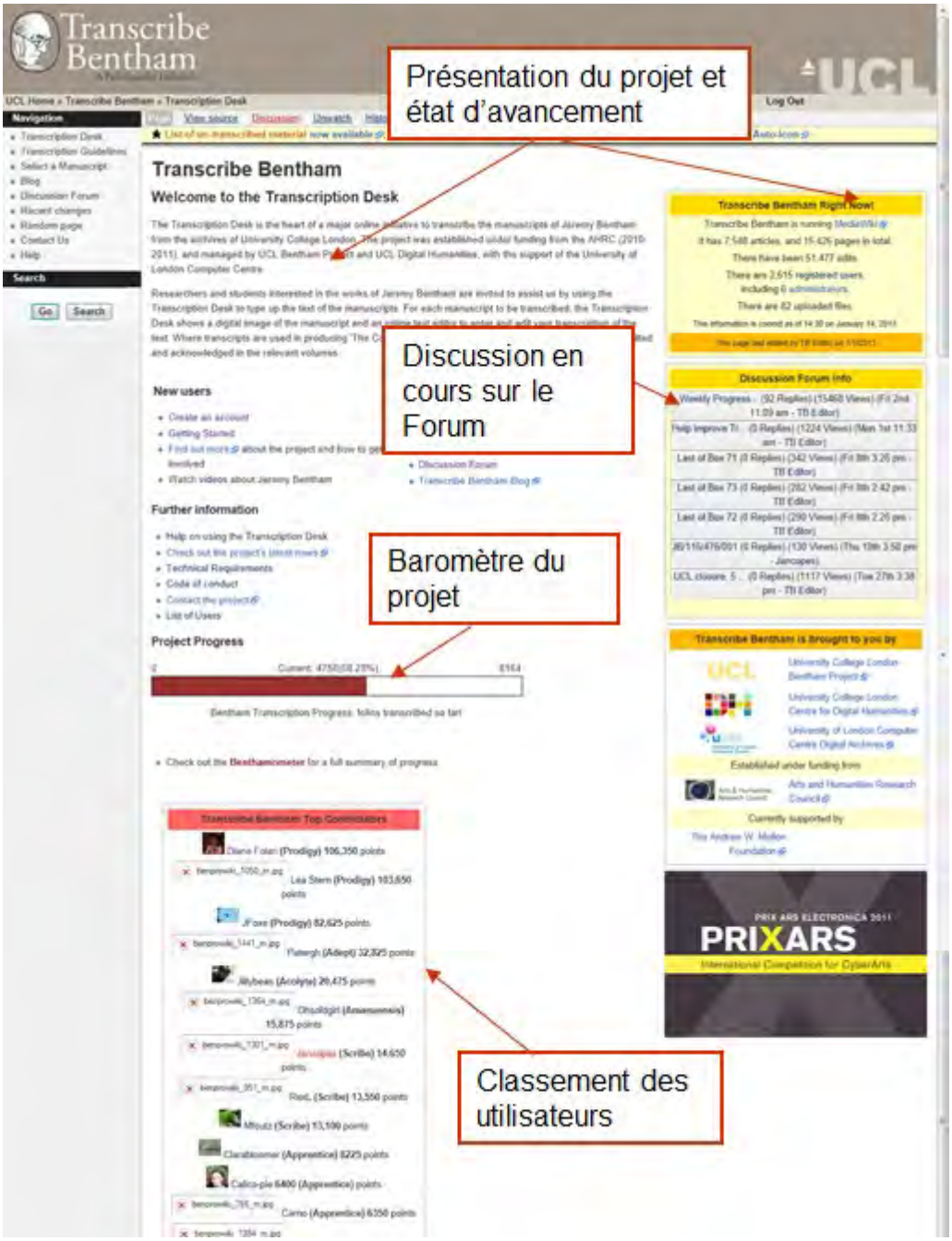
Suite à la publicité qui a entouré le lancement du projet, 1 115 visites ont été réalisées sur le site au cours de la première semaine. Puis grâce à la parution d'un l'article du New York Times, le site est passé de 11 visites le 26 décembre à 1140 visites le 27 décembre 2010. Pour poursuivre une présence visible et interactive en ligne, le blog du projet met à jour régulièrement des rapports d'étape, les détails de la couverture médiatique et des présentations à venir. Les médias sociaux tels que Twitter et Facebook semblent avoir eu peu d'impact pour générer du trafic directement sur le site, mais ont permis d'animer la communauté.

Un sondage a été réalisé et a indiqué que la plupart des utilisateurs ont été motivés de participer par un projet d'intérêt général. Mais de nombreuses personnes ont également trouvé le défi intellectuel de la transcription motivante.
Une grande majorité de ceux qui ont visité le site ne sont pas devenus des utilisateurs actifs. Certains ont été découragés par la difficulté de déchiffrer l'écriture de Bentham. Pour beaucoup, le codage de texte a rajouté à la complexité de la tâche.

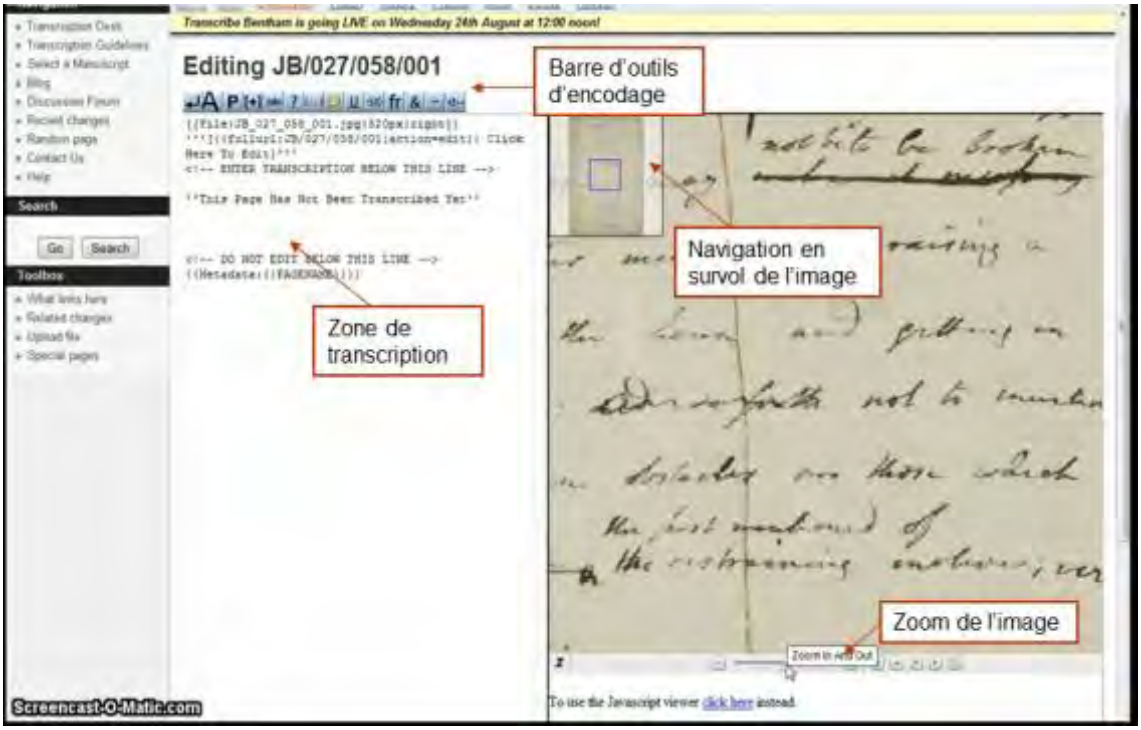
Peu de collaboration entre contributeurs a été constatée. Les fonctions sociales du site n'ont pas été vraiment utilisées. Mais le sondage a révélé le besoin pour les novices de demander de l'aide aux utilisateurs expérimentés et les outils à disposition ne semblaient pas adaptés.

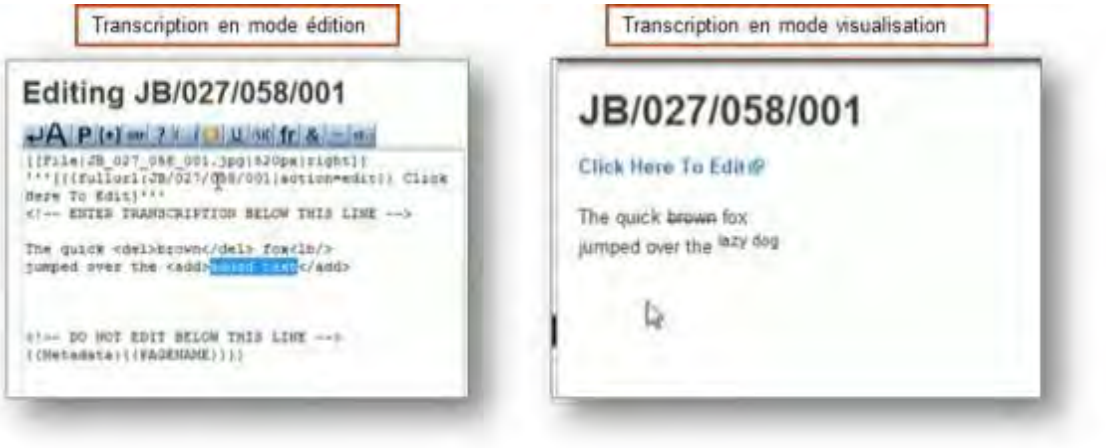
Copies d'écran

- Accueil



- Interface de transcription





- Page de sélection d'un manuscrit

Bibliographie / webographie

Les sites du projet Transcribe Bentham :

- Transcribe Bentham: http://www.ucl.ac.uk/Bentham-Project/transcribe_bentham
- *Transcription Desk* : http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham
- Blog : http://blogs.ucl.ac.uk/transcribe-bentham/
- Page Facebook: https://www.facebook.com/TranscribeBentham
- Twitter : https://twitter.com/TranscriBentham
- *Transcribe Bentham* video: http://www.youtube.com/watch?v=CtEqW4WwMHU .

Causer, Tim, and Wallace, Valerie. 'Building a volunteer community: results and findings from Transcribe Bentham', *Digital Humanities Quarterly*, vol. 6, no. 2, 2012.

Causer, Tim, Tonra, Justin, and Wallace, Valerie. 'Transcription maximized; expense minimized: crowdsourcing and editing The Collected Works of Jeremy Bentham', *Literary and Linguistic Computing*, vol. 27, no. 2, 2012.

Achaleke, Beatrice, Graham Harwood, Aaron Koblin, Liu Yan, and Tiago Peixoto, 'Guinea Pigs and Apples: Digital Communities category jury summary", in Hannes Leopoldseder, Christine Schöpf, and Gerfried Stocker, Cyberarts 2011: the International Compendium of the Prix Ars Electronica 2011. Hatje Cantz, 2011.

Benotmane, Ghita, *Je transcris, tu transcris, nous crowdsourçons : le projet Transcribe Bentham,* Introduction aux humanités numériques, 07/11/2012. http://archinfo41.hypotheses.org/93

II.7. Ancient Lives, un projet de « sciences citoyennes »

Description du projet

Ancient Lives, lancé en 2011, est issu d'une collaboration entre chercheurs et institutions spécialisés en papyrologie, et est coordonné par l'Université d'Oxford. Le projet s'insère dans le réseau Zooniverse, maintenu par la *Citizen Science Alliance*, un partenariat entre des universités et des musées dont l'objectif est d'associer la participation d'amateurs à des travaux de relevés, de dépouillements et d'identification scientifiques. Si la plupart de ces projets de « sciences citoyennes » (dont le tout premier, *Galaxy Zoo*, destiné à l'identification de galaxies) s'exercent dans le domaine des sciences dures ou des sciences naturelles, deux d'entre eux s'appuient sur des collections patrimoniales :

- *Old Weather* : projet de transcription et géolocalisation collaborative des relevés météorologiques manuscrits réalisés par les navires de la Marine royale au début du XX^{e} siècle et conservés aux Archives nationales du Royaume-Uni. L'objectif est de disposer de bases de données météorologiques complètes et fiables, sur lesquelles les météorologues pourront appuyer des études scientifiques pour comprendre et modéliser le climat d'aujourd'hui et ses évolutions demain.
- *Ancient Lives*, transcription collaborative de centaines de milliers de fragments de papyri de l'Egypte gréco-romaine (uniquement en grec dans une première étape), afin de les identifier, de les publier et de les mettre à disposition des chercheurs.

