L'expérience RECITAL, un tremplin vers le développement du Crowdsourcing pour les Humanités

Benjamin HERVY[*,**], Guillaume RASCHIA[*]

[*]LS2N
prénom.nom@ls2n.fr
http://www.ls2n.fr/
[**]Mazedia
http://www.mazedia.fr

1 Introduction

Le crowdsourcing consiste à mobiliser la créativité, l'intelligence et le savoir-faire d'un grand nombre de personnes pour réaliser une tâche. Traditionnellement, cette tâche est fragmentée en micro-tâches indépendantes, et soumise pour résolution au plus grand nombre, par l'intermédiaire de plates-formes numériques ouvertes.

2 Pourquoi le Crowdsourcing dans les Humanités ?

Porté par le mouvement des humanités numériques, un nombre croissant de programmes de recherche en SHS requiert en préambule la constitution d'une base de données électronique à même d'être soumise à des mécanismes d'enrichissement, de filtrage, de visualisation, d'exploration ou d'analyse. Or, la collecte et le formatage des données, terreau des nouvelles problématiques de recherche, est une activité à la fois fastidieuse, coûteuse et peu valorisée en matière de résultats et de carrières scientifiques. À certains égards, la tâche ne requiert que très peu d'expertise pour être accomplie, même s'il est prudent d'évaluer au cas par cas cette hypothèse avant de soumettre un projet de collecte à une communauté de citoyens-bénévoles. Parmi les tâches de crowdsourcing adaptées à des questions de recherche, nous citons pêle-mêle la transcription de corpus littéraires pour l'édition, la catégorisation et l'annotation d'archives documentaires, l'observation de terrain, l'indexation de registres, la transcription d'enregistrements vocaux, etc.

3 Les plates-formes de crowdsourcing emblématiques

À ce jour, et à notre connaissance, Zooniverse est la version la plus aboutie de plateforme de science citoyenne en régime de production. Il s'agit d'un service hébergeant

88 projets de production participative, dont une très grande majorité est classée en santé, biologie, environnement ou astronomie, et comportant une vaste communauté de bénévoles (environ 1,8 million de comptes), pour la plupart anglophones.

Parallèlement, les technologies pour le crowdsourcing ont acquis une maturité indéniable, doublement attestée par d'une part, l'avènement de plates-formes mondiales de production participative monétisée telles Amazon Mechanical Turk, et d'autre part, le regain d'intérêt de la communauté scientifique en informatique pour les questions liées à la mise en œuvre de telles plates-formes.

Néanmoins, l'essor des sciences participatives peine à s'inscrire dans le champ des humanités numériques, malgré un besoin grandissant de ce type d'outils pour les projets de recherche en SHS.

Parmi les initiatives emblématiques figurent des travaux de transcription d'œuvres littéraires tels Transcribe Bentham à propos des manuscrits inédits de Jérémy Bentham (21 609 pages à transcrire dont 20 779 ont été certifiées depuis 2012). En France, le projet Testaments de Poilus hébergé par la TGIR Huma-Num, propose la transcription et l'encodage en TEI de 354 documents des fonds des Archives nationales et des archives départementales des Yvelines, pour une édition électronique. Enfin, nous signalons l'existence de la plate-forme RECITAL d'annotation et de transcription de 25 250 pages de registres du théâtre de la Comédie-Italienne. On y trouve de la catégorisation de documents en 8 classes distinctes, de l'annotation selon 133 types d'information à identifier dans la page, de la transcription de texte, et une tâche singulière de vérification. La vérification consiste à évaluer les propositions de transcription soumises par d'autres bénévoles, afin de réduire l'erreur dans les données produites. Au 12 mars 2019, RECITAL comptabilisait un peu moins de 200 000 actions réalisées par des bénévoles.

Les institutions liées au patrimoine culturel et les bibliothèques tendent elles aussi à multiplier les initiatives en matière de transcription participative, suivant des agendas qui leur sont propres et indépendamment de toute question scientifique. Par exemple, la plate-forme Transcribathon propose la transcription participative de 51 652 documents d'archive de la Première Guerre Mondiale, issus des fonds Europeana. Au Royaume-Uni, la British Library expérimente le dispositif libCrowds comportant à ce jour deux collections à transcrire ou annoter. En substance, et malgré l'existence de plates-formes de production participative mutualisées comme Zooniverse ou "à usage unique" telles Transcribe Bentham, la communauté de recherche en SHS se trouve être le parent pauvre d'un mouvement de fond qui a pourtant toute sa place dans le champ des humanités numériques.

4 RECITAL : les défis d'une mise en œuvre

Bien que satisfaisante, l'activité RECITAL [1] souffre de fortes irrégularités, dues pour très grande part à la difficulté à constituer et entretenir une communauté de bénévoles spécifique au projet : les comptes utilisateurs sont propres à la plate-forme, et l'animation, sous forme de lettre d'information mensuelle, de forum de discussions, ou de journées de transcription, est entièrement à la charge du projet.

1. quantité d'actions réalisées par les bénévoles sur une période donnée (heure, jour, mois, etc.)

RECITAL a été construit à partir du framework open source *ScribeAPI*, une initiative conjointe de la New York Public Library et de Zooniverse. Néanmoins, l'effort de développement spécifique est significatif. Il est en effet intéressant de pouvoir faire évoluer certaines fonctions de la plate-forme, comme la stratégie d'affectation des tâches aux bénévoles, ou l'évaluation de la fiabilité des bénévoles et des données produites. Un suivi de l'activité, global et individualisé, sous forme de tableau de bord est également souhaitable, ou encore des mécanismes de ludification (badges, parcours d'apprentissage, défis, grades, classement, etc.) pour contribuer à mobiliser et fidéliser les citoyens-bénévoles. Outre les compétences techniques requises pour réaliser cette maintenance corrective et évolutive, il y a également un effort important de production de contenu pour rendre la plate-forme plus accessible et attractive (enrichissement continu des tutoriels, de l'aide en ligne sous forme textuelle, sonore ou vidéo).

Par ailleurs, le développement d'une plate-forme de crowdsourcing requiert une attention particulière du point de vue de la conformité au Règlement Général sur la Protection des Données et de la propriété intellectuelle des données collectées.

Enfin, les données issues de la collecte doivent faire l'objet de nombreux traitements pour leur nettoyage, leur mise en forme, et leur intégration dans des écosystèmes *Findable-Accessible-Interoperable-Reusable* prêts à l'emploi en termes de filtrage, de statistique descriptive et de visualisation. De surcroît, un défi essentiel qu'il est impératif de relever pour favoriser une large adhésion des dispositifs de science citoyenne auprès de la communauté SHS, consiste à assurer la transparence des traitements de données de sorte à être en mesure de produire à tout instant la trace intégrale depuis la source jusqu'à la donnée. Cette exigence de traçabilité s'accompagne d'une autre, tout aussi essentielle, qui impose de produire des indicateurs de fiabilité combinant l'incertitude issue du processus de collecte (crowdsourcing) avec celle provenant de la chaîne de post-traitements.

5 Conclusion

En définitive, la plate-forme RECITAL, « à usage unique », délivre un service de crowdsourcing utile au projet de recherche qu'elle sert, techniquement avancé, mais coûteux à faire vivre et évoluer, et dont la plupart des coûts observés tirerait profit d'une mutualisation au sein d'une plate-forme partagée par un ensemble de projets. Ce constat ambivalent tiré de l'expérience RECITAL