Avant-propos

Au lieu d'externaliser certaines tâches auprès de prestataires ayant recours à des pays dont la main-d'œuvre est bon marché, les bibliothèques dans le monde font de plus en plus appel aux foules d'internautes, rendant plus collaborative leur relation avec les usagers. Après un chapitre conceptuel sur les conséquences de ce nouveau modèle économique sur la société et sur les bibliothèques, un panorama des projets est présenté dans les domaines de la *numérisation à la demande*, de la correction participative de l'OCR notamment sous la forme de jeux (*gamification*) et de la folksonomie. Ce panorama débouche sur un état de l'art du *crowdsourcing* appliqué à la numérisation et aux bibliothèques numériques, et sur des analyses dans le domaine des sciences de l'information et de la communication.

Remerciements

Imad Saleh, professeur au laboratoire Paragraphe de l'université Paris 8, pour avoir accepté d'encadrer mon projet de thèse, pour sa gentillesse et pour ses conseils tout du long du projet.

Samuel Szoniecky, maître de conférences à l'université Paris 8 – Saint Denis, pour avoir accepté d'être codirecteur de ma thèse et pour m'avoir invité à intervenir auprès de ses étudiants.

Ghislaine Chartron, professeure au Conservatoire national des arts et métiers, pour avoir accepté d'être rapporteur de ma thèse.

Stéphane Chaudiron, professeur à l'université Charles-de-Gaulle, Lille 3, pour avoir accepté d'être rapporteur de ma thèse.

Céline Paganelli, maître de conférences – HDR à l'université Paul Valéry Montpellier 3, pour avoir accepté d'être examinateur de ma thèse.

Alain Garnier, directeur général de Jamespot et référent crowdsourcing auprès du Groupement français des industries de l'information (GFII) pour avoir accepté d'être examinateur de ma thèse.

François Houllier, président de l'Institut national de la recherche agronomique, pour m'avoir permis de participer à ses côtés à un groupe de travail sur les sciences citoyennes afin de remettre un rapport sur le sujet à la demande de nos ministres de tutelles.

Odile Hologne, de la direction de la valorisation, information scientifique et technique de l'Institut national de la recherche agronomique, pour avoir encouragé les expérimentations autour du projet Numalire à l'Inra dans le cadre de mon travail.

Filippo Gropallo et Denis Maingreaud, de la société Orange et de la société Yabé, pour leur projet Numalire auquel ils m'ont permis de participer et pour leur collaboration tout au long de ce travail de recherche.

Marc Maisonneuve et Emmanuelle Asselin, de la société de *consulting* TOSCA, pour leur collaboration dans le livre que nous avons publié ensemble sur les logiciels et les plateformes pour développer des bibliothèques numériques.

Gaëtan Tröger, de l'Ecole nationale des ponts et chaussées, pour sa collaboration dans l'étude que nous avons menée sur la visibilité et les statistiques de consultation des bibliothèques numériques.

Pauline Rivière, de la bibliothèque Sainte-Geneviève, et Anaïs Dupuy-Olivier, de l'Académie de médecine, pour leur collaboration dans le retour de l'expérience Numalire que nous avons rédigé ensemble.

Robert Miller, d'Internet Archive, pour la collaboration que nous avons eue à la bibliothèque Sainte-Geneviève qui est devenue la première bibliothèque en France à participer à Internet Archive.

Stéphane Ipert, du Centre de conservation du livre, pour les collaborations et les intéressantes discussions que nous avons eues.

Pierre Beaudoin et Rémi Mathis, précédent et actuel présidents de Wikimédia France, association avec laquelle des collaborations avec Wikisource ont été concrétisées (Ecole

nationale vétérinaire de Toulouse en 2008) ou seulement envisagées (bibliothèque Sainte-Geneviève).

Valérie Chansigaud, historienne des sciences et contributrice Wikipédia, avec qui un premier contact avait été établi au Muséum puis une expérimentation pilote de numérisation et de correction participative de l'OCR avait été conduite dès 2008 à l'Ecole nationale vétérinaire de Toulouse.