Entre juillet 2011 et décembre 2012, plus de 1,5 million de tâches de transcription ont ainsi été réalisées, qui ont permis l'identification d'une centaine de textes, dont des œuvres littéraires de Plutarque et d'Euripide.

Pour chaque document, trois interfaces différentes sont proposées, qui correspondent à des actions différentes de l'utilisateur :

- Interface de transcription
- Interface de mesure des marges et de l'espacement des colonnes. Cette activité de structuration du texte est un outil pour l'identification du contenu des documents, par le repérage d'auteurs ou d'ateliers d'écriture qui pratiquent les mêmes règles de structuration des papyri.
- Interface sociale

La correction peut se faire en mode identifié ou non identifié. En mode identifié, il s'agit d'un profil utilisateur commun à l'ensemble des projets du réseau Zooniverse. Les fonctionnalités de transcription et de mesure des documents sont ouvertes à tous, mais il est nécessaire d'être identifié pour bénéficier des fonctionnalités sociales.

Transcription et mesure

Les deux interfaces correspondant aux activités de *crowdsourcing* (transcription et mesure) sont particulièrement attractives et intuitives, presque ludiques, avec des possibilités de personnalisation (changement des couleurs, mais aussi affichage d'une « *lightbox* » c'est-à-dire une galerie d'images regroupant les fragments sur lesquels un utilisateur a travaillé) et de nombreuses fonctionnalités facilitant la lecture (zoom, rotation d'image, défilement horizontal, etc.). La prise en main est très aisée, et un tutoriel clair et illustré est proposé lors de la première utilisation.
Un bouton « *Issue* » permet à l'usager de signaler des problèmes sur le document : image trop sombre ou trop claire, fragments qui ne sont pas en grec, etc.
Il s'agit fonctionnellement d'une succession de micro-tâches (transcription d'une lettre, mesure d'une marge), relativement accessibles à des non-spécialistes (à condition de connaître l'alphabet grec), mais les commentaires montrent que de nombreux contributeurs ont une bonne connaissance de la langue et de la paléographie grecques et peuvent s'appuyer sur la compréhension des mots pour déchiffrer les caractères difficilement lisibles.
Le travail d'identification des textes, pour lequel aucune interface spécifique n'est prévue à part la page d'interactions sociales, est facilité par une fonctionnalité de « *Match* », qui permet de rechercher les occurrences des mêmes groupes de lettres dans des corpus de textes grecs en ligne.
Les fragments sont proposés à la transcription de façon semble-t-il aléatoire, le correcteur ne peut pas les choisir (en réalité, ils sont poussés par les responsables du projet en fonction de leurs priorités). Les mêmes fragments sont attribués à plusieurs correcteurs pour croiser les transcriptions et améliorer la qualité des résultats.
Il n'est pas possible de consulter les transcriptions réalisées par d'autres correcteurs ou de faire des recherches dans le corpus.

Fonctionnalités sociales

Les fonctionnalités sociales, ouvertes uniquement aux usagers authentifiés, sont accessibles sur un sous-site dédié appelé « *Talk* » (mais il est alors nécessaire de se ré-identifier, ce qui est pénible).
La page d'accueil de ce sous-site regroupe et met en avant les interactions entre usagers, les commentaires sur les documents, les interventions des administrateurs, les forums apportant des conseils aux correcteurs.
On y trouve pour chaque document une page d'interactions ouvrant des espaces de discussion et de commentaires (la différence entre les deux types d'échanges n'est pas claire et n'est pas explicitée). Les correcteurs peuvent y saisir des informations sur l'identification du texte présent sur le papyrus, mais aussi soulever des problèmes ou poser des questions. Ces espaces sont fortement animés par les administrateurs du site, des chercheurs spécialistes eux aussi, qui répondent aux questions et guident les correcteurs.
L'espace social permet également aux utilisateurs de se créer des collections, en regroupant des fragments de papyri.

On peut consulter les profils des autres contributeurs, qui regroupent leurs discussions, leurs commentaires, leurs collections, savoir s'ils sont en ligne et les contacter par un service de messages directs.

Communication

Une intense communication a accompagné le lancement du projet en 2011, dans la presse, mais aussi avec une émission sur la BBC. Un compte Twitter @ancientlives a été ouvert à ce moment-là, mais il n'est plus utilisé depuis.
Depuis, un blog assure avec régularité la communication auprès des transcripteurs volontaires : informations sur l'avancement du projet, conseils pour la transcription et l'utilisation de l'interface, informations paléographiques ou diplomatiques pour aider à la transcription (par exemple des billets sur les abréviations utilisées à cette époque, sur les types de documents, les formules de politesse, etc.).

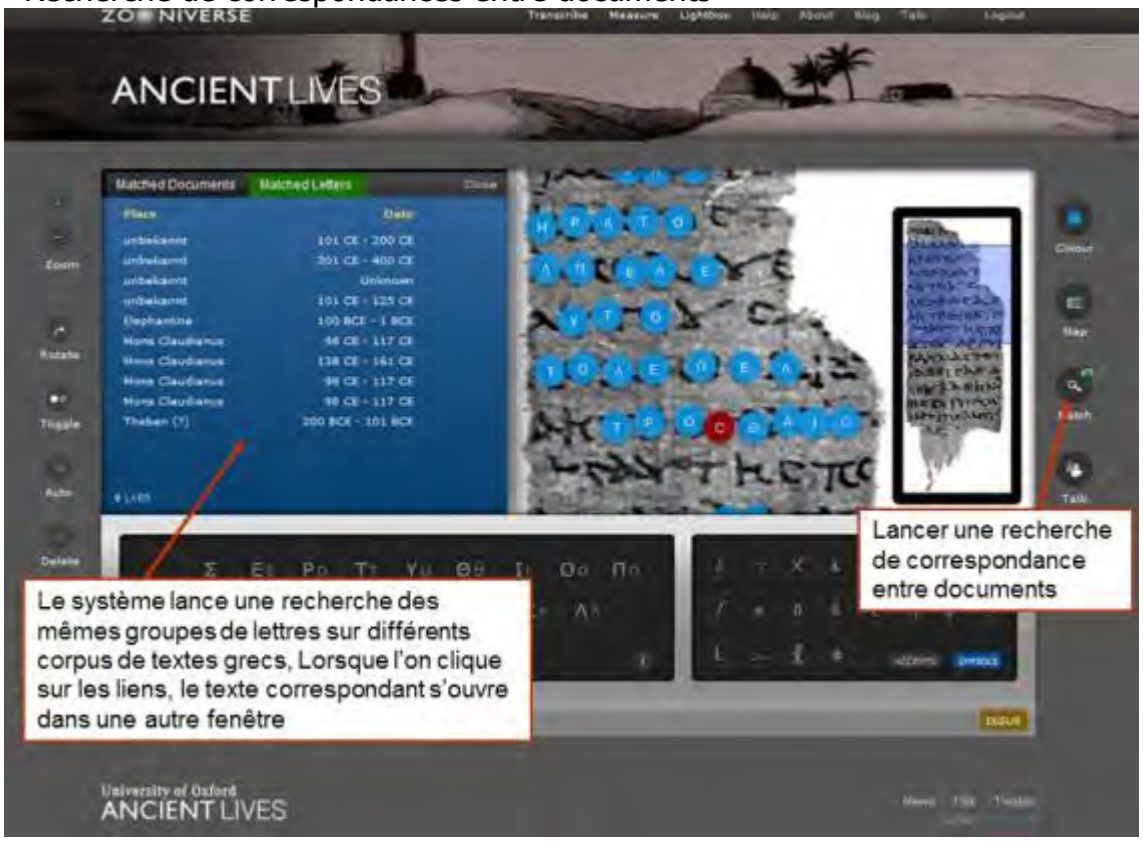
Facteurs de succès ou d'échec

La qualité des interfaces de transcription et de mesure constitue le principal atout de ce projet.
La communication, l'animation de la communauté des contributeurs dans l'espace social, ainsi que l'intégration dans le réseau plus vaste de Zooniverse, ont permis le recrutement de nombreux contributeurs, bien que le sujet puisse sembler aride et scientifiquement complexe.
Les interfaces sociales, particulièrement riches pour un projet de ce type, permettent de nombreuses interactions entre les correcteurs et les documents, entre les correcteurs entre eux, et avec les administrateurs, très présents. Toutefois, les fonctionnalités sociales, déportées sur un sous-site spécifique, souffrent d'un manque d'intégration dans l'interface de transcription (nécessité de s'identifier deux fois, mais aussi de sortir de la page de correction pour voir les commentaires) et d'un éclatement difficile à comprendre (discussions / commentaires / forums).
Il est également décevant de n'avoir aucune visibilité sur l'avancement du projet : ni consultation des textes transcrits (qui seront publiés ultérieurement, sur un tout autre site vraisemblablement), ni informations claires sur la quantité de fragments transcrits, mesurés, identifiés.

