Bibliothèques numériques et crowdsourcing : analyses bibliométriques et text mining

Mathieu Andro, Imad Saleh, Samuel Szoniecky

Bibliothèques numériques et crowdsourcing : analyses bibliométriques et text mining

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

(1) DIST, Institut National de la Recherche Agronomique. mathieu.andro@versailles.inra.fr

(2) Paragraphe, Université Paris 8

Résumé
Dans le cadre d'une thèse traitant du crowdsourcing appliqué aux bibliothèques numériques, un corpus spécifique exhaustif et représentatif du sujet a été constitué. Cet article présente les analyses bibliométriques et text mining de ce corpus.

Mots clés : crowdsourcing, bibliothèques numériques, numérisation, bibliométrie, text mining, fouille de textes

A l'instar d'un nombre croissant d'organisations publiques et privées, les bibliothèques peuvent désormais bénéficier des contributions de leurs usagers grâce au *crowdsourcing*. Dans les bibliothèques numériques, le *crowdsourcing* reste actuellement principalement limité à la correction participative de l'OCR (Andro, 2015, 2). Quelques expérimentations de numérisation à la demande par financements participatifs (*crowdfunding*) ont également déjà été conduites (Andro, 2014).

S'il est généralement bénévole, le *crowdsourcing* en bibliothèque fait parfois peut parfois aussi faire l'objet de récompenses ou de rémunérations via des plateformes. Il peut aussi être réalisé sous la forme de jeux. On parle dans ce cas de *gamification* (Andro, 2015, 1). Le travail peut aussi être réalisé de manière involontaire sur le modèle des *reCAPTCHA*. On peut parler dans ce cas de *crowdsourcing* implicite.

Une thèse de recherche-action a été conduite autour de ce sujet au sein du Laboratoire Paragraphe. Au cours de ce travail, un corpus au sujet du *crowdsourcing* appliqué aux bibliothèques numériques a été construit. Dans le présent article, nous présentons les résultats d'analyses bibliométriques et *text mining* de ce corpus original et représentatif du sujet.

Méthodologie

Un corpus a été initialement et classiquement constitué à partir de recherches bibliographiques sur le Web of Science, ScienceDirect et Google Scholar. Les équations de recherche utilisées ont été améliorées de manière itérative au fur et à mesure que de nouvelles expressions ont été identifiées et nous ont permis de développer notre vocabulaire se rapportant à la numérisation, aux bibliothèques numériques, aux diverses formes de crowdsourcing évoquées en introduction, aux noms de projets, de plateformes, d'auteurs ou de revues spécialisées identifiés notamment grâce aux analyses bibliométriques.

Le vocabulaire utilisé a été géré à partir de l'outil de text mining Luxid WebStudio. C'est un outil de construction de vocabulaires à partir de corpus de textes. Le vocabulaire que nous avons construit comporte les termes précédemment évoqués sous les formes anglaises et françaises avec leurs synonymes. Afin d'en permettre l'éventuelle réutilisation, ce vocabulaire a été déposé en tant que donnée de la recherche au format XML SKOS sur la plateforme Zenodo (Andro, 2016, 1).

Le corpus a également été complété de manière plus classique à partir des bibliographies des publications que nous avons lues. Afin de ne pas avoir à répéter périodiquement nos équations de recherches sur nos sources au risque de faire des erreurs, de repérer plusieurs fois les mêmes publications ou d'en ignorer, nous avons également mis en place un dispositif de veille éditoriale avec l'outil DigiMind. Ce dispositif nous a permis d'enregistrer et de capitaliser les sources et le vocabulaire d'interrogation de ces sources et d'être alerté rapidement de toute nouvelle publication relative au sujet.

Les notices bibliographiques obtenues via ces divers moyens de collecte ont été importées sous EndNote, dédoublonnées grâce aux fonctionnalités de l'outil, nettoyées manuellement, retravaillées avec Open Refine et analysées statistiquement avec Excel afin de produire les analyses bibliométriques.