Gilonne d'Origny, de la société ondemandbook.com, avec laquelle une collaboration pour une première implantation en France d'une Espresso Book Machine a bien failli se concrétiser.

Daniel Teeter, de la société Amazon, pour l'intéressante opportunité de partenariat que nous avons failli construire.

Juan Pirlot de Corbion, fondateur de Chapitre.com et de YouScribe, pour les passionnants échanges que nous avons eus au cours de nos rencontres.

Daniel Benoilid, fondateur de la société de crowdsourcing rémunéré Foule Factory, pour les discussions que nous avons eues.

Jean-Pierre Gerault, directeur général de la société I2S, leader dans le domaine de la fabrication de scanner pour la numérisation patrimoniale, président du Comité Richelieu et directeur général de Publishroom, pour les intéressantes discussions que nous avons eues.

Arnaud Beaufort, de la Bibliothèque nationale de France, rencontré à l'occasion de journées Wikimédia à l'Assemblée nationale et avec lequel j'ai eu un intéressant entretien par la suite.

Silvia Gstrein et Veronika Gründhammer, de l'université d'Innsbruck, pour m'avoir invité à intervenir à la conférence *Ebooks on Demand 2014*.

Yves Desrichard et Armelle de Boisse, de l'Ecole nationale supérieure des sciences de l'information et des bibliothèques, pour m'avoir permis d'intervenir aux journées « Quoi de neuf en bibliothèques ? » ces cinq dernières années.

Thierry Claerr, du ministère de la Culture et de la Communication, qui m'a permis d'intervenir régulièrement à l'ENSSIB, qui m'a sollicité pour la rédaction d'un ouvrage collectif et avec lequel j'ai eu des discussions très enrichissantes.

Jean-Marie Feurtet, de l'Agence bibliographique de l'enseignement supérieur, pour la collaboration que nous avons eue autour d'un projet de mutualisation d'une bibliothèque numérique et pour m'avoir invité à intervenir aux journées ABES 2011.

Nicolas Turenne, de l'Institut national de la recherche agronomique, pour m'avoir invité à exposer les premiers résultats de ses travaux au séminaire de l'axe « traces digitales » (groupe Cortext, Institut francilien recherche, innovation, société).

Pierre-Benoît Joly, directeur de l'Institut francilien recherche, innovation, société (IFRIS), pour m'avoir invité à donner un cours au master Etudes numériques et innovation (NUMI).

La SNCF, pour le confort des voyages en train pendant lesquels la thèse a été rédigée.

Google, pour le service Google Drive qui a été utilisé pour rédiger la thèse tout en donnant accès en temps réel à la rédaction du document à mon directeur de thèse, à mes collaborateurs et à mes contacts qui avaient ainsi la possibilité d'y ajouter des commentaires.

Mon épouse Véronique et mes trois enfants Terence, Orégane et Eloïse.

Je tiens également à remercier les personnes suivantes pour les commentaires constructifs qu'elles ont mis sur le texte de la thèse diffusée dans sa version première en Google Drive : Christine Young (une relecture d'article en anglais), Wilfrid Niobet (une idée, huit pistes, six corrections), Célya Gruson-Daniel (trois pistes, quatre corrections), Olivia Dejean (neuf corrections), Michaël Jeulin (sept corrections), Catherine Thiolon (dix pistes), Caroline Dandurand (cinq pistes), Diane Le Hénaff (trois pistes), Sophie Aubin (deux pistes), Nicolas Ricci (une piste), Pauline Rivière (une piste), Frédérique Bordignon (une piste), Sylvie Cocaud (une piste), Marjolaine Hamelin (une piste), Silvère Hanguehard (une piste), Christine Sireyjol (une piste), Odile Viseux (une piste), Véronique Decognet (une piste), Dominique Fournier (deux corrections), et tous les « soldats inconnus » qui sont restés anonymes dans leurs commentaires (quatre-vingt-deux corrections).