Copies d'écran

- Transcription d'un fragment

- Recherche de correspondances entre documents

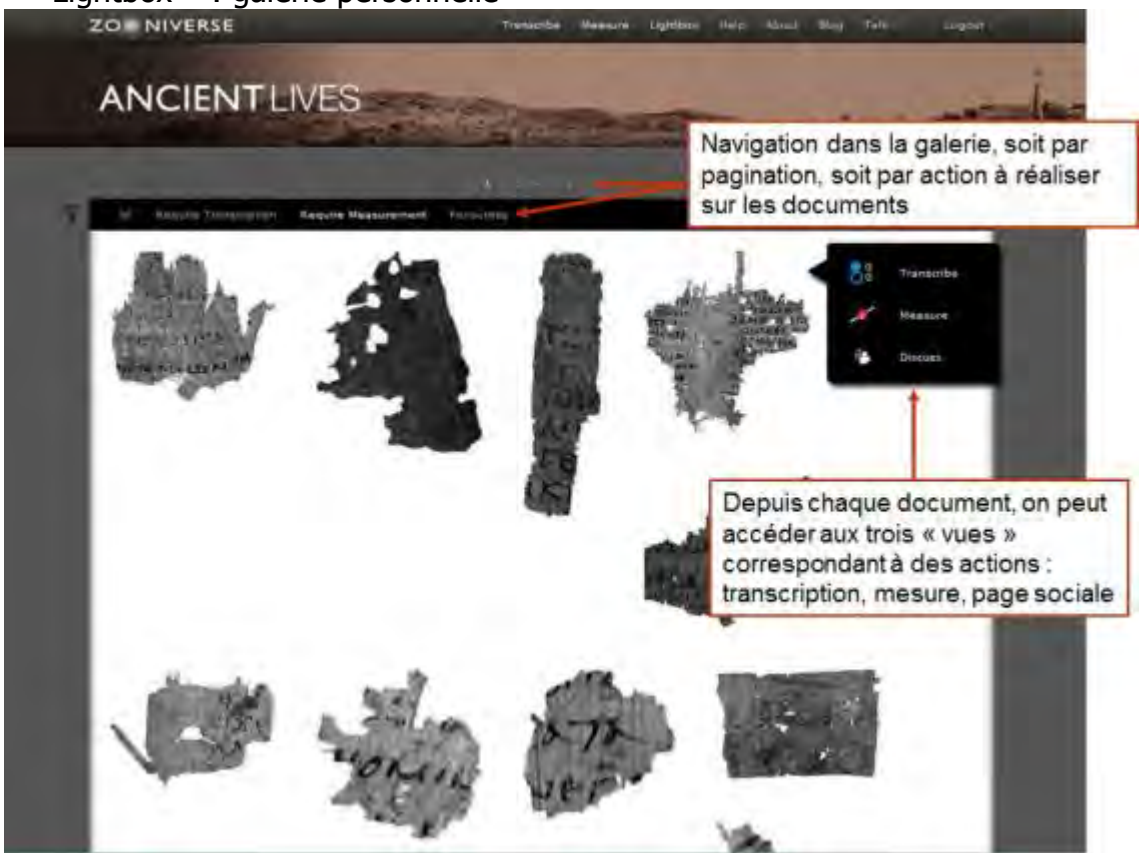


- Interface de mesure des marges

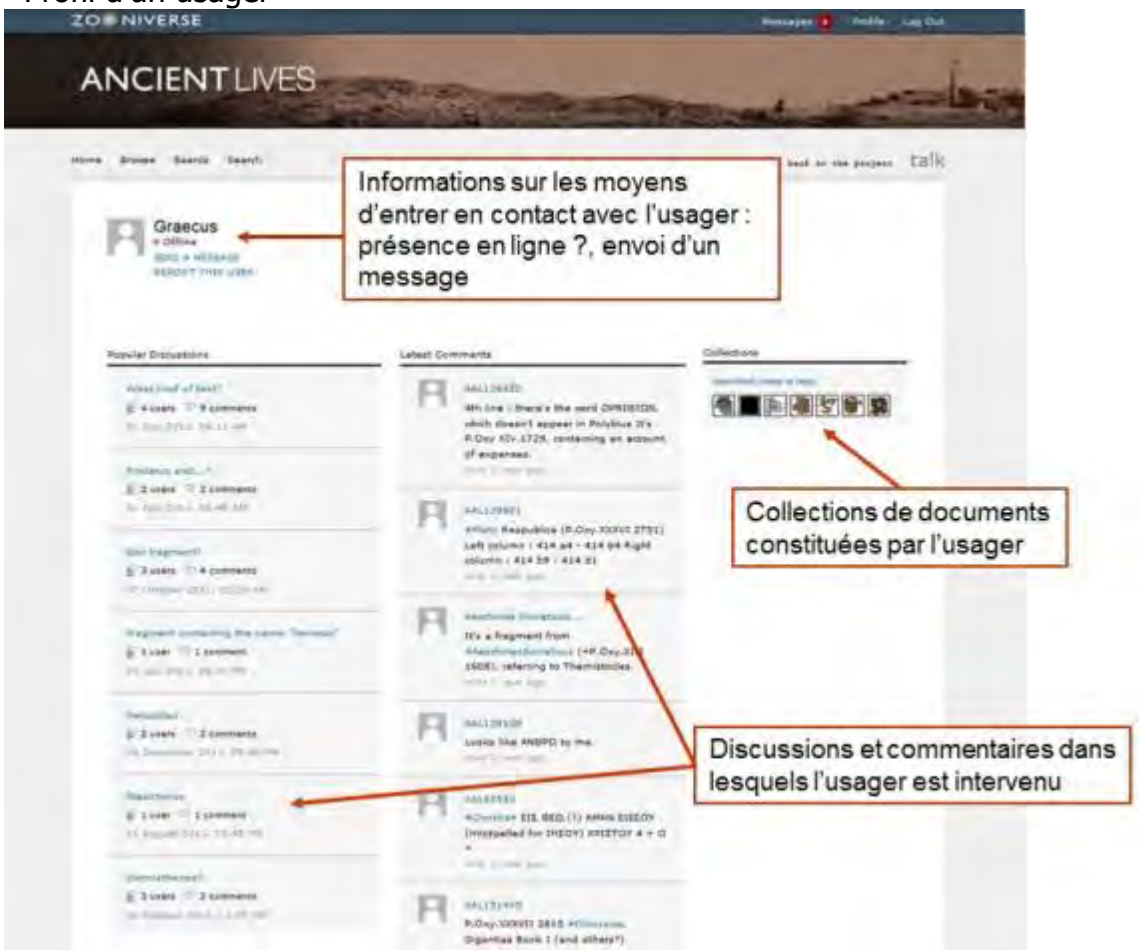
- Espace social

- Page sociale autour d'un document

- « Lightbox » : galerie personnelle



- Profil d'un usager



Bibliographie / webographie

Ancient Lives http://ancientlives.org

Plateforme sociale du projet Ancient Lives http://talk.ancientlives.org/

Blog du projet Ancient Lives http://blog.ancientlives.org/

Projet Old Weather http://www.oldweather.org

II.8. What's on the menu? *: transcription collaborative à la* New York Public Library *(Etats-Unis)*

Description du projet

Avec environ 45 000 menus de restaurants datant des années 1840 à nos jours, la collection de la *New York Public Library* est l'une des plus vastes au monde. Environ un quart des menus ont été numérisées et sont proposés en mode image. Afin d'améliorer l'accès à ces contenus, la bibliothèque s'est engagée dans un programme de transcription ouvert au public. La transcription a été préférée à l'OCR pour deux raisons :

- la nature des textes, qui sont en grande partie des manuscrits ;

- l'objectif de structuration des contenus, la granularité étant le plat, ce qui conduit à une transcription plus orientée données que texte : nom des restaurants et des plats, prix, localisation, etc.

L'objectif initial consistait à traiter environ 9 000 menus, ce qui a était accompli en trois mois. Depuis, de nouveaux menus numérisés sont régulièrement ajoutés pour être retranscrits (16 000 sont disponibles actuellement, soit environ 800 000 plats). Le but final est de transcrire la collection complète. Il est prévu de solliciter les utilisateurs pour d'autres tâches, telles la géolocalisation et la catégorisation des restaurants ou l'ajout de liens entre données. La bibliothèque réfléchit également aux moyens d'élargir l'étendue de sa collection grâce à des partenariats avec d'autres bibliothèques et services d'archives disposant de collections du même type.

La transcription est réalisée en deux temps :
- Pause d'une étiquette dans la vue image pour identifier un texte à saisir.
- Saisie du texte de l'étiquette.

Notons qu'à une saisie peut être associée une caractéristique d'incertitude (lisibilité).

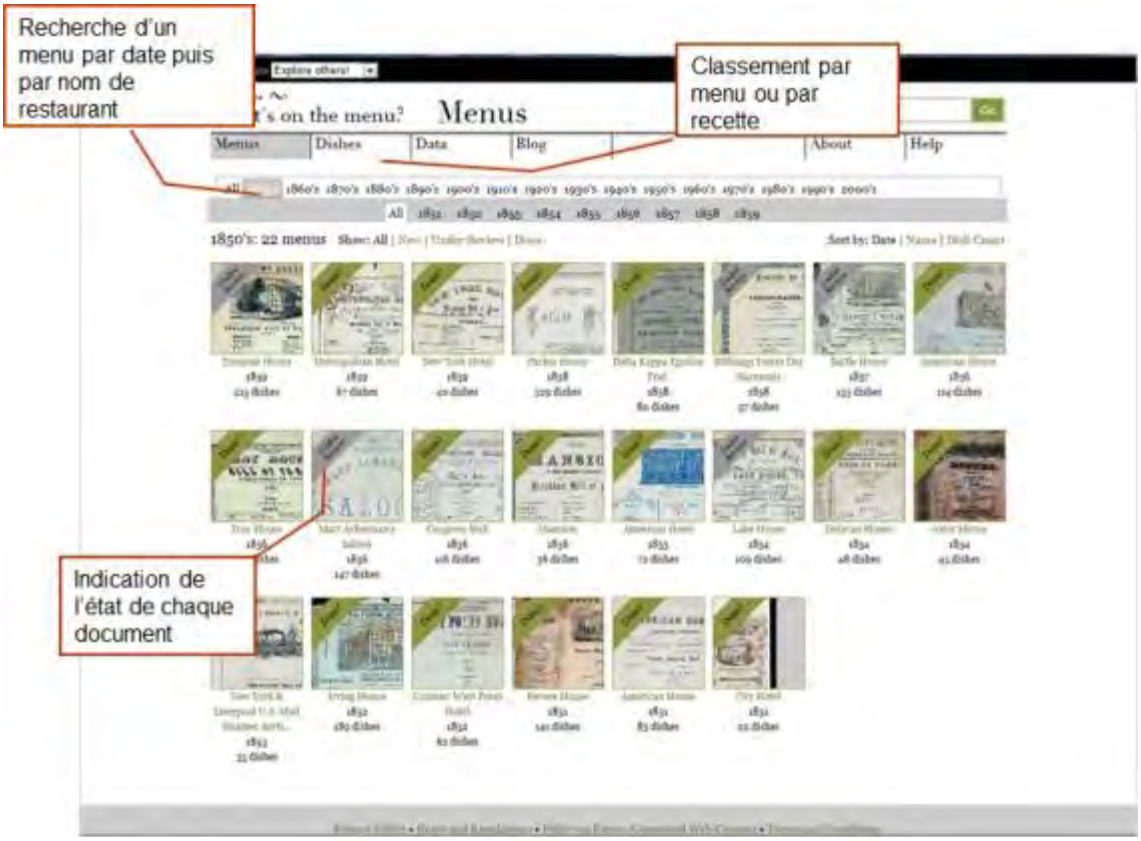
Facteurs de succès ou d'échec

+ La collection est monothématique mais le thème touche à la fois passionnés et grand public
+ L'interface de correction est simple et efficace
+ La transcription est réalisée sous la forme de micro-saisies : l'utilisateur peut transcrire ici et là, sans être attaché à un document particulier
+ Pas d'identification d'accès
- Pas d'OCR préalable, pas de segmentation : l'utilisateur a aussi la tâche d'identifier les textes à saisir par clic écran, et la liaison image-texte peut donc être imprécise.
- Pas de traces et statistiques
- Pas de fonctions sociales

Copies d'écran

- Accueil

- Consultation d'un document



- Correction d'un document

Bibliographie / webographie

Federman, Rebecca. *Happy Birthday to... Us! A Year of Menus*, avril 2012, [blog en ligne] http://www.nypl.org/blog/2012/04/20/happy-birthday-to-us-menus

II.9. Monasterium (ICARUS)

Description du projet

Monasterium est un portail numérique de sources diplomatiques de l'Europe médiévale porté par le réseau européen ICARUS (International Centre for Archival Research). Il donne accès à 250 000 documents originaux conservés dans plus de 50 institutions (services d'archives, universités, bibliothèques) situées dans 10 pays d'Europe. Il propose aux chercheurs médiévistes de participer à la transcription mais aussi à l'édition critique et scientifique des documents, sur la plateforme MOM-CA (Monasterium Collaborative Archive).
Cette plateforme fournit de très nombreuses fonctionnalités nécessaires à des travaux universitaires de recherche, comme le regroupement d'extraits de manuscrits en collections personnelles qui peuvent être publiées avec leurs propres descriptions.

Développé en 2002, MOM-CA propose aujourd'hui plus de 100 000 documents décrits selon des normes spécialisées (le standard XML CEI - *Charters Encoding Initiative*). De plus, les documents peuvent être progressivement transcrits et annotés (description matérielle, description du sceau, identification des personnes et lieux cités) grâce à l'éditeur EditMOM développé spécifiquement pour le projet.

Tout internaute intéressé par le projet peut contribuer, à condition de s'inscrire, bien que dans la réalité seuls des érudits y participent (150 inscrits aujourd'hui, historiens, étudiants, et quelques amateurs éclairés).
Afin d'assurer la qualité scientifique des contributions, chaque document est vérifié par des experts (une équipe de 14 personnes) avant d'être mis en ligne. Par ailleurs, un contrôle technique assure de la conformité des annotations avec le schéma XML CEI.