Afin de produire les analyses, deux corpus ont été utilisés : un corpus de 219 références de la bibliographie de la thèse afin de produire les analyses bibliométriques déjà évoqués et un corpus de 1071 textes intégraux afin de produire les analyses text mining. Ce corpus contient les textes des 219 références de la bibliographie de la thèse mais également des publications sur le sujet qui ont été lues mais n'ont pas été retenues en bibliographie. Les analyses text mining ont été produites grâce aux outils Luxid. Cette technologie nous a permis de produire des analyses complémentaires à celles obtenues avec la bibliométrie à partir des champs structurés des bases bibliographiques.

Ces corpus constitués entre décembre 2013 et juin 2015 résultent d'une sélection par nos soins après lecture des publications. Ils diffèrent donc qualitativement des corpus généralement construits afin de réaliser des analyses bibliométriques et qui, n'ayant pas toujours fait l'objet d'une sélection manuelle, peuvent comporter du bruit ou du silence documentaires et qui ont pour effet de biaiser les résultats obtenus. Nous estimons que la méthodologie que nous venons de décrire nous a permis d'obtenir un corpus exhaustif mais aussi un corpus fiable et sans bruit, chaque publication ayant été lue. Dans ces conditions, nous estimons que les analyses produites ci-après sont bien représentatives du crowdsourcing en bibliothèque.

Résultats

Analyses bibliométriques

Quelle est la dynamique de ce sujet de recherche ?

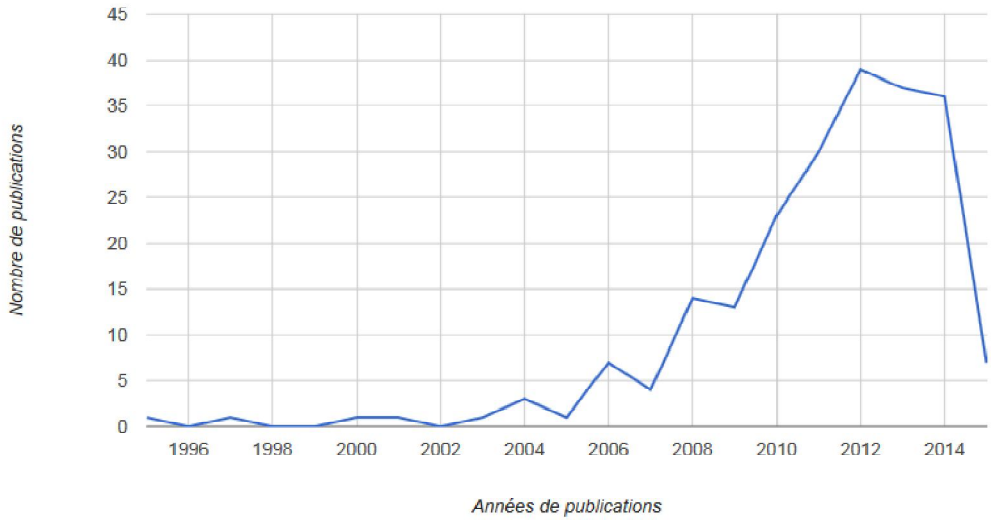

Figure 1. Nombre de publications par an dans la bibliographie de la thèse

Le sujet du crowdsourcing appliqué aux bibliothèques numériques est un sujet de recherche croissant, principalement à partir de 2006. Néanmoins, on observe une stagnation depuis 2013 qui coïncide probablement avec le passage d'une phase expérimentale à une phase plus industrielle.

Qui travaille sur ce sujet ?

Parmi les 361 auteurs du corpus, voici les plus spécialisés sur le sujet :

Holley R. 9 occurrences	Deterding, S 3
McKinley, D. 7	Dijkshoorn, C 3
Von Ahn, L 7	Gstrein, S 3
Aroyo, L 4	Ipeirotis, P. G 3
Smith-Yoshimura, K 4	Mühlberger, G 3
Alam, S. L 3	Tonra, J 3
Blum, M. 3	Wallace, V. 3
Campbell, J. 3	

Parmi les 149 institutions d'auteurs de notre bibliographie de thèse, voici les plus représentées :

Australie - National Library of Australia – 11 occurrences
Nouvelle Zélande - Victoria University of Wellington - 9