Introduction

Les bibliothèques ont déjà eu recours à l'externalisation de certaines tâches de saisies de notices bibliographiques, de catalogage, d'indexation ou encore de correction de l'OCR auprès de prestataires dans des pays où la main-d'œuvre est dite à bas coût. Cette externalisation est demeurée dans un cadre contractuel et limité et n'a pas bouleversé en profondeur le mode de fonctionnement sur lequel reposent les bibliothèques. Mais, avec le développement du *crowdsourcing*, il pourrait être envisagé d'externaliser (*outsourcing*) certaines de ces tâches, non plus auprès de prestataires, mais auprès de foules (*crowd*) d'internautes et donc de faire faire une partie du travail des professionnels par des amateurs. Le crowdsourcing modifie ainsi le paradigme sur lequel reposent des bibliothèques largement centrées sur la constitution et la conservation de collections. Il modifie également le rapport entre les producteurs d'un service que sont les bibliothécaires et ses consommateurs que sont les usagers, ces derniers devenant également des producteurs actifs du service. Le crowdsourcing pourrait aussi interroger les politiques documentaires des bibliothèques qui anticipent les besoins dans une logique d'offre qui n'est pas directement et immédiatement déterminée par la demande. C'est particulièrement le cas avec la numérisation à la demande par *crowdfunding*, une forme de crowdsourcing faisant appel, non pas au travail des foules, mais à leurs ressources financières ou avec l'impression à la demande qui lui est indissociable. Avec ces modèles économiques à la demande, la politique documentaire est finalement partagée avec les usagers qui décident de ce qui sera numérisé et/ou imprimé. Les collections deviennent ainsi l'œuvre des usagers.

Cet ouvrage a pour objet d'apporter des éléments de réponse à la question du recours au crowdsourcing à destination des professionnels des bibliothèques, mais aussi des étudiants, des chercheurs en sciences de l'information et de la communication et plus généralement les personnes intéressées par les projets d'intelligence collective. Il est le résultat d'une thèse en sciences de l'information et de la communication comportant à la fois une recherche-action, une expérimentation et une analyse de la littérature

[AND 16]. Cette thèse elle-même a fait précédemment l'objet d'un article en reprenant les principaux apports [AND 17].

Au-delà des questions coûts/bénéfices et avantages/inconvénients, la question d'une évolution du métier de bibliothécaire recentré sur ses compétences singulières sera abordée. Cet ouvrage a également pour objectif scientifique d'apporter une contribution à la connaissance du crowdsourcing sur le plan théorique et conceptuel autour des modèles économiques.

Cet ouvrage se limite à l'application du crowdsourcing dans le domaine de la numérisation et des bibliothèques numériques. Depuis les années 1990, la numérisation des documents s'est généralisée dans les bibliothèques. Aujourd'hui, avec la numérisation de masse et le développement de bibliothèques numériques géantes comme Google Books qui a dépassé le seuil des 30 millions de livres, ou encore Internet Archive, Hathi Trust, Europeana, le « moissonneur » de bibliothèques numériques européennes, il devient de plus en plus difficile d'identifier des imprimés n'ayant pas déjà été numérisés et méritant encore de l'être parmi les 130 millions[1] d'imprimés existants depuis l'invention de l'imprimerie.