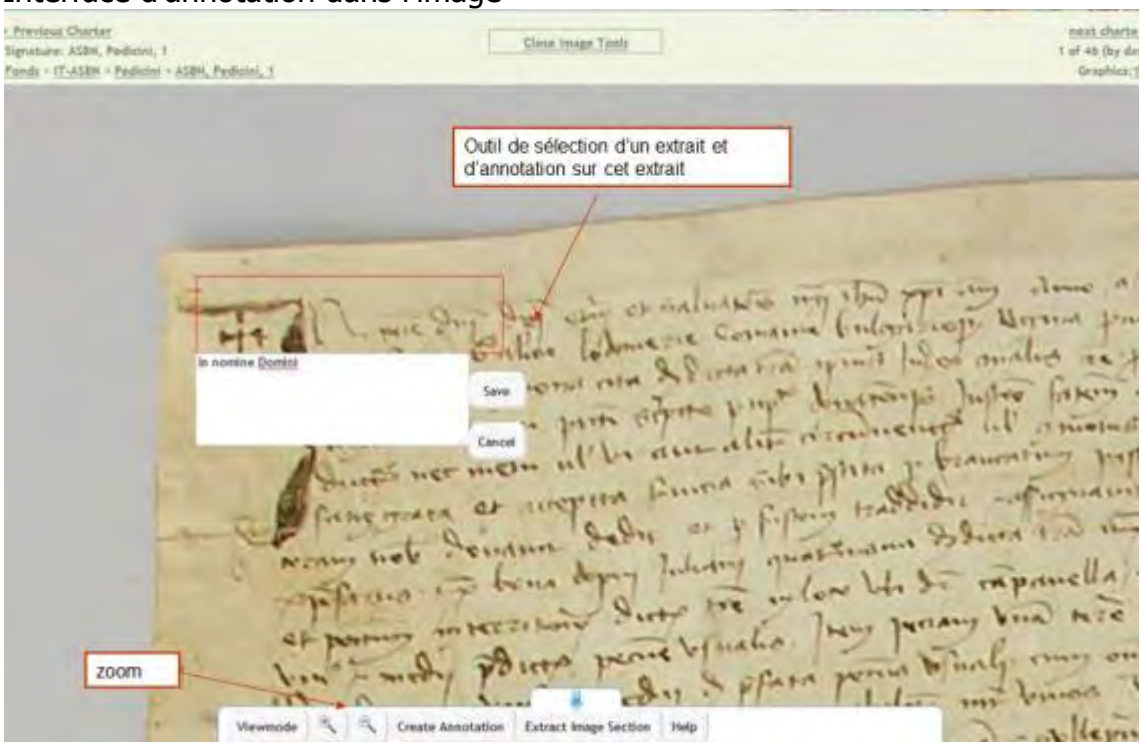
Facteurs de succès ou d'échec

- Une communauté ciblée pour laquelle la plateforme répond à un besoin professionnel : disposer d'outils techniques pour réaliser des travaux scientifiques dans un cadre professionnel, et bénéficier de fonctionnalités avancées de collaboration, personnalisation et diffusion des résultats de leurs recherches
- Une qualité exemplaire des résultats du crowdsourcing, grâce à une vérification systématique des corrections par des experts du sujet
- Une grande richesse des données produites, grâce à l'utilisation d'un standard XML spécifique (établissement de la transcription, structure, mais aussi édition critique et commentaires)
- En revanche, l'interface, bien que fournissant de très nombreuses fonctionnalités de correction, est complexe à prendre en main, et la plateforme de correction est complètement distincte du site de consultation des documents. Plusieurs interfaces différentes sont possibles à partir d'un même document, pour des usages de contribution différents, ce qui n'est pas toujours très intuitif : zoom dans la page / lecteur flash, correction sur la page / interface d'annotation dans l'image.

Copies d'écran

- Consultation d'une charte

- Interface d'annotation dans l'image



- Interface d'édition du texte

Bibliographie / webographie

Monasterium http://www.monasterium.net/

Plateforme MOM-CA (Monasterium Collaborative Archive) http://www.mom-ca.uni-koeln.de/mom/home

Wiki de développement du projet Monasterium http://www.mom-wiki.uni-koeln.de/

D. Jeller, *Presentation of the use of MOM-CA (Monasterium.Net- Collaborative Archive)*, 2011, http://fr.slideshare.net/icaruseu/bratislava-mom-capres13102011djeller.

G. Vogeler, *Lessons from Monasterium.net*, http://colab.mpdl.mpg.de/mediawiki/images/6/6f/Vogeler_Berlin120223.pdf.

II.10. ArcHIVE : transcription collaborative aux Archives nationales d'Australie

Description du projet

arcHIVE – Transcribe National Archives of Australia est un projet de transcription collaborative et de correction collaborative d'OCR (suivant les documents, dont certains sont manuscrits) des Archives nationales australiennes. Il vise à créer un catalogue d'archives, c'est-à-dire que la correction/transcription porte sur des listes de documents archivés, et non sur les documents eux-mêmes.
Ce projet présente quelques fonctionnalités intéressantes :

- mise en avant des *top contributors*, badges,
- récompenses (marque-page, poster, copie d'archive) quand on gagne assez de points,
- classement des documents par difficulté (3 niveaux)

Par contre, l'interface de transcription est perfectible, car peu ergonomique :

- les zones image et texte ne sont pas en regard
- pas de lien texte-image par clic
- pas de gestion des tableaux ou de structuration du texte

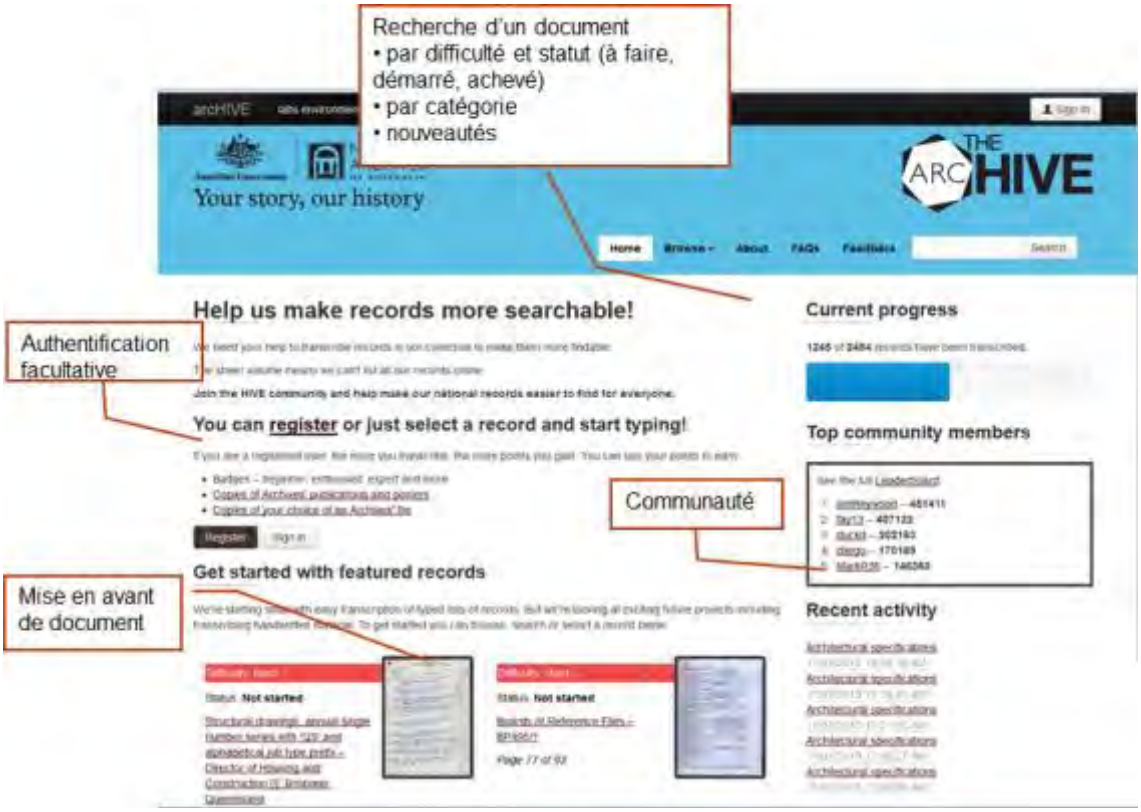
En termes de succès public, sur les 800 listes de documents d'archives disponibles au lancement, plus de 300 ont été complétées lors des deux premières semaines.

Facteurs de succès ou d'échec

+ Authentification facultative
+ Classement des documents par difficulté de correction
+ Compteur de score
+ Système de récompense
- Interface de correction peu pratique
- Aide en ligne minimaliste

Copies d'écran

- Accueil



- Recherche d’un document

- Correction / transcription du texte

Bibliographie / webographie

arcHIVE http://transcribe.naa.gov.au/

Holley, Rose, *National Archives of Australia embraces crowdsourcing and releases 'The Hive'*. http://rose-holley.blogspot.fr/2012/11/national-archives-of-australia-embraces.html

Help Transcribe the National Archives of Australia's Records with archive, http://www.gouldgenealogy.com/2013/01/help-transcribe-the-national-archives-of-australias-records-with-archive/

II.11. Do it Yourself History : transcription collaborative de l'Université de l'Iowa (Etats-Unis)

Description du projet

Ce projet vise à transcrire les collections spécialisées des bibliothèques de l'université de l'Iowa et des archives de l'université de l'Iowa. Les collections actuelles comprennent notamment des manuscrits culinaires, le journal intime d'une habitante de l'Iowa au XIX^{e}

siècle, des lettres et journaux de combattants de la Guerre de sécession américaine, des fonds photographiques, etc.
Les premiers travaux de transcription de manuscrits ont démarré au printemps 2011 et l'annotation de photographies en octobre 2012 (à l'aide de Flickr). Un module de correction d'OCR est prévu dans le futur. Les premiers résultats reportés sont de 15 000 pages transcrites par 1 000 participants.
La motivation des contributeurs a été particulièrement prise en compte :

- choix de corpus attractifs (histoire locale, livres de cuisine, contenus narratifs accrocheurs : l'usager a envie de connaître la suite de l'histoire…). Par ailleurs, le choix de corpus définis permet de fédérer des contributeurs qui ont les mêmes centres d'intérêts et sont plus susceptibles de créer de véritables communautés ;
- importance de l'animation de communauté (un animateur de communauté à temps plein), profil Twitter, projet de blog où les internautes pourraient publier des photos des recettes de cuisine qu'ils auraient testées

Par ailleurs, les transcripteurs des contenus historiques (Guerre de sécession) ont été confrontés à des difficultés de lecture spécifiques :

- acronymes techniques (militaires),
- vocabulaire et argot du XIXe
- lisibilité des manuscrits

Facteurs de succès ou d'échec

+ La collection est multithématique et touche à la fois passionnés et grand public…
+ L'interface de transcription est simple mais efficace
+ Identification d'accès optionnelle
+ Fonctions sociales : tweets, news
- Traces et statistiques seulement si identification
- Pas de structuration du texte saisi
- Pas de dictionnaire d'aide à la transcription

Copies d'écran

- Accueil

- Consultation d'un document

- Correction d'un document

Bibliographie / webographie

DIY History, http://diyhistory.lib.uiowa.edu/

Wolfe, Jen, *University of Iowa Libraries Offers Crowdsourcing History Research Opportunities* http://iowacity.patch.com/articles/university-of-iowa-libraries-offers-crowdsourcing-history-research

Jao, Carren, *DIY History crowdsources the transcription of 17th century cookbooks*, http://www.wired.co.uk/news/archive/2012-12/03/open-source-culinary-history

II.12. Tableaux de synthèse

NB : Les tableaux de synthèses ci-dessous sont proposés dans l'optique d'apporter une lecture transverse des projets étudiés au travers d'une grille de thématiques clés. Les commentaires exprimés restent subjectifs et les données chiffrées récoltées de doivent pas être abordées dans une perspective comparative.

