Bibliothèques, *crowdsourcing*, métadonnées sociales

Pauline Moirez
Bibliothèque nationale de France
pauline.moirez@bnf.fr

Pauline Moirez, *archiviste-paléographe, a commencé sa carrière en tant que conservatrice du patrimoine (spécialité archives) aux Archives nationales (2003-2007) puis au Service interministériel des Archives de France (2007-2010), en charge de projets de diffusion numérique. Elle a rejoint en 2011 la Bibliothèque nationale de France où elle exerce une expertise sur les métadonnées (en particulier interopérabilité, web de données) et les usages numériques (bibliothèques numériques et interfaces des catalogues,* crowdsourcing*).*

L'intégration des bibliothèques dans l'écosystème du web permet d'envisager des possibilités inédites et innovantes d'interactions avec les usagers et d'enrichissement des métadonnées descriptives des collections en s'appuyant sur la participation des internautes.

Le terme le plus couramment utilisé pour désigner ce type de projets, qui peut s'appliquer largement au-delà du monde des bibliothèques et de la culture, est celui de *crowdsourcing*, c'est-à-dire des contenus ou informations (*source*) produits par la foule (*crowd*) des usagers. Ce terme générique met plus l'accent sur le volume des participants, sur la notoriété et l'ampleur des projets, sur la constitution de communautés de contributeurs, que sur la valeur de leurs contributions. On pourra ainsi désigner sous ce terme aussi bien des projets qui font appel à la sensibilité, à la subjectivité de l'utilisateur, comme la notation ou la critique d'ouvrages ou de films, que de véritables programmes scientifiques.

Les institutions culturelles, bibliothèques, archives, musées, s'attachant à la qualité et à la valeur ajoutée des contenus apportés par les utilisateurs, investissent particulièrement un sous-ensemble de ces activités de *crowdsourcing*, que l'on peut plus spécifiquement qualifier par l'adjectif de « participatives [1] », c'est-à-dire qu'elles sollicitent la mise en œuvre de véritables compétences et connaissances des usagers, de caractère scientifique, qui contribuent à l'enrichissement de la description des collections : indexation collaborative (en particulier, folksonomies [2]), identification de documents iconographiques ou audiovisuels, correction collaborative d'OCR, transcription collaborative, co-création de contenus scientifiques [3]. On parlera alors aussi de « métadonnées sociales [4] » pour insister davantage sur l'enrichissement et l'amélioration de la description des collections et donc de l'accès des utilisateurs à ces collections. Ce sont ces interactions de haut niveau qui feront l'objet principal de cet article.

Les projets de *crowdsourcing* en bibliothèque, et plus largement dans les établissements culturels, correspondent cependant à une pratique encore jeune, dont les mises en œuvre restent largement innovantes et expérimentales.

1. 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, « Exploring the participatory archives », dans *Archives Next*, billet publié le 30 août 2011. www.archivesnext.com/?p=2319

2. Olivier Le Deuff, « Folksonomies : les usagers indexent le web », *Bulletin des bibliothèques de France*, 2006, n° 4. En ligne : http://bbf.enssib.fr/consulter/bbf-2006-04-0066-002
– Olivier Ertzscheid, *Folksonomies et indexation sociale : le monde comme catalogue*, 2008. http://fr.slideshare.net/olivier/oe-abes-mai2008

3. Une typologie des différents types de projets de *crowdsourcing*, ainsi que de nombreux exemples analysés, sont disponibles dans : Pauline Moirez, Jean-Philippe Moreux et Isabelle Josse, *État de l'art en matière de* crowdsourcing *dans les bibliothèques numériques*, 2013. www.bnf.fr/documents/crowdsourcing_rapport.pdf

4. Voir l'étude de l'OCLC sur les métadonnées sociales dans les institutions culturelles : OCLC, « Sharing and Aggregating Social Metadata », 2011-2012. www.oclc.org/research/activities/aggregating.html

Par ailleurs, l'intégration de fonctionnalités d'enrichissement collaboratif dans les catalogues ou bibliothèques numériques françaises reste peu fréquente, et rencontre rarement le succès escompté [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.

C'est pourquoi il est intéressant de s'appuyer sur les retours d'expérience des institutions qui ont mis en œuvre de tels projets, pour essayer d'analyser les atouts et enjeux de ces projets pour les bibliothèques, ainsi que les risques et défis à relever, afin de donner des pistes de réalisation à nos bibliothèques... et l'envie de tenter l'aventure ?

