Ethique et construction collaborative de données lexicales par des GWAPs (quelques leçons tirées de l'expérience JeuxDeMots)

M. Lafourcade[1], N. Le Brun[2]
[1] LIRMM, université Montpellier [2] Imagin@t, 55 rue Danton 34400 Lunel

La construction de données lexico-sémantiques a trouvé dans les approches collaboratives de nouveaux moyens de s'exprimer. Que ce soit dans les approches purement coopératives (du type Wiktionnaire ou Wikipedia) ou celles se réalisant à travers des jeux de type GWAP (Games With A Purpose) la question de l'éthique se pose, aussi bien au niveau du processus de construction des données que de leur devenir.

Le projet JeuxDeMots, démarré officiellement en septembre 2007, a été l'occasion d'un débat autour des questions d'éthique en rapport avec la création et l'exploitation de ressources lexicales. Le jeu principal (jeuxdemots.org) est accompagné d'une collection de "contre-jeux" ainsi que d'un environnement directement contributif (DIKO : http://www.jeuxdemots.org/diko.php). Cet environnement dans sa globalité participe à la création d'une ressource unique qui prend la forme d'un réseau lexico-sémantique. Depuis septembre 2007, plus de 300 000 termes ont été introduits dans le réseau ainsi que 13 millions de relations lexicales et sémantiques. Une grande partie des termes introduits l'ont été sur l'initiative de joueurs ou contributeurs, et environ 50 % des relations créées sont directement issues des 1,5 million de parties jouées depuis le lancement du jeu. Les autres relations proviennent des contre-jeux et des mécanismes de l'environnement.

Les objectifs du projet doivent être annoncés

Il semble légitime de présenter clairement les tenants et aboutissants du jeu dans le contexte du projet de création de ressources, dès lors qu'il apparaît clairement comme l'instrument d'un programme de recherche. Une proportion non négligeable de joueurs ont en effet la curiosité de savoir 1) comment leur activité ludique peut générer des données, 2) quelles peuvent être les applications des ressources ainsi produites. De plus, au-delà de l'exigence de transparence, l'assurance de participer, via un jeu, à quelque chose d'utile et de sérieux peut non seulement être puissamment motivante pour certains joueurs, mais également déculpabiliser ceux pour qui "jouer" rimerait avec perte de temps… Ainsi, sans pour autant rentrer dans un exposé exhaustif des mécanismes qui sous-tendent le jeu, un lien vers un résumé succinct du projet, accompagné d'éléments montrant que l'activité des joueurs génère des résultats, voire des retombées concrètes, peut-être perçu à la fois comme une reconnaissance et un encouragement à poursuivre. Avoir connaissance, même de manière concise, des résultats générés est aussi une forme de justification du temps investi par le joueur, qui peut tirer une certaine fierté d'avoir, en collaborant au travail de chercheurs, apporté sa pierre à un édifice en construction. JeuxDeMots tout comme ZombiLingo [Fort et al., 2014] donne un accès direct et immédiat aux données produites.

Notons ici que la motivation liée à la fierté de servir une "noble cause" est un paramètre particulièrement et soigneusement entretenu par les concepteurs des jeux qui ont comme retombées potentielles des applications médicales, comme Foldit, Eyewire, ou EteRNA.

Bien entendu, certains joueurs ne s'intéresseront pas du tout pas aux buts scientifiques qui sous-tendent le jeu et focaliseront leur intérêt sur ses qualités ludiques et l'intérêt du challenge proposé, sans se préoccuper du reste.

Des données personnelles protégées

Il semble également capital que le joueur sollicité pour produire des données via un GWAP soit totalement rassuré quant à la sécurité et la confidentialité des informations qu'il va fournir pour s'inscrire. Certes, on peut jouer en "invité" mais s'enregistrer (login + mot de passe + adresse mail) donne accès à beaucoup plus de possibilités, aussi les joueurs sont-ils vivement incités à le faire. Beaucoup de gens répugnent à donner leur adresse mail de peur d'être inondés de spam (comme c'est souvent le cas avec les jeux gratuits), c'est pourquoi ils doivent recevoir l'assurance que leurs données ne seront ni communiquées à des tiers, ni accessibles. De plus, toujours dans un souci de protéger l'anonymat des joueurs, aucune information personnelle n'est demandée à l'inscription ou par la suite, et à tout moment, un joueur peut demander la suppression de son compte, qui consistera à anonymiser login et mot de passe et à supprimer l'adresse mail.