Profil projet

	Organisme	**Objectifs du projet** (utilisation des données produites)	**Nature du corpus**	**Chiffres clés**	**Publics visés**	**Durée du projet**
TROVE	Bibliothèque Nationale d'Australie	Correction collaborative d'OCR automatiquement prises en compte en consultation des collections	Journaux australiens de 1803 à 1954 numérisés		Tous publics	Août 2008 à aujourd'hui
Wikisource	Wikimédia France / BnF	Correction collaborative d'OCR	Documents textuels présentant des qualités d'OCR variées	1416 documents dont 38 seulement ont été entièrement corrigés	Utilisateurs de la plateforme Wikisource Tous publics	2010 à aujourd'hui
CDNC	California Digital Newspaper Collection, University of California, Riverside.	Programme collaboratif pour l'amélioration de l'OCR des collections patrimoniales	Presse quotidienne de la Californie, ≈400 000 pages	450 correcteurs, 750 000 lignes corrigées	Tous publics	Fin 2011 à aujourd'hui
Digitalkoot	Bibliothèque nationale de Finlande avec Microtask	Programme collaboratif pour l'amélioration de l'OCR des collections patrimoniales	Documents anciens dégradés	8 millions de corrections de mots réalisées 110 000 participants	Tous publics	Févr. 2011 à novembre 2012
CONCERT	IBM dans le	Moteur de correction d'OCR	Textes anciens		Collaborateurs des	2009 à 2012

	Programme européen IMPACT	adaptatif			entreprises utilisant l'outil	
Transcribe Bentham	University College of London	Transcription brute des textes pour les chercheurs voire pour une réédition	Manuscrits inédits de Jeremy Bentham	60 000 manuscrits 7 548 articles et 15 426 pages corrigés 51 873 modifications enregistrées 2 663 utilisateurs dont 6 administrateurs	Etudiants, chercheurs mais aussi tout public	Déc. 2010 à aujourd'hui
Ancient Lives	Université d'Oxford	Transcription collaborative, mesure, identification des fragments	Fragments de papyri en grec	1,5 million de tâches de transcriptions entre juillet 2011 et décembre 2012	Tous publics Chercheurs spécialisés en papyrologie et littérature grecque	Juillet 2011 à aujourd'hui
What's on the menu	The New York Public Library	Transcription collaborative	Menus des restaurants de NY, ≈16 000 menus	9 000 menus en 3 mois	Tous publics	Avril 2011 à aujourd'hui
Monasterium	ICARUS (International Centre for Archival Research)	Transcription et édition scientifique collaborative	Manuscrits médiévaux	Plus de 100.000 documents mis en ligne 150 participants inscrits	Chercheurs et érudits médiévistes	2002 à aujourd'hui
ArcHIVE	Archives nationales australiennes	Transcription collaborative	Catalogues d'archive	1500 sur 3500	Tous publics	Avril 2011 à aujourd'hui
Do it Yourself History	Bibliothèques et archives de l'université de l'Iowa	Transcription et annotation collaborative (correction à venir)	Manuscrits culinaires, lettres et journaux intimes, fonds photographiques	30 000 pages transcrites par 1 000 participants	Tous publics	Printemps 2011 à aujourd'hui

Organisation générale

	Intégration dans le site référent (même univers, externalisé, fusionné)	**Authentification** (oui, non, facultative)	**Distribution de rôles** (animateur, expert, novice…)	**Sélection des documents** (par date, thème, difficulté, aléatoire…)	**Accueil éditorialisé** (présentation du projet, Mises en avant)	**Vérification qualité** (Validation, stat. de réponses)
TROVE	Intégration complète sur le site référent	facultative	non	Par date, par titre d'article, par état, par catégorie, par tags	Zoom sur un journal du jour + outils de sélection + tableau des scores	Historique de toutes les modifications effectuées afin qu'un administrateur puisse les annuler
Wikisource	Export complet sur le site spécialisé Wikisource	facultative	non	Liste alphabétique	Page de présentation du projet	Double correction Historique des corrections, page à page
CDNC	Intégration complète sur le site référent	oui	non	Par date, titre, mot-clé	Zoom sur un journal du jour + top 10 des correcteurs	*non documenté*
Digitalkoot	Site Digitalkoot. complètement externalisé	Authentification via Facebook	non	Mots présentés de façon aléatoire au joueur	Présentation projet + top 10 des joueurs	Vérification de l'efficacité des correcteurs par soumission de mots test + redondance des corrections
CONCERT	Outil indépendant	oui	Attribution des tâches (caractère, mot, page) en fonction des compétences des utilisateurs	Sur liste	Non	Contrôle de la qualité des correcteurs par insertion de pseudo erreurs
Transcribe Bentham	Site différent mais appartenant au même univers que le site du projet Bentham	oui	Classement allant de "stagiaire" à "prodige"	Par thématique, période, document non encore retranscrit ou partiellement retranscrit, niveau de difficulté ou aléatoire	Présentation du projet et de son avancée / liens vers les aides en ligne Top 10 des contributeurs Discussions en cours sur le	Validation effectué par les chercheurs hors outil de transcription

					Forum	
Ancient Lives	Site dédié, intégré dans le réseau Zooniverse	Facultative Obligatoire pour les fonctionnalités sociales	Non (mais interventions des administrateurs du projet)	Attribution aléatoire (pour l'usager, en fait priorisation des administrateurs)	Non	Non documenté
What's on the menu	Site dédié de transcription et de consultation	facultative	Non, mais tâches de transcription et de validation distinguées dans l'interface	Par date, par plat, par avancée	Présentation du projet et de son avancée Menus thématiques, plats du jour	Rôles de transcripteur et de validateur
Monasterium	Site dédié	obligatoire	Non (mais 14 experts administrateurs)	Par institution, par collection, ou via un moteur de recherche	Pas sur la plateforme de correction mais sur le site Monasterium (actualités, mise en avant de documents remarquables)	Une équipe de 14 experts valide chaque correction
ArcHIVE	Site dédié de transcription	facultative	non	Par nouveauté, avancée, difficulté	Présentation du projet et de son avancée Documents du jour Top 5 des contributeurs	*non documenté*
Do it Yourself History	Site dédié de transcription et d'annotation	facultative	Non, mais tâches de transcription et de validation distinguées dans l'interface	Par thème, par date, par avancée, par pages récemment éditées	Présentation du projet et de ses corpus thématiques	Rôles de transcripteur et de validateur

Prise en main de l'interface

	Interface facile, intuitive	**Double affichage (image /texte)**	**Approche scénarisée / gamification / multimédia**	**Tableau de bord utilisateur** (statistique, liste de documents, profil…)	**Accompagnement** (aide en ligne, tutoriaux, expert…)
TROVE	oui	oui	non	Statistiques, listes de documents, gestion du profil	Guide d'utilisation
Wikisource	Oui pour les corrections ponctuelles, moins pour les corrections de structure (saut de page, etc.)	oui	non	Profil personnalisable (commun aux autres projets Wikimédia), liste de suivi des documents auxquels on a contribué	Pages d'aide détaillée sur l'interface de contribution
CDNC	oui	oui	non	Statistiques, gestion du profil	Pas d'aide sur la correction
Digitalkoot	oui	oui	Deux petits jeux de validation et de saisie de mots	*non documenté*	*non documenté*
CONCERT	Interface encore à l'état de prototype mais assez intuitive	oui, ligne à ligne	non	non	Présentation d'écrans déjà corrigés
Transcribe Bentham	Interface user friendly, la tâche d'encodage est facilitée par des outils intuitifs	oui	non	Compte utilisateur : - Mes discussions - Mes préférences - Mes Favoris - Mes contributions	Guide d'utilisation et des vidéos de démonstration
Ancient Lives	oui	Oui (les caractères se superposent à l'image)	Interface multimédia	Galerie des documents auxquels on a contribué Profil regroupant les discussions	Tutoriel simple et interactif lors de la connexion sur le site Pages d'aide détaillées, FAQ

				de l'utilisateur	Conseils scientifiques et techniques complémentaires sur le blog Intervention des administrateurs sur l'espace social (forum, discussions autour des documents) pour guider les correcteurs
What's on the menu	oui	oui	non	non	Guide d'utilisation
Monasterium	Oui (interface d'annotation dans l'image), non (interface d'édition scientifique)	oui	non	Profil utilisateur Listes de suivi des chartes auxquelles on a participé ou que l'on a bookmarkées, des annotations que l'on a soumises à validation ou qui ont été validées	Pages d'aide très détaillées, en particulier sur le standard XML CEI utilisé
ArcHIVE	Non, peu ergonomique	oui	non		
Do it Yourself History	oui	oui	non	Statistiques, gestion du profil	Court guide d'utilisation

Outils de correction

	Correction dans le contexte du document	**Micro-tâche** (caractère/mot)	**Transcription de bloc de texte**	**Fonctionnalité de structuration**	**Mise en doute / signalement d'erreur**
TROVE	oui	non	oui	non	non
Wikisource	oui	Non	Oui	Faible (seulement séparation des pages)	non

CDNC	oui	non	oui	non	non
Digitalkoot	non	Micro tâche sous forme de deux jeux permettant de valider les résultats de l'OCR ou de saisir des mots	Non	non	non
CONCERT	Oui pour la vue « Page » et la vue « Caractère »	Micro tâche de vérification de formes similaires permettant de valider, rejeter ou corriger l'OCR au niveau caractère Micro tâche de validation de mots non fiables	non	Correction de segmentation des zones oubliées	Les mots non validés dans la vue « Mot » sont considérés comme non fiables dans la vue «Page »
Transcribe Bentham	oui	non	oui	Encodage du texte saisi pour signaler les exergues, ratures, passages à la ligne	Possibilité de coder le texte en « lecture discutable » ou « illisible »
Ancient Lives	Oui	Oui Correction caractère à caractère sous forme de microtâches, qui n'exclue pas un travail linéaire systématique	Oui Correction caractère à caractère, qui n'exclue pas un travail linéaire systématique	Oui (mesure des marges et de l'espacement des colonnes)	Pas au niveau du texte, mais possibilité de signaler un fragment comme non conforme au corpus (pas en grec, par exemple) Dans la page de discussion autour d'un document, possibilité de signaler un doute
What's on the menu	oui	oui	oui	Encodage de la position des textes dans la page	Mise en doute de la saisie
Monasterium	Oui	Non	Oui	Non	non
ArcHIVE					
Do it Yourself History	oui	non	oui	non	non

Fonctions sociales

	Forum	**Messagerie**	**Discussion instantanée**	**Tableau des scores**	**Système de récompense**
TROVE	oui	Messagerie liée au site global	non	oui	non
Wikisource	Page de discussion attachée à chaque document	non	non	non	non
CDNC	non	non	non	oui	non
Digitalkoot	non	non	non	oui	
CONCERT	non	non	non	non	Envisagé par le système
Transcribe Bentham	oui	oui	non	oui	
Ancient Lives	Forum Page de discussion attachée à chaque document	Possibilité d'envoyer des messages privés aux autres contributeurs	non	non	non
What's on the menu	non	non	non	non	non
Monasterium	non	non	Non	Non	Non
ArcHIVE	oui	non	non	oui	oui

Do it Yourself History	news	non	non	non	non

Communication projet / médiation

	Réseaux sociaux	**Blog**	**Campagne de communication** (relations presse, publicité, achat d'espace…)	**Autre type de médiation**
TROVE	Compte twitter	oui		Présence sur YouTube Publications scientifiques Participation à des séminaires, conférences…
Wikisource	non	non	Communiqués de presse BnF et Wikimédia France lors du lancement du projet. Rien depuis	Présentation du projet lors des Rencontres Wikimedia 2010. Rien depuis
CDNC	Partage de liens Facebook, lien Twitter	non	non	
Digitalkoot	Page Facebook	Microtask Blog	Nombreux articles de presse finlandaise et internationale (The New York Times, Wired…)	
CONCERT	Compte twitter du projet IMPACT vimeo Linkedin			Publications scientifiques
Transcribe Bentham	Page Facebook Compte twitter	Blog publiant régulièrement les progressions du site	Campagne de communication avec fortes retombées presse dont The New York Times	Organisation de rencontres présentielles (échec) Mailing diffusé auprès de la communauté

			Achat de Google AdWords (échec) Conception d'un dépliant présentant le projet et distribué lors de conférences	universitaire et professionnelle Participation à des séminaires, conférences
Ancient Lives	Compte Twitter, inactif	oui	Communication intense lors du lancement du projet (presse, radio)	non
What's on the menu	Page Facebook Compte twitter	oui	non	Export des données au format CSV, mise à disposition d'une API
Monasterium	Non	Non	Non	Présentation dans de nombreuses conférences scientifiques (monde de la recherche, monde des bibliothèques et des archives)
ArcHIVE	non	Blog des Archives australiennes	non	
Do it Yourself History	Compte twitter	non	non	