Les enjeux des bibliothèques participatives : nouveaux usages, nouveaux besoins

Des données bibliographiques et des collections : autant d'atouts pour des projets de *crowdsourcing* en bibliothèque

Les bibliothèques disposent d'atouts significatifs pour mettre en œuvre de tels projets, et en premier lieu une conscience de l'importance des métadonnées – de leur exhaustivité comme de leur qualité – pour l'accès aux collections et – par voie de conséquence – pour l'amélioration des services aux usagers. Les bibliothèques ont également une bonne expérience de la récupération de données produites ailleurs (dérivation de notices bibliographiques, récupération de données en provenance des éditeurs, par exemple) : les métadonnées sociales peuvent aisément s'insérer dans ces processus de récupération et de rediffusion de données de provenances diverses.

La participation des usagers, qui peut s'appliquer à de simples données bibliographiques, présente un intérêt renforcé par la mise en ligne massive de documents numériques. En effet, la mise à disposition des usagers de documents numérisés, images, voire textes OCRisés, permet des opérations de *crowdsourcing* ambitieuses qui enrichissent notablement la description des documents : indexation, identification de photographies, correction d'OCR ou encore transcription collaborative. De plus, la masse, la richesse et la variété de ces collections numérisées par les bibliothèques (manuscrits, livres, images fixes, images animées, documents sonores ou audiovisuels) multiplient les opportunités d'expérimentations et d'interventions d'utilisateurs aux intérêts, formations et qualifications divers.

Enfin, pour les contenus édités, l'existence de plusieurs exemplaires d'un même document ouvre la voie à des réutilisations d'enrichissements sociaux réalisés sur d'autres exemplaires. C'est le modèle des médias sociaux spécialisés dans les échanges autour des livres, des films ou de la musique, comme Babelio, Sens Critique, LibFly ou encore LibraryThing, qui disposent d'un large vivier de contributeurs, et dont l'intense activité de recommandation peut permettre d'atteindre la masse critique de contributions nécessaire à l'enrichissement du signalement des bibliothèques [7].

Inventer de nouvelles interactions avec les usagers

Les bibliothèques s'inscrivent 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 [8], 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 [9].

L'enjeu des bibliothèques est donc, au-delà d'une réponse à cette attente des internautes, de faire le meilleur usage possible des contributions des usagers pour enrichir les métadonnées descriptives de leurs collections et pour améliorer l'expérience de recherche et de navigation des utilisateurs. Il serait en effet dommage de n'utiliser les potentialités du web social que de façon « cosmétique », sans en faire véritablement bénéficier le signalement des collections et l'interface de recherche de la bibliothèque.

Pour aller plus loin, le développement des « sciences citoyennes [10] », associant la participation d'amateurs

5. Lionel Dujol, « Le catalogue 2.0 ou le mythe de l'usager participatif ? », dans *La bibliothèque apprivoisée*, billet publié le 14 octobre 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 ? », dans *Carnet de notes*, billet publié le 11 février 2012. http://bccn.wordpress.com/2012/02/11/des-utilisateurs-aux-collaborateurs-une-fausse-bonneidee

6. Pauline Moirez, « Archives participatives », dans *Bibliothèques 2.0 à l'heure des médias sociaux*, dir. Muriel Amar et Véronique Mesguich, Éditions du Cercle de la librairie, 2012, p. 187-197.

7. Voir en particulier Eymeric Manzinali, « Babelthèque à la bibliothèque de Toulouse : observations sur les OPAC 2.0. », dans *Le Monde du livre*, billet publié le 29 février 2012. http://mondedulivre.hypotheses.org/477

8. La « pyramide de la participation », dite aussi 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 trois quarts de contributeurs au moins occasionnels au Royaume-Uni, par exemple). Voir Aref Jdey, « La règle des 90/9/1 est désormais dépassée », dans *Demain la veille*, billet publié le 2 juillet 2012. www.demainlaveille.fr/2012/07/02/la-regle-des-9091-est-desormais-depassee

9. Étienne Cavalié, « Les tags dans les Opac : ce n'est pas parce que personne ne s'en sert que ça ne sert à rien », dans *Bibliothèques [reloaded]*, billet publié le 19 février 2012. 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

10. NDLR : lire à ce sujet, dans ce dossier, l'article de Marc Pignal et Eva Pérez, « Numériser et promouvoir les collections d'histoire naturelle », p. 27-31.

Ancient Lives **: une interface de transcription attractive et intuitive (université d'Oxford).**

à des travaux de relevé, de dépouillement, d'identification scientifiques, peut s'appliquer aux domaines patrimoniaux[11], et en particulier aux collections de bibliothèques.