Aucun gain en numéraire ou lots quelconques

Outre le fait que la participation à un GWAP doive impérativement rester gratuite, il semble aussi important, pour préserver la dimension éthique de l'approche ludo-collaborative, et la qualité des données récoltées, que les joueurs ne gagnent rien d'autre que des points, qu'ils ne reçoivent en particulier ni lot concret, ni gain en numéraire qui pourrait faire assimiler le jeu à un

"travail", à une source de revenus. Sans développer les questions relatives à la législation du travail, il faut garder à l'esprit que le crowdsourcing peut soulever des problèmes délicats quand il devient une activité rémunérée. Notons ici que l'activité de crowdsourcing à la base de ReCaptcha [Ahn *et al.*, 2008] est carrément qualifiée de "travail forcé" par [Good et Su, 2013].

Devoir payer des volontaires pour s'assurer leur concours, y compris via des approches ludiques, est une forme de dérive de la science dite "citoyenne" qu'on peut légitimement craindre, selon Eric Hand, journaliste à la revue Nature [Hand, 2010] : Pour le moment, le concept de GWAP est encore relativement récent et les expériences en cours profitent de la vague de popularité qui accompagne et soutient la notion de science participative et assure l'adhésion du public, valorisé de participer à des projets scientifiques. Mais après la phase d'engouement suscité par la très forte médiatisation de projets comme Foldit ou Eyewire, les gens risquent à la fois de se lasser et, en prenant conscience de leur valeur et de leur rôle, de devenir exigeants au point de se sentir exploités et manipulés si on sollicite leur concours bénévole. C'est pourquoi certains chercheurs craignent que dans un futur pas si lointain, il ne soit nécessaire de rémunérer les gens pour s'assurer leur concours via les jeux.

Aucune publicité

Aucune publicité n'est présente dans le site du projet JeuxDeMots, et curieusement quelques joueurs s'en sont étonnés (sans pour autant s'en plaindre). L'absence de publicité renforce la crédibilité du projet en montrant qu'il n'a aucune visée commerciale.

La méthode de création des données doit être transparente...

Dans quelle mesure les mécanismes sous-jacents du jeu permettant la création des données doivent-ils être rendus publics ?

...dans leurs grandes lignes
Les règles du jeu sont présentées dans leur globalité. Certains documents, produits par les joueurs eux-mêmes, décortiquent le jeu et fournissent des explications et des conseils stratégiques pour en exploiter au mieux les possibilités afin d'être parmi les mieux classés. Ainsi, une communauté de joueurs se crée spontanément autour du jeu en tant que projet, et leurs échanges peuvent influer sur l'évolution des mécanismes de jeu.

...pas nécessairement dans les détails de la mécanique mise en œuvre
Tous les mécanismes internes du jeu ne sont pas pour autant dévoilés aux joueurs. La raison principale est la crainte qu'ils ne tentent de les exploiter afin d'augmenter de façon abusive leurs propres performances. Ceci pourrait avoir deux effets indésirables : (a) l'agacement légitime de joueurs non avertis qui ne parviendraient pas à égaler les mieux classés tout en percevant qu'ils exploitent une astuce, (b) des performances de jeu accrues, mais au détriment de la qualité des données, moindre que lors d'une expérience de jeu normale.

Ceci amène à la question générale de la triche. La triche est un facteur de démotivation pour les joueurs qui ne trichent pas, et doit donc être évitée, même si certaines formes de triche peuvent en fait accélérer le processus d'acquisition lexicale. Dans ce dernier cas, il faut se demander si la méthode de triche employée peut être intégrée au jeu et en avertir les joueurs.

Un projet collectif

Les données créées par la foule reviennent à la foule
Les données produites dans le contexte du projet JeuxDeMots sont accessibles au public et libres de droit. Mises à disposition sous licence libre CCO, elles font donc partie du domaine public. Ce point a été particulièrement apprécié par de nombreux joueurs, qui, se sentant collectivement impliqués dans un projet de construction d'une ressource libre, ne se considèrent pas comme un rouage d'une entreprise dont le produit leur échapperait.

La question de l'accessibilité des données n'est jamais évoquée dans de très nombreux GWAP, parmi lesquels les GWAP les plus populaires en biologie que sont Foldit, EteRNA ou encore Eyewire. Inversement, Jérôme Waldispülh, créateur de Phylo, GWAP d'alignement de séquences génétiques, soulève le problème du statut des données issues de crowdsourcing et regrette que dans la plupart des cas, elles ne soient pas accessibles. Il va plus loin en créant une structure baptisée Open-Phylo, qui permet aux chercheurs du monde entier d'utiliser Phylo pour faire aligner leurs propres séquences par les joueurs [Kwak et al., 2013]. Cependant, bien qu'il milite pour l'accessibilité des données, et mette en évidence l'interface de téléchargement des séquences à faire aligner via le jeu, on cherche en vain comment se procurer les données issues du jeu. Nous n'avons pas pu mettre en évidence l'accessibilité des données pour les jeux suivants : Foldit, EteRNA, Nanocrafter, Fraxinus, Eyewire, Citizen Sort, Nightjar Project, Nanodoc, Dizeez, The Cure, Malaria Training Game.