Une partie non négligeable de ce qui est numérisé par les bibliothèques n'a jamais été mise en ligne. Elle génère des numérisations en doublons et « dort » sur des CD-ROM, des DVD ou des disques durs externes, dont la durée de vie est limitée. Le développement d'une bibliothèque numérique peut, en effet, être coûteux en logiciels, administration de logiciels et serveurs et le résultat peut être décevant en fonctionnalités, en pérennité, en coûts et en visibilité. En 2012, nous avions publié une étude consacrée aux logiciels YooLib (Polinum), Invenio (CERN), ORI-OAI (universités), DSpace (DuraSpace), DigiTool (Ex Libris), Mnesys (Naoned), ContentDM (OCLC), Eprint (université de Southampton), Greenstone (université de Waikato) et Omeka (université George Mason) [AND 12]. Dans cette étude nous constations qu'il était plutôt avantageux pour les bibliothèques de participer à une bibliothèque numérique mutualisée comme Internet Archive tant du point de vue des coûts (gratuité), des fonctionnalités (océrisation et conversion en EPUB et MOBI pour liseuses directement implémentées sur archive.org), de l'archivage pérenne (serveurs miroirs multiples dans le monde) que de celui de la visibilité. En effet, la position d'un site web dans la liste d'une requête Google dépend de son *PageRank*. Celui-ci dépend, en grande partie, du nombre de liens pointant vers son nom de domaine. Dans ses conditions, une bibliothèque numérique ayant beaucoup de contenu aura mécaniquement un meilleur PageRank et une meilleure visibilité

1. Le nombre de livres qui ont été imprimés depuis l'invention de l'imprimerie par Gutenberg est estimé à 129 864 880 par Leonid Taycher, un ingénieur de Google, d'après un article publié sur son blog le 5 août 2010.

sur le web et générera un trafic web plus important qu'une petite bibliothèque numérique avec peu de contenu.

Comme l'affirmait [WAI 08], il existe donc deux écoles, une vieille école qui considère que chaque bibliothèque doit créer sa propre bibliothèque numérique et chercher à y attirer des internautes et une nouvelle école qui considère plutôt, au-delà de la communication institutionnelle et pour mieux satisfaire les besoins des internautes, que les bibliothèques feraient mieux de participer à des bibliothèques numériques collectives déjà fréquentées par les internautes comme Internet Archive ou encore Flickr. C'est également notre point de vue. Fortes d'un trafic web suffisant, les bibliothèques pourront susciter la participation des internautes.

La partie introductive de l'ouvrage permet d'expliciter son contexte et la méthodologie qui a été utilisée.

Le premier chapitre conceptuel aborde les représentations philosophiques, politiques, économiques du crowdsourcing et ses conséquences sur le mode de fonctionnement des bibliothèques. Ce chapitre conceptuel contient, en particulier :

– une discussion critique à propos de la définition de crowdsourcing ;

– une chronologie originale de ses origines historiques ;

– une analyse au sujet de ses origines conceptuelles auprès de courants philosophiques parfois diamétralement opposés et, en particulier, un apport conceptuel autour de la loi de la valeur ;

– une réflexion sur le concept de sagesse des foules ;

– une analyse des diverses critiques du crowdsourcing appliqué aux bibliothèques numériques que certains pourraient qualifier, aujourd'hui, de « ubérisation » des bibliothèques numériques.

Le second chapitre contient une sélection de projets par types de tâches avec :

– la mise en ligne et la curation participatives ;

– la numérisation et l'impression à la demande sous forme de crowdfunding ;

– la correction participative de l'OCR et la transcription participative de manuscrits ;

– la folksonomie.

Ce chapitre contient des données et des informations récoltées dans la littérature pour chaque projet.

Des analyses originales pour chaque grand type de projets sont données en conclusion de ce second chapitre.

En troisième chapitre, des analyses du point de vue des sciences de l'information et de la communication et un état de l'art sont proposées avec, en particulier :

– une taxonomie originale du crowdsourcing en bibliothèque numérique distinguant crowdsourcing explicite (ou conscient) bénévole et rémunéré, crowdsourcing implicite (ou inconscient), gamification et crowdfunding ;

– une analyse des motivations des bibliothèques et des conditions nécessaires au développement de projets de crowdsourcing ;

– une taxonomie des motivations des internautes qui contribuent à leurs projets ;

– des analyses sur les récompenses et rémunérations éventuelles ;

– un éclairage à propos de la communication nécessaire au recrutement ;

– des développements sur le *community management* spécifique de ce type de projets ;

– des analyses sur la question de la qualité et de la réintégration des données produites ;

– une réflexion sur l'évaluation des projets de crowdsourcing.