Au sein du réseau Zooniverse, maintenu par la *Citizen Science Alliance* dont l'objet est d'associer universités et musées dans des projets de sciences citoyennes, le projet *Ancient Lives* (voir illustration ci-dessus), coordonné par l'université d'Oxford, propose la transcription collaborative de centaines de milliers de fragments de papyrus de l'Égypte gréco-romaine, afin de les identifier, de les publier et de les mettre à disposition des chercheurs[12]. Entre 2011 et 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.

Répondre à l'évolution des besoins de recherche des usagers

Les métadonnées sociales permettent de répondre à des besoins différents et d'offrir aux usagers et aux chercheurs des services différents de ceux permis par les métadonnées produites par les catalogueurs professionnels. Métadonnées professionnelles et métadonnées sociales ne sont pas concurrentes, mais complémentaires, pour répondre à l'ensemble des besoins de recherche des internautes dans les collections des bibliothèques.

Ainsi, devant l'évolution des pratiques de recherche induites par les moteurs de type Google, il devient de plus en plus important de pouvoir fournir aux usagers une recherche en plein texte dans les collections textuelles. Des projets de correction d'OCR[13] ou de transcription collaborative *ex nihilo* ont vu le jour pour répondre à ce besoin, dont le plus abouti à ce jour est sans doute celui mis en place pour les périodiques dans la bibliothèque numérique Trove de la Bibliothèque nationale d'Australie[14]. Il propose aux internautes de participer à l'amélioration de la transcription de plus de 8 millions de pages ; 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.

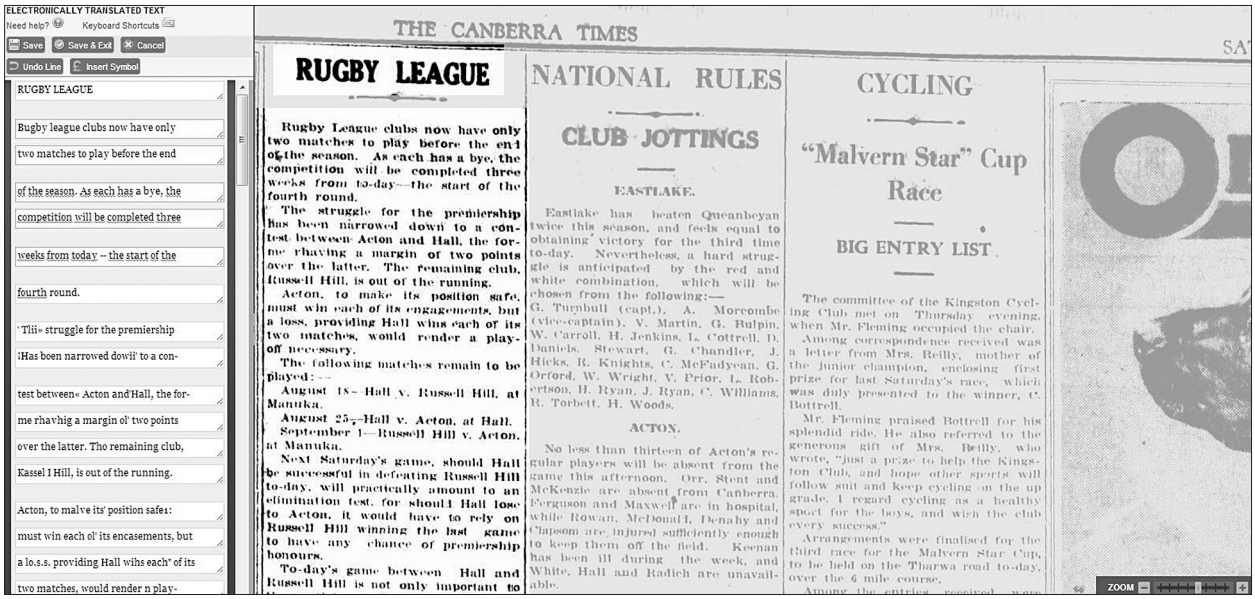
De même, des projets de *crowdsourcing* permettent d'offrir aux usagers une granularité de description des collections plus fine, et tout particulièrement pour ce qui concerne les collections iconographiques auxquelles il est impossible d'accéder par un moteur de recherche si elles ne disposent pas d'un minimum de données descriptives. La bibliothèque municipale de Lyon propose ainsi l'identifica-

11. Ainsi les Archives nationales des États-Unis ont mis en place une stratégie globale de projets participatifs regroupés sous le concept de « *citizen archivist* » (www.archives.gov/citizen-archivist) : indexation, transcription de documents, rédaction d'articles scientifiques, numérisation.