Précisons tout de même que nous n'avons pas contacté les auteurs pour leur demander s'ils étaient d'accord pour communiquer les données issues de leur GWAP. Un "travers" possible est de déclarer que les données sont librement

accessibles, mais sans mettre explicitement à disposition un fichier téléchargeable, et de ne pas répondre aux mails demandant lesdites données…

Associer les joueurs au projet
Associer la communauté de joueurs aux publications scientifiques est la stratégie adoptée par les 3 grands GWAPs biochimiques américains (Foldit, EteRNA et Eyewire) et peut être vu comme une manière de rendre hommage aux joueurs en leur conférant le statut de collaborateurs à part entière, de maillon indispensable sans lequel les résultats n'auraient pas vu le jour. Sur le plan éthique, ça coupe court à toute accusation d'"exploitation", puisque la communauté de joueurs est reconnue et désignée, et que le fait de l'associer aux résultats est très largement médiatisé dans la presse généraliste.

Conclusion

La transparence dans les objectifs et les modalités d'un GWAP, ainsi que le statut et le devenir des données construites sont des éléments éthiques majeurs. Ceux-ci ne sont pas pour autant une garantie de la réussite du projet, le jeu devant par ailleurs être attractif et intéressant. On insistera, en particulier, sur la mise à disposition régulière, et de façon libre de droit, des données obtenues lors du projet. Il s'agit à la fois de donner la possibilité à quiconque d'en évaluer les résultats et de fournir aux joueurs un juste retour sur le temps investi dans le jeu. Ne pas le faire, c'est risquer au mieux ne pas arriver à fidéliser les joueurs et au pire d'alimenter une suspicion légitime sur le projet.

Il faut par ailleurs insister sur le fait que tous les types de ressources relevant du TAL ne se prêtent pas à l'utilisation de GWAP pour leur construction. Il faut être capable de présenter la tâche à réaliser de façon simple, didactique et ludique, de faire en sorte qu'elle présente effectivement un intérêt pour le joueur avec un bon équilibre entre difficulté et faisabilité. Enfin, il faut clairement identifier le couplage entre "bien jouer" et "produire des données de qualité". Les éléments éthiques présentés ci-dessus sont déterminants dans ce type d'entreprise, l'expérience montrant que leur non-respect aboutit assez rapidement à une mort du projet, faute de joueurs. Le grand cimetière des GWAP passés en est le témoin [Lafourcade et al., 2015].

Références

[Ahn *et al.*, 2008] Luis von Ahn, Ben Maurer, Colin McMillen, David Abraham and Manuel Blum (2008). "reCAPTCHA: Human-Based Character Recognition via Web Security Measures" Science 321 (5895), pp. 1465–1468.

[Fort *et al.*, 2014] Fort K., Guillaume B., Stern V. "ZOMBILINGO : manger des têtes pour annoter en syntaxe de dépendances." Dans TALN - Traitement Automatique des Langues Naturelles (2014), pp. 15-16.

[Good et Su 2013] Good B. M. and Su A. I.(2013) "Crowdsourcing for Bioinformatics", in Bioinformatics (2013) 29 (16): 1925-1933.

[Hand 2010] Hand E.(2010) "Citizen science : People power" Nature(2010) 466 pp. 685-687.

[Kwak *et al.*, 2013] Kwak D., Kam A., Becerra D., Zhou Q., Hops A, Zarour E., Kam A., SarmentaL., Blanchette M., and Waldispühl J. (2013) "Open-Phylo : a customizable crowd-computing platform for multiple sequence alignment." Genome Biology (2013) 14 :R116.

[Lafourcade et Joubert, 2010] M. Lafourcade, A. Joubert (2013) "Bénéfices et limites de l'acquisition lexicale dans l'expérience JeuxDeMots." In: Gala, Núria et Michael Zock (dir.), Ressources Lexicales: Contenu, construction, utilisation, évaluation. 2013. Linguisticae Investigationes, Supplementa 30, John Benjamins, 364 pages. (pp. 187–216).

[Lafourcade et al., 2015] M. Lafourcade, N. Le Brun, et A. Joubert (2015) "Les Gwaps" Editions ISTE, 125 p. (à paraître).