USA - Carnegie Mellon University - 8

UK - University College of London - 6

Danemark - University of Copenhagen - 5

Autriche - Innsbruck University - 4

Pays Bas - University Amsterdam - 4

USA - OCLC - 4

USA - University of Illinois – 4

USA - University of Iowa - 4

Parmi les 31 pays de ces auteurs, on obtient la répartition suivante sur une carte :

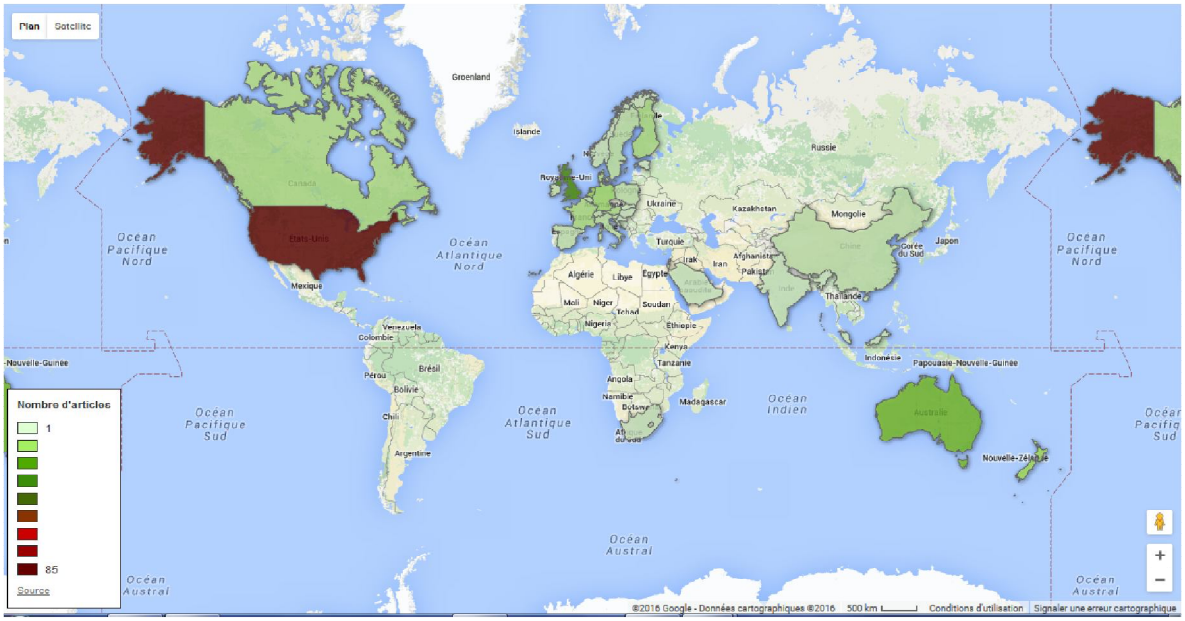

Figure 2. Poids des pays des auteurs dans la bibliographie de la thèse

On observe une nette domination des pays anglo-saxons, Etats-Unis puis Royaume Uni en tête. On observe aussi l'absence des pays en voie de développement et des pays du Sud. Concernant la France, nous n'avons considéré que les textes en langue internationale afin de ne pas biaiser les résultats.

Analyses text mining

Quels types de *crowdsourcing* sont évoqués ?

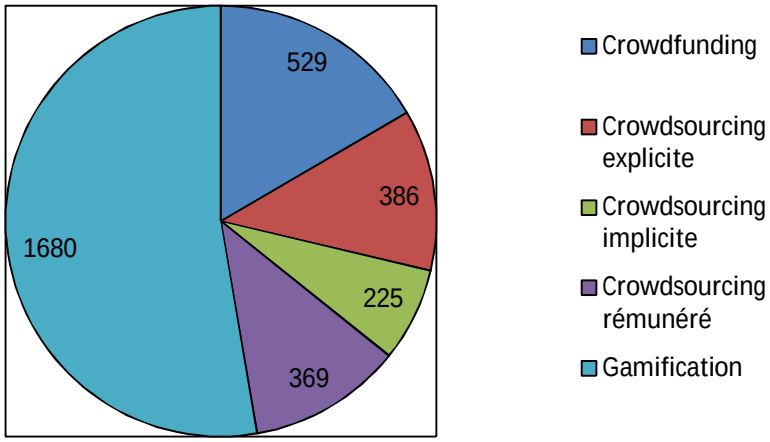

Figure 3. Nombre d'occurrences par type de crowdsourcing dans le corpus text mining