12. http://ancientlives.org

13. Voir à la fin de l'article l'encadré « La BnF engagée dans un projet de R&D pour la conception de la plateforme Correct (Correction et Enrichissement Collaboratifs de Textes) ».

14. http://trove.nla.gov.au
Voir Rose Holley, « Many Hands Make Light Work : Public Collaborative OCR Text Correction in Australian Historic Newspapers », 2009. www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf



Trove, une interface de correction au sein même de la bibliothèque numérique (Bibliothèque nationale d'Australie).

tion collaborative de photographies[15], tandis que la New York Public Library permet aux usagers de géoréférencer les cartes anciennes[16].

Bibliothèques au défi du *crowdsourcing*

Un projet de *crowdsourcing* nécessite généralement un important investissement en temps et/ou en argent. Afin d'assurer un réel retour sur investissement, il est nécessaire d'avoir conscience des risques de ces projets… et de relever leurs défis.

Recruter et motiver les contributeurs

Les projets de *crowdsourcing* n'ont de sens qu'à partir du moment où l'on atteint une masse critique de contributions qui permet de remplir les objectifs de correction que l'institution s'est fixés et d'améliorer véritablement l'accès aux collections.

Afin de recruter les usagers, de les faire venir sur l'interface de *crowdsourcing*, de les convaincre de participer, voire de devenir des contributeurs réguliers, il est donc nécessaire d'identifier les principaux leviers de contribution, et de les utiliser aussi bien dans les campagnes de communication et de médiation accompagnant le projet, qu'au sein de l'interface de contribution elle-même : intérêt scientifique, participation à une cause « citoyenne », envie de jouer, sentiment de communauté, etc.[17]

Ainsi, Digitalkoot, programme de correction collaborative d'OCR de la Bibliothèque nationale de Finlande, s'appuie sur la « gamification » pour engager les contributeurs à effectuer les tâches de correction[18] : deux jeux permettent de valider les résultats de l'OCR et de réaliser de la saisie de mots. Grâce à cette approche ludique, Digitalkoot a été un grand succès : près de 110 000 participants ont généré plus de 8 millions de tâches de correction de mots (voir illustration page suivante).

Le choix des corpus ouverts à la correction est également un levier de motivation fréquemment utilisé par les institutions culturelles. La New York Public Library cible par exemple les gourmands et gourmets grâce à son projet *What's on the menu ?* qui ouvre à la transcription collaborative 45 000 menus de restaurants datant des années 1840 à nos jours[19].

Assurer la qualité des données produites

La coexistence dans les catalogues et bibliothèques numériques de données produites par des professionnels et de données produites par les internautes nécessite de porter une grande

15. BM de Lyon, photographes en Rhône-Alpes. Portail de la BM : http://collections.bm-lyon.fr/photo-rhone-alpes
NDLR : voir la présentation qui en est faite dans l'article de Nicolas Gros et Pierre Guinard, « Numelyo, la bibliothèque numérique de Lyon », p. 12-14.
16. New York Public Library, Map Wraper. Portail de la NYPL : http://maps.nypl.org/warper

17. Voir aussi les « Tips for crowdsourcing » dans : Rose Holley, « Crowdsourcing : how and why should libraries do it ? », *D-Lib Magazine*, vol. 16, n° 3/4, 2010. http://dlib.org/dlib/march10/holley/03holley.html
18. National Library of Finland. Digitalkoot. www.digitalkoot.fi/index_en.html
– Voir : 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
19. NYPL. What's on the menu ? http://menus.nypl.org

vigilance à la qualité des données créées par ces derniers.

Il convient donc de mettre en œuvre des processus pour assurer la qualité et la fiabilité des contributions : formation et assistance des contributeurs, évaluation de leurs compétences et distribution de rôles différenciés, corrections multiples des mêmes données, vérification systématique ou échantillonnée par des professionnels [20].