Résultats notables

	Points remarquables	**Les plus**	**Les moins**
TROVE	▪ Interface intuitive ▪ Organisation de la page d'accueil permettant de sélectionner un document par diverses options et mise en avant d'un journal paru à la date du jour	▪ L'application met à jour son contenu lorsque l'utilisateur l'a modifié ▪ Historique de toutes les modifications effectuées ▪ l'intérêt du sujet à traiter est un facteur de motivation ▪ Authentification facultative	▪ La zone à corriger n'est pas complétement identifiée au sein de l'image, seul le début de la ligne est indiqué
Wikisource	▪ Plateforme spécifiquement dédiée à ce type de projets : faible coût pour l'institution, recrutement potentiel des contributeurs habituels de la plateforme,	▪ Authentification facultative ▪ Excellente qualité des résultats grâce à la double correction	▪ Faiblesse de la communication et donc du recrutement ▪ interface peu intuitive,

	mais problèmes de réintégration dans la bibliothèque numérique d'origine		▪ format de correction en TXT non réintégrable dans le catalogue de Gallica
CDNC	▪ L'interface de correction est totalement intégrée à l'interface de consultation de la bibliothèque numérique (pas de plateforme ou d'outil externe)	▪ Accès au contenu au niveau article ▪ Format interne standard (METS/ALTO) ▪ Traces et statistiques pour les administrateurs, statistiques et classement pour les utilisateurs	▪ Peu ou pas de fonctions sociales
Digitalkoot	▪ Grosse couverture médiatique du fait de son choix de la gamification ▪ Projet couronné de succès	▪ Gamification ▪ Système de vérification de l'efficacité des correcteurs ▪ Mise en œuvre d'une large redondance des corrections pour obtenir un taux OCR de 99 %, malgré la difficulté (police Fraktur)	▪ Peu ou pas de fonctions sociales, mais envisagé pour le futur
CONCERT	▪ Décomposition et contextualisation des tâches de correction ▪ Le programme de recherche a été prolongé par la création du centre de compétence *SUCCEED* qui a pour vocation de valoriser les outils développés dans *IMPACT*	▪ Système capable d'apprendre de ses erreurs de reconnaissance ▪ Dictionnaire central adaptatif	▪ Interface restée à l'état de « prototype ». Le design et l'ergonomie n'étant pas encore optimisés ▪ Pas d'organisation globale de la collaboration
Transcribe Bentham	▪ Large communication et forte couverture médiatique mais toutes les stratégies n'ont pas eu les retombées espérées ▪ Le site a reçu de nombreuses visites qui n'ont pas toujours été transformées par une participation	▪ "*Benthamometer*" affiche les progrès de la transcription par Boite de manuscrits ▪ Multiple possibilité de sélectionner un document	▪ Difficulté de transcrire une écriture et une pensée complexe + ajout de la tâche d'encodage ▪ Fonctions sociales peu ou mal utilisées
Ancient Lives	▪ Intégration dans le réseau Zooniverse : soutien à la communication et au recrutement	▪ Qualité des interfaces ▪ Large communication ▪ Nombreuses fonctionnalités sociales permettant	▪ Séparation des interfaces de correction et des fonctionnalités sociales ▪ Aucune visibilité sur l'avancement du projet

	▪ Qualité des interfaces de correction, multimédia et ludiques	d'encadrer et d'aider les correcteurs	
What's on the menu	▪ Les données collectées sont rendues publiques (export Excel et API) ▪ Pas d'identification d'accès ▪ Corpus original, se prêtant bien à une éditorialisation	▪ Site bien designé, très éditorialisé ▪ Très bonne ergonomie de l'IHM de transcription	- Pas d'OCR préalable, pas de segmentation : l'utilisateur a aussi la tâche d'identifier les textes à saisir par clic écran, et la liaison image-texte peut donc être imprécise. - Pas de traces et statistiques, pas de fonctions sociales
Monasterium	▪ Outil spécialisé répondant aux besoins du public des chercheurs médiévistes, en termes d'outils d'annotation comme de consultation	▪ Communauté ciblée ▪ Excellente qualité des résultats grâce aux vérifications d'un groupe d'experts	▪ Interfaces multiples et complexes
ArcHIVE	▪ Classement des documents par difficulté de correction ▪ Présence d'un forum	▪ Système de récompense des contributeurs ▪ Compteur de score	- Interface de correction peu pratique et documents complexes à transcrire - Pas de structuration des documents - Aide en ligne minimaliste
Do it Yourself History	▪ Corpus multi-thématique ▪ Transcription, correction (à venir) et annotation dans le même projet ▪ Fonctions sociales : tweets, news	▪ Interface de transcription simple mais efficace ▪ Bonne éditorialisation	- Pas de structuration des documents - Pas de dictionnaire d'aide à la transcription (américain du XIXe, argot militaire)

III. Enjeux et pistes de réflexion

Les projets de *crowdsourcing* en bibliothèques et plus largement dans les établissements culturels sont une pratique encore jeune, dont les mises en œuvre restent largement innovantes et expérimentales. Les établissements qui se lancent dans l'aventure du *crowdsourcing* utilisent souvent leurs expériences comme un « bac à sable » permettant de tester de nouvelles formes d'interaction avec leurs usagers, et d'envisager ensuite, de façon progressive et itérative, la mise en place de projets plus ambitieux.
Les retours d'expérience des projets étudiés dans ce document permettent d'esquisser des pistes de réponse à la question : qu'est-ce qui fait qu'un projet de *crowdsourcing* est considéré comme « réussi » ?

III.1. Comment motiver les usagers à contribuer à un projet de crowdsourcing *?*

Afin de motiver les usagers à contribuer à un projet de *crowdsourcing*, et d'obtenir de leur part un investissement suffisant à remplir les objectifs de correction que l'institution s'est fixés, il est possible d'actionner des « leviers » de motivation à plusieurs niveaux :

- en amont, avant l'arrivée de l'usager sur la plateforme de contribution : comment faire connaître le projet, comment faire venir les contributeurs potentiels ?
- sur le site, à l'arrivée de l'usager : comment le convaincre de contribuer ?
- sur le site, après les premières contributions de l'usager : comment le convaincre de rester, de revenir, de devenir un contributeur régulier ?

« Recrutement » : comment faire connaître le projet, comment faire venir des contributeurs potentiels ?

Qu'est-ce qui incite les internautes à venir sur une application de correction ou de transcription ? L'équipe en charge du projet *Transcribe Bentham* a fortement investi dans la communication autour du projet, et a également communiqué sur leur retour d'expérience en ce domaine. Si la formule n'est pas forcément applicable en l'état à d'autres contextes (institutionnels, géographiques ou stratégiques), elle est toutefois intéressante à analyser pour mettre en avant la complémentarité des outils de communication et pour montrer qu'ils ne fonctionnent pas forcément tous :

- La campagne de presse a été un vrai succès. Les projets de *crowdsourcing* étant encore nouveaux et expérimentaux, ils peuvent encore être facilement relayés dans la presse.
- en revanche, les médias sociaux tels que Twitter ou Facebook, importants pour l'animation de la communauté, semblent avoir eu peu d'impact pour générer du trafic directement sur le site.

- de même, l'achat de Google Adwords a été un échec,
- et une communication ciblée vers des publics que l'on pouvait penser intéressés ou concernés (étudiants, chercheurs…) n'a pas toujours fait mouche.

Identifier les leviers de motivation des usagers en amont, avant leur arrivée du le site de *crowdsourcing*

L'un des enjeux est de mettre en place une communication claire vers des publics différents (chercheurs, étudiants, érudits locaux, usagers de la bibliothèque, simples curieux, public local, etc.). Il faut savoir présenter les enjeux et objectifs du projet en fonction des différentes sensibilités que l'on peut rencontrer. Les actions de communication doivent être précises et pertinentes et être adaptées aux différents types de motivation que peut susciter le projet chez ces différents publics[33], en particulier :

- L'intérêt scientifique / l'engouement pour le sujet abordé (Transcribe Bentham, DIY History, What's on the menu ?, Monasterium)
- La participation à une cause « citoyenne » / action de bénévolat (TROVE, les « sciences citoyennes » du réseau Zooniverse et en particulier Ancient Lives)
- La curiosité / l'intérêt des nouvelles technologies (Wikisource)
- L'envie de jouer (DigitalKoot)
- La volonté d'améliorer son e-réputation (ArcHIVE, Transcribe Bentham)
- Le sentiment de communauté (Transcribe Bentham, Monasterium)

Il conviendra ainsi de s'attacher à bien présenter le projet sur la page d'accueil et dans les différentes communications diffusées pour faire adhérer les participants aux enjeux et aux objectifs du projet (*What's on the menu, Transcribe Bentham*), pour « faire sens » aux yeux des contributeurs potentiels.

Pour simplifier le recrutement et susciter une motivation ciblée, il est possible de se limiter à un corpus spécifique et de définir une stratégie de communication vers un ou deux publics cibles (Monasterium, qui cible le public des chercheurs médiévistes, limite sa communication aux colloques et conférences scientifiques spécialisés).

Mettre en place les canaux de communication et de médiation adéquats

- Médias traditionnels (presse, radio) : communiqué de presse, pour atteindre un large public (DigitalKoot, Transcribe Bentham, Ancient Lives)
- Publications scientifiques ou participation à des conférences pour cibler des utilisateurs spécialistes du sujet ou du corpus (Transcribe Bentham, Monasterium)
- Mise en place de *teasers* ou export de *widgets* de jeu pour approcher les « *gamers* » (sur le site de l'institution et/ou sur ses outils de médiation sociale)
- Médias sociaux : blogs (Transcribe Bentham), réseaux sociaux (Twitter, Facebook), blogs et sites des communautés susceptibles d'être intéressées pour relayer l'information et capter tous types de publics (Wikisource, Monasterium), etc.
- Captation des usagers traditionnels de la bibliothèque via le site institutionnel (TROVE, CDNC)

Motivation / adhésion : comment convaincre l'usager de contribuer ?

[33] Voir aussi *Mais pourquoi contribue-t-on ?* http://donneesouvertes.info/2012/11/22/mais-pourquoi-contribue-t-on/ Données ouvertes, Le site du livre "L'open data, comprendre l'ouverture des données publiques" Simon Chignard, Fyp Editions, 2012.

Au-delà de la réussite du recrutement des bénévoles vers le site, il faut savoir le transformer en une participation effective. Le succès de Digitalkoot[34], service de correction collaborative d'OCR de la Bibliothèque nationale de Finlande (en un an, 101 614 visiteurs ont passé 328 376 minutes pour réaliser 6 461 659 micro-tâches de correction) provient ainsi autant de sa forte couverture médiatique en amont que de son approche ludique sur le site, qui finit de convaincre les internautes de s'investir.

S'adresser à tous les types de contributeurs ?

On constate souvent que même pour des projets qui accueillent de nombreux participants, la majorité des travaux (« jusqu'à 80% dans certains cas » nous dit Rose Holley[35]) est réalisée par 10% des utilisateurs.
Tim Causer et Valérie Wallace[36] du projet *Transcrire Bentham* reprennent la distinction développée par Caroline Haythornthwaite[37] entre « foule » et « communauté », qui repose sur deux modèles d'engagement différents dans les projets de *crowdsourcing* : l'engagement anonyme, simple et sporadique de la « foule » se différencie de l'engagement de la « communauté » qui va répondre à des tâches plus complexes et des lignes directrices détaillées. Les leviers pour motiver ces deux types de contributeurs ne sont pas les mêmes : alors que la foule va se satisfaire d'un retour purement quantitatif sous forme de statistiques, la communauté va attendre le soutien d'experts et une organisation basée sur des paliers de progression pour entretenir leur motivation.

Ces deux modèles sont complémentaires et les stratégies de recrutement et de motivation doivent être adaptées à chacun. Il s'agit de pouvoir faire appel non seulement à un large réseau d'anonymes qui vont produire des contributions ponctuelles et irrégulières basées sur des tâches simples, voire des micro-tâches, mais également de permettre la constitution d'un noyau dur de contributeurs engagés qui interagissent et s'entraident pour faire face à des tâches complexes.