La gamification est un domaine très représenté dans la littérature scientifique. Il semble intéresser la recherche scientifique bien que le nombre de projets en bibliothèque reste encore assez faible par rapport aux projets classiques de *crowdsourcing*. Le *crowdsourcing* explicite reste, en effet, très majoritaire mais semble moins intéresser le monde académique du fait de son caractère désormais moins innovant. La même observation peut être faite aussi pour le *crowdfunding*, le *crowdsourcing* rémunéré ou le *crowdsourcing* implicite qui sont encore peu expérimentés en bibliothèques.

Quels types de motivations sont évoqués ?

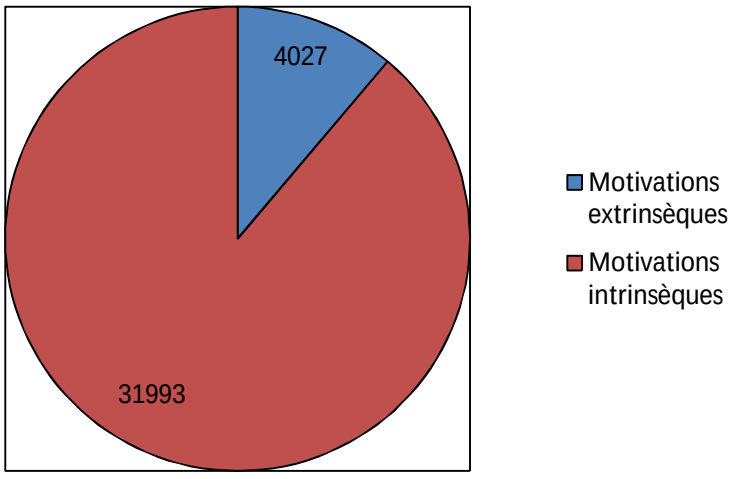

Figure 4. Nombre d'occurrences relatives pour les motivations extrinsèques et intrinsèques dans le corpus text mining

Nous définissons les motivations intrinsèques dans le domaine du *crowdsourcing*, comme des motivations qui poussent les internautes à contribuer pour le plaisir et le seul intérêt du travail proposé de manière désintéressée. Au contraire, les motivations extrinsèques les poussent à contribuer afin d'obtenir un bénéfice extérieur au travail lui-même comme la reconnaissance, les récompenses symboliques, matérielles ou financières.

Ce dernier type de motivation semble moins étudié. Le crowdsourcing rémunéré, en particulier, a encore très peu été expérimenté et étudié dans les bibliothèques.

Conclusion

Les analyses que nous avons produites portent sur un corpus qui nous semble le plus exhaustif et fiable chaque publication ayant été consultée et sélectionnée. On observe une croissance de la littérature sur le crowdsourcing appliqué aux bibliothèques numériques, en particulier dans les pays anglo-saxons. Le domaine de la gamification intéresse, en particulier, beaucoup les communautés scientifiques. Celui du crowdsourcing rémunéré nous semble encore avoir fait l'objet d'assez peu d'expérimentations. Il serait donc judicieux d'en envisager l'expérimentation prochaine.

Bibliographie

Andro, M., Rivière, P., Dupuy-Olivier, A., Gropallo, F., Maingreaud, D. (2014). Numalire, une expérimentation de numérisation à la demande du patrimoine conservé par les bibliothèques sous la forme de financements participatifs (crowdfunding). Bulletin des Bibliothèques de France, contribution du 2 octobre 2014, 9 p.

Andro, M., Saleh, I. (2015, 1). Bibliothèques numériques et gamification : panorama et état de l'art. I2D - Information, données & documents. 52(4), p. 70-79.

Andro, M., Saleh, I. (2015, 2). La correction participative de l'OCR par crowdsourcing au profit des bibliothèques numériques. Bulletin des Bibliothèques de France, Contribution du 16 juin 2015.

Andro, M. (2016, 1). Crowdsourcing thesaurus. http://dx.doi.org/10.5281/zenodo.59507