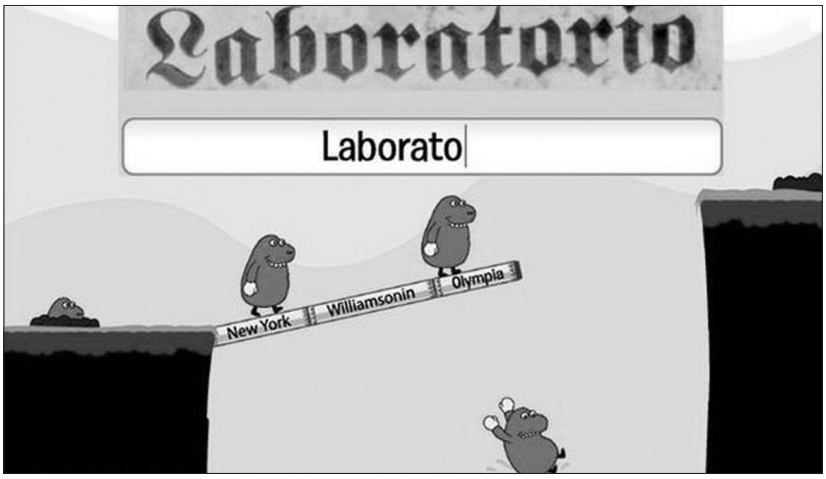
La qualité des contributions du projet *Transcribe Bentham*, transcription des 60 000 pages du philosophe anglais Jeremy Bentham, initié par l'University College of London, s'appuie ainsi sur une communauté soudée autour du projet, et sur une validation par des experts [21] : lorsqu'un manuscrit a été étudié par un nombre suffisant d'utilisateurs, il est soumis à la validation d'une équipe de chercheurs.

Dans le projet *Old Weather* [22], les internautes sont invités à transcrire les relevés météorologiques manuscrits réalisés par les navires de la Marine royale anglaise au début du xx^{e} siècle, afin de disposer de bases de données météorologiques complètes et fiables, pour comprendre et modéliser le climat et ses évolutions. Les relevés sont systématiquement soumis à deux contributeurs, et à un troisième en cas de différence entre les deux premiers.

Réintégrer les contributions dans les catalogues

Il faut également rester vigilants à éviter l'écueil d'un *crowdsourcing* « cosmétique », réalisé pour se conformer aux codes du web et pour donner l'image d'une institution innovante et à l'écoute de ses utilisateurs, 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, dans les bibliothèques numériques, 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 (comme dans la bibliothèque numérique Trove) ou déportés sur des médias externes (plateforme dédiée comme pour Digitalkoot, ou plateforme distincte, préexistante et spécialisée dans ce type d'activité, comme Wikisource [23]).



Digitalkoot, le jeu au service de la correction d'OCR (Bibliothèque nationale de Finlande).

Les nouveaux enjeux du *crowdsourcing*

Pour conclure, il convient de souligner que le *crowdsourcing* en bibliothèque, et plus largement dans les institutions culturelles, reste un domaine d'innovation, où des projets bien installés côtoient des projets de recherche aussi bien technologiques (interfaces hommes-machines, fouille automatique de données) que d'usages (des collectivités territoriales françaises envisagent par exemple de renforcer leur politique d'*open data* grâce à des projets de *crowdsourcing*).

L'articulation entre *crowdsourcing* et web de données semble ainsi particulièrement prometteuse, avec des projets comme HdA-Lab [24], collaboration entre l'Institut de recherche et d'innovation (IRI) et le ministère de la Culture et de la Communication, qui expérimente le « tagging sémantique » des ressources du portail Histoire des Arts en utilisant les entrées de Wikipédia comme référentiel d'indexation. ●

Septembre 2013

20. Voir aussi : Ben W. Brumfield, « Control for Crowdsourced Transcription », dans *Collaborative Manuscript Transcription*, billet publié le 5 mars 2012. http://manuscripttranscription.blogspot.com/2012/03/quality-control-forcrowdsourced.html

21. University of London, *The Bentham Project*. www.ucl.ac.uk/Bentham-Project/transcribe_bentham

22. Le projet *Old Weather* est intégré au réseau Zooniverse, sur des collections des Archives nationales du Royaume-Uni : www.oldweather.org

23. http://fr.wikisource.org/wiki/Wikisource:Accueil. La Bibliothèque nationale de France (http://fr.wikisource.org/wiki/Wikisource:Partenariats/Biblioth%C3%A&que_nationale_de_France) ou encore les Archives départementales des Alpes-Maritimes (http://fr.wikisource.org/wiki/Wikisource:Partenariats/Archives_D%C3%A9partementales_des_Alpes-Maritimes) ont mis en œuvre des partenariats avec Wikimédia France pour des projets de transcription collaborative sur Wikisource.

24. http://hdalab.iri-research.org/hdalab