Offrir aux contributeurs différents modes d'appropriation des documents

- Autoriser les corrections par des utilisateurs anonymes (La Bibliothèque nationale d'Australie en a fait le pari avec succès. Pour Rose Holley la clé de l'engagement est la confiance. *"If you give them a high level of responsibility, they repay that trust tenfold*"[38].)
- Favoriser l'appropriation du sujet par tous types d'utilisateurs grâce à des fonctionnalités de sélection des documents, multiplier les points d'accès aux documents à corriger pour que chacun s'y retrouve :
    - Recherche d'un document par date, auteur, titre (*Trove*, *Transcribe Bentham*) par zone géographique (*Trove)*, par tags ou mots clés (*Trove*, CDNC)
    - Sélection des documents selon leur niveau de difficulté (*Transcribe Bentham, ArcHIVE*)

34 Nora Daly, IMPACT Final Conference-Crowdsourcing in the Digitalkoot Project, 2011, http://impactocr.wordpress.com/2011/10/24/impact-final-conference-crowdsourcing-in-the-digitalkoot-project/
35 Holley, Rose, "Crowdsourcing: How and Why Should Libraries Do it?", dans : D-Lib Magazine. Vol. 16, n°s 3/4, 2010, [en ligne] http://dlib.org/dlib/march10/holley/03holley.html
36 Causer, Tim, and Wallace, Valerie. 'Building a volunteer community: results and findings from Transcribe Bentham', Digital Humanities Quarterly, vol. 6, no. 2, 2012.
37 Haythornthwaite, Caroline "Crowds and Communities: Light and Heavyweight Models of Peer Production", Proceedings of the 42nd Hawaiian Conference on System Sciences. Waikola, Hawaii, IEEE Computer Society (2009): 1-10
38 Conrad Walters Volunteers with an eagle-eye on the news The Age, February 7, 2011 http://www.smh.com.au/technology/technology-news/volunteers-with-an-eagleeye-on-the-news-20110206-1aifk.html

    - Choix des documents en fonction de leur état d'avancement (*ArcHIVE* ; les utilisateurs de *Transcribe Bentham* qui veulent démarrer sur une page vierge peuvent sélectionner les documents non encore retranscrits.)
- Donner la possibilité de ne pas faire de choix en proposant une sélection aléatoire des documents à corriger (*Transcribe Bentham, Digitalkoot*)
- Dégager des thématiques lisibles pour permettre la constitution de communautés. Certains thèmes ou supports semblent attirer plus de volontaires que d'autres. On pourra citer :
    - histoire, généalogie,
    - sciences
    - presse
    - archives ou documents manuscrits de personnes célèbres ou de périodes historiques marquantes (guerres, etc.)

<u>Proposer une interface ergonomique et fonctionnelle</u>

Les qualités ergonomiques et fonctionnelles du système de correction/transcription jouent un rôle majeur dans la réussite globale. Les utilisateurs doivent être accueillis par une ergonomie de saisie simple et intuitive qui facilite la prise en main et les travaux de correction ou transcription. Des outils de personnalisation facilitent également la prise en main par les contributeurs.

- Mettre en place un tableau de bord personnel pour que l'utilisateur s'approprie complétement l'interface et le projet :
    - Gestion de son profil (*Do it Yourself History, CDNC)*
    - Personnalisation de son profil *(Wikisource, Transcribe Bentham)*
    - Statistiques (*Monasterium, Trove…)*
    - Listes ou galeries de documents sélectionnés *(Ancient Lives, Transcribe Bentham)*
    - Favoris (*Transcribe Bentham, Monasterium)*
- Aider les nouveaux utilisateurs à appréhender l'outil, les conseiller et les guider
    - Présentation d'écrans déjà corrigés par d'autres (*CONCERT*)
    - Tutoriaux (*Trove*, *Transcribe Bentham, What's on the menu.?, Ancient Lives)*
    - FAQ, rubrique « Aide » (*Trove*)
    - Possibilité de demander de l'aide à un expert (Ancient Lives)
    - Bac à sable (*Wikisource*)
- Proposer une interface ludique pour « accrocher » l'utilisateur. L'enquête réalisée auprès des utilisateurs de *Transcribe Bentham* indique que de nombreuses personnes se sont passionnées par la tâche de transcription elle-même, elles se sont prises au jeu du défi de l'énigme. Sur des projets comme *Galaxy Zoo*[39] ou *Trove* les « super » contributeurs se décrivent comme « dépendant » ou « accro ».
    - Gamification (DigitalKoot)
    - Approche « scénarisée » (Old Weather)
    - Micro-tâches simples et addictives (CONCERT, Ancient Lives)

39 Gallaxy ZOO (www.galaxyzoo.org/) est un projet collaboratif d'astronomie de plusieurs universités internationales. Les membres du public sont invités à aider au classement des millions de galaxies à partir de photos numériques. Chris Lintott un membre de l'équipe de Galaxy Zoo dit: «"One advantage [of helping] is that you get to see parts of space that have never been seen before. These images were taken by a robotic telescope and processed automatically, so the odds are that when you log on, that first galaxy you see will be one that no human has seen before". Le but est d'avoir chaque galaxie classés par 30 utilisateurs différents. Cette classification multiple permet de construire une base de données précises et fiables, qui répondent aux normes élevées de la communauté scientifique.

Animation / cohésion : comment maintenir l'implication des usagers, les convaincre de devenir des contributeurs réguliers ?

Créer de la nouveauté, faire évoluer les contenus

L'équipe de Trove a constaté que la mise à disposition de nouveaux documents entraine toujours un pic d'activité sur le site.

De plus, le projet gagne en dynamisme si le contenu est éditorialisé :

- Ajout régulier de nouveaux contenus
- Editorialisation autour d'actualités liées au projet ou extérieures (*Do it yourself history, ArcHIVE*)
- Mise en avant de corpus et de documents sur la page d'accueil (journal de la date du jour sur *Trove* et CDNC ; menus thématiques et plats du jour sur *What's on the menu ? ;* mise en avant de documents remarquables sur *Monasterium ou ArcHIVE*, présentation des corpus thématiques sur *DIY History*)
- Présentation de l'état d'avancement du projet ("Benthamometer" de *Transcribe Bentham*, What's on the menu ?, ArcHIVE)

Créer du lien entre les contributeurs

Si les fonctionnalités sociales peuvent jouer un rôle de levier pour les fonctions de contribution, elles ne rencontrent pas toujours le succès escompté : l'enquête réalisée auprès des bénévoles du projet *Transcribe Bentham* montre qu'ils n'ont pas été intéressés par le fait de communiquer des informations personnelles via leur profil et se sont limité en majeure partie aux informations obligatoires.

- Fournir des lieux d'échanges et de discussion entre usagers (forums) (*Transcribe Bentham, Ancient Lives)*
- Mettre en avant les échanges entre usagers pour susciter davantage de réponses (Mise en avant sur la page d'accueil des Discussions en cours sur le Forum de *Transcribe Bentham* ou de *Ancient Lives*)
- Développer la socialisation par affinités révélées par le projet (travail sur la même page, sur le même thème, même auteur ou période) en affichant le nom des personnes qui ont corrigé un document (Trove affiche le nom des correcteurs et donne la possibilité de voir les corrections effectuées), ou en ouvrant des pages de discussion autour des documents (Wikisource, Ancient Lives)
- Offrir la possibilité de contacter directement d'autres participants ou de faire appel à un expert, par messagerie ou chat, pour sortir le contributeur de son isolement (*Ancient Lives*)

Créer du challenge entre les contributeurs

Plusieurs analyses de projets[40] montrent qu'il peut être motivant pour les utilisateurs de leur apporter une dose de challenge, en développant une saine concurrence.

- Proposer un Tableau des scores (TROVE, CDNC, DigitalKoot, Transcribe Bentham, ArcHIVE)
- Attribuer un qualificatif ou un rôle en fonction des interventions de l'usager sur l'application (classement allant de "stagiaire" à "prodige" sur *Transcribe Bentham*)
- Proposer un système de récompenses (*ArcHIVE)*

[40] Causer, Tim, and Wallace, Valerie. 'Building a volunteer community: results and findings from Transcribe Bentham', Digital Humanities Quarterly, vol. 6, no. 2, 2012. http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html
- Holley, Rose, "Crowdsourcing: How and Why Should Libraries Do it?", dans : D-Lib Magazine. Vol. 16, n°s 3/4, 2010, [en ligne] http://dlib.org/dlib/march10/holley/03holley.html

Valoriser le travail effectué

Au-delà de la mise en concurrence des contributeurs, il est important de savoir les remercier et de valoriser le travail effectué :

- Mettre en avant les contributeurs les plus actifs (« *top correctors* ») sur la page d'accueil (*Trove, Transcribe Bentham, Digitalkoot*). Le tableau des scores de *Digitalkoot* affiche même, en plus du nom des contributeurs, une évaluation quantitative de leur contribution (nombre de tâches réalisées, nombre d'heures passées)
- Proposer des statistiques personnelles dans le profil des utilisateurs.
- Réintégration dans le catalogue de la bibliothèque numérique des corrections effectuées, afin que les contributeurs puissent constater l'utilité réelle de leur participation (TROVE, CDNC)
- Présenter l'état d'avancement du projet, ce qui valorise collectivement les contributeurs

Conseils pour réussir un projet de *crowdsourcing* dans les bibliothèques ou les archives (par Rose Holley[41], manager de Trove) :

The Thing	The System	The People	The Content
Clear goal	Easy and Fun	Acknowledge	Interesting
Big challenge	Reliable and quick	Reward	New
Progress	Intuitive	Team support	Lots
Results	Options	Trust	History/Science

Le facteur « projet »

Le projet de correction/transcription collaborative doit pouvoir être facilement communiqué auprès de son public cible, tant au lancement public du projet (recrutement initial) que pendant son fonctionnement (information aux volontaires) :

- L'objectif doit être clair : quel(s) corpus, dans quel but ?
- L'objectif peut constituer un challenge
- La progression au regard de l'objectif doit pouvoir être mesurée et donc communiquée, tant au niveau du projet (« déjà fait », « reste à faire », etc.) que pour chacun de ses contributeurs (statistiques personnelles).

Le facteur « contenu »

La nature du contenu proposé à la correction/transcription joue à plusieurs niveaux :

- Un corpus homogène centré autour d'un thème unique pourra être plus facilement éditorialisé qu'un contenu hétérogène
- Un corpus associé à un thème, un territoire, etc., attirera des utilisateurs a priori intéressés, voire passionnés

41 Holley, Rose, "Crowdsourcing: How and Why Should Libraries Do it?", dans : D-Lib Magazine. Vol. 16, n°s 3/4, 2010, [en ligne] http://dlib.org/dlib/march10/holley/03holley.html

- Un corpus associé à un thème « pointu » aura un public potentiel moindre que celui d'un thème grand public

Le facteur humain

Un grand soin doit être apporté aux participants, puisqu'ils sont la force vive du projet. Les projets réussis réunissent en général ces critères :

- Un système d'identification permettant aux utilisateurs de capitaliser leur travail et d'obtenir de la reconnaissance de la part de la communauté des volontaires, sous différentes formes (classements ou récompenses). L'identification peut être optionnelle, pour ne pas effrayer les utilisateurs occasionnels.
- Introduire une (petite) dose de compétition entre utilisateurs, pour favoriser la participation globale
- Des outils de communication et de collaboration pour encourager la constitution d'une communauté d'usagers. Cette communauté peut être animée par les utilisateurs eux-mêmes en plus de la médiation des membres du projet
- Ecouter les utilisateurs, en favorisant par exemple les échanges avec les bibliothécaires
- Faire confiance aux utilisateurs. Si un système de détection des « mauvais correcteurs » est mis en œuvre, il doit être transparent pour les « bons » utilisateurs

Le facteur « système »

Les qualités principales attendues pour le système sont :

- La simplicité : le système doit être utilisable intuitivement, ou avec l'aide de quelques consignes. Il doit être rapide et fiable
- Le *fun* : le système doit être plaisant à utiliser et doit donner envie de l'utiliser à nouveau
- L'adaptabilité : le système doit pouvoir s'adapter aux choix et préférences des utilisateurs, en particulier :
    - Choix entre tâche de correction et tâche de validation,
    - Choix du niveau de difficulté de la tâche : documents faciles, moyennement difficiles, difficiles, etc.
    - Choix de la granularité de la tâche : mot, paragraphe, page. Par exemple, la correction d'une page de journal peut sembler effrayante au premier abord. Mais un découpage de la page en articles permet de présenter à l'utilisateur une petite portion de texte.

III.2. Quels sont les bénéfices d'un projet de **crowdsourcing** *pour l'institution culturelle ?*

Pour assurer un réel retour sur investissement pour l'institution culturelle, il est nécessaire que le *crowdsourcing* ne soit pas seulement un projet « décoratif » qui se contente d'améliorer l'image de marque de l'établissement en renforçant les interactions avec ses usagers, mais qu'il contribue réellement à enrichir les contenus numériques et à améliorer les services offerts aux usagers.

Pour cela, il convient de mettre en œuvre des processus pour assurer la qualité et la fiabilité des contributions produites, mais aussi de s'efforcer de réintégrer ces contributions au cœur du catalogue pour les offrir à la consultation.

Comment assurer la qualité des contributions ?

Formation des contributeurs

- Mise à disposition des contributeurs de guides d'utilisation, de pages d'aide, de tutoriels, de boîtes à outils
- Forums et espaces de discussion pour aider et conseiller les correcteurs (Ancient Lives)

Evaluation de la compétence des contributeurs

- Identification des contributeurs (enregistrement ou adresse IP)
- Recrutement et travail avec une communauté ciblée de spécialistes (Monasterium, Transcribe Bentham, Ancient Lives)
- Evaluation du niveau de compétence sur la base des corrections réalisées (Transcribe Bentham, CONCERT), par les administrateurs (choix des experts de Monasterium), par tests (Digitalkoot, ou, hors panel : les Archives départementales de l'Ain soumettent leurs indexeurs volontaires à un test de paléographie avant de leur attribuer des documents à indexer dont la difficulté sera adaptée à leur niveau)
- Distribution de rôles différenciés suivant le niveau de compétence (CONCERT, Transcribe Bentham)
- Classement des documents par difficulté (ArcHIVE)

Corrections multiples

- Soumettre, sans qu'ils le sachent, les mêmes corrections à plusieurs indexeurs pour croiser les résultats (Digitalkoot, projets du réseau Zooniverse et en particulier Ancient Lives)
- correction par un usager ensuite systématiquement vérifiée par un autre (Wikisource), le cas échéant en séparant des rôles de correcteur et de validateur (What's on the Menu ?, DIY History)
- Transparence sur les corrections effectuées (historique, versions) pour offrir une vérification collaborative des compétences des usagers et des modifications apportées (Wikisource) ou pour qu'un administrateur puisse les annuler (TROVE)

Signalement d'erreurs

- Proposer aux internautes de signaler les erreurs qu'ils constatent
- Proposer aux correcteurs de signaler les corrections dont ils ne sont pas sûrs (Transcribe Bentham, What's on the Menu ?)

Vérification par des professionnels ou des experts (Monasterium, Transcribe Bentham)

Pour compléter cette analyse, on peut voir la typologie des méthodes de contrôle de la qualité dans les projets de *crowdsourcing* identifiées par Ben W. Brumfield, développeur spécialisé dans les outils de transcription collaborative : « *Quality*

Control for Crowdsourced Transcription », *Collaborative Manuscript Transcription*, 2012, http://manuscripttranscription.blogspot.com/2012/03/quality-control-for-crowdsourced.html :

« ***Single-track methods*** » : le document ne fait l'objet que d'une seule transcription (par un seul contributeur ou de façon collaborative ensemble sur le même document)

- « Open-ended community revison » (Wikipédia) : les utilisateurs peuvent continuer à modifier le texte transcrit, sans limite dans le temps. Un historique des modifications permet de revenir à la version précédente et d'éviter le vandalisme.
- « Fixed-term community revision » (Transcribe Bentham) : convient pour des projets d'édition plus traditionnels, dont l'objectif est la publication d'une "version finale". Quand une transcription atteint un niveau acceptable, validée par les experts, elle est close et publiée.
- « community-controlled revision workflows » (Wikisource) : la transcription est considérée comme une "version finale" non plus par des experts, mais parce qu'elle a traversé un workflow collaboratif de correction/révision/validation
- « transcriptions with "known-bad" insertions before proofreading » : dans une première phase, les correcteurs sont invités à transcrire. Puis d'autres correcteurs révisent la transcription en la comparant au texte original ; pour s'assurer que la seconde lecture est bien réalisée, des erreurs sont ajoutées dans le texte : si toutes les « fausses erreurs » sont corrigées, le système déduit que les « vraies erreurs » ont dû être corrigées aussi.
- « single-keying with expert review » : lorsqu'une transcription a été réalisée par un contributeur, elle est validée ou rejetée par un expert (soit un professionnel de l'institution à l'origine du projet, soit un contributeur sélectionné). Si la correction est rejetée, elle est soit à nouveau soumise à correction, soit corrigée par l'expert et validée.

« ***Multi-track methods*** » : ces méthodes conviennent particulièrement à des corrections portant sur des données structurées ou des micro-tâches. La même image de départ est présentée à plusieurs contributeurs qui transcrivent chacun à partir de zéro. Généralement, les contributeurs ne savent pas s'ils sont les premiers correcteurs ou si d'autres transcriptions ont déjà été soumises. Puis les données ainsi collectées sont comparées automatiquement.

- « *triple-keying with voting* » (Old Weather, ReCAPTCHA) : l'image est présentée à 3 contributeurs, la majorité l'emporte (au depart, Old Weather proposait l'image à 10 contributeurs, mais ils se sont aperçus que la pertinence était sensiblement la même avec 3 qu'avec 10 contributeurs)
- « *double-keying with expert reconciliation* » : la même donnée est présentée à deux contributeurs, et, s'ils ne sont pas d'accord entre eux, un expert tranche.
- « *double-keying with emergent community-expert reconciliation* » (FamilySearch Indexing) : la method est presque similaire à la précédente, sauf que l'expert qui tranche entre deux corrections divergentes est lui-même un contributeur, qui a été promu conciliateur grâce à l'analyse automatique de ses contributions (volume, pertinence).
- « *double-keying with N-keyed run-off votes* » : si les deux contributeurs ne sont pas d'accord, la correction est re-proposée à un nouveau duo/trio d'usagers.

Comment réintégrer les contributions ?

Choisir l'endroit où se déroule l'activité de *crowdsourcing*

- sur une plateforme distincte, préexistante et spécialisée dans ce type d'activité, et gérée par un autre acteur, (Wikisource, mais aussi pourquoi pas Mecanichal Turk)
- sur une plateforme distincte, gérée par un autre acteur, mais spécialement développée pour le projet (Ancient Lives sur le réseau Zooniverse)
- sur une plateforme dédiée, développée par l'institution porteuse du projet (Monasterium, Digitalkoot, Transcribe Bentham, What's on the Menu ?, ArcHIVE, DIY History)
- complètement intégrée dans le catalogue de l'institution (TROVE, CDNC)

Intégrer les résultats du *crowdsourcing* à l'offre de services aux usagers

- Une plateforme distincte et préexistante sera bien entendu moins coûteuse à mettre en œuvre, et pourra bénéficier d'une plus grande visibilité, mais on n'aura pas la main sur les fonctionnalités ni les formats, ce qui pourra rendre complexe la réintégration des données produites (partenariat Wikisource / BnF). De plus, la multiplication des plates-formes augmente les risques de fragmentation, de perte de lisibilité pour l'offre globale de la bibliothèque, ce qui peut induire une perte de motivation pour les contributeurs.
- Dans tous les cas, il faut veiller aux liens avec le catalogue de l'institution : liens entrants (recrutement via le catalogue de contributeurs qui pourraient constater des erreurs et vouloir les corriger), et liens sortants (réintégration des résultats des corrections dans le catalogue)
- La réintégration ne doit pas être purement cosmétique, mais être véritablement intégrée dans le *workflow* d'alimentation du catalogue (indexation par le moteur de recherche, affichage dans les résultats), afin que les contributeurs constatent que leur participation a réellement servi : la mise en valeur du travail effectué est un facteur de valorisation des contributeurs.

Comment estimer la réussite d'un projet de crowdsourcing pour l'institution ?

Le recrutement et l'animation d'une masse critique de contributeurs, ainsi que la mise en place d'une réflexion sur les retours exacts du projet pour la bibliothèque sont donc bien évidemment des éléments clefs pour la réussite d'un projet. Mais celle-ci doit également être lue à l'aune des objectifs fixés initialement pour le projet (enrichir les descriptions ?, améliorer la qualité des contenus ?, développer de nouvelles formes d'interaction avec les usagers ?, améliorer la visibilité et la notoriété de la bibliothèque numérique ?, offrir de nouveaux services aux usagers ?) et du contexte institutionnel et géographique (une petite bibliothèque locale ne peut sans doute pas attendre le même nombre de contributeurs qu'une bibliothèque nationale ; un projet portant sur un corpus scientifique très limité n'a pas non plus besoin d'une grand volume de contributeurs, mais nécessite des compétences plus ciblées, etc.

Mesurer quantitativement la réussite d'un projet de correction collaborative implique ainsi de mettre en regard :

- le périmètre du projet, lié au public potentiel pour un corpus donné (national/local, grand public/érudit)
- le public qui a effectivement participé au projet
- le pourcentage de correction du corpus à la fin du projet

Ainsi, un projet de correction collaborative peut s'avérer être un succès malgré une faible participation (quelques centaines de volontaires) si le corpus est corrigé dans les délais prévus (Transcribe Bentham). En revanche, un projet d'ambition nationale sera un succès si une part significative de la population a participé (Digitalkoot).

Recommandations de l'OCLC pour un projet réussi de « métadonnées sociales[42] »

- établir des objectifs clairs pour le projet : interactions avec les communautés d'usagers ou enrichissement des collections ?
- motiver les usagers à contribuer et exploiter leur enthousiasme : identifier les raisons de contribuer (sujets intéressants, interfaces ludiques, contribution au bien commun, challenge, participation à une communauté)
- regarder d'autres projets pour glaner des idées avant de se lancer
- « Se lancer ! », ne pas se laisser intimider par la crainte des spam ou des malveillances : si les contributeurs sont bien encadrés, il y a peu de risques
- Mettre en place un règlement pour cadrer les interactions des usagers (aussi bien définition des comportements acceptables dans les commentaires par exemple, que définition des conditions légales de réutilisation des données produites par les usagers)
- Former et sensibiliser le personnel des institutions culturelles, à la fois en termes de technique et d'outils que de médiation avec les usagers
- Mettre en place des indicateurs pour mesurer le succès du projet (quantitatifs et qualitatifs)
- Analyser les atouts et les risques de déporter le projet sur un site tiers (par exemple Flickr)
- Utiliser de préférence des outils open-source
- Mettre en place des expérimentations avec les usagers avant et après le lancement du projet
- Ajouter régulièrement de nouveaux contenus, pour contribuer à maintenir l'intérêt de la communauté des contributeurs
- Rendre accessibles les données produites par les contributeurs, les indexer, les intégrer au catalogue
- Utiliser les projets de *crowdsourcing* pour construire de véritables communautés d'utilisateurs
- Utiliser des identifiants stables et pérennes pour désigner les objets numériques, afin de favoriser leur dissémination et leur visibilité
- Prévoir un plan de migration des contenus au cas où l'on voudrait changer de plate-forme
- Faire indexer les contenus par les moteurs de recherche (Google)
- Etre réactifs, répondre rapidement aux demandes des usagers

[42] *Social Metadata for Libraries, Archives, and Museums: Executive Summary*
http://www.oclc.org/content/dam/research/publications/library/2012/2012-02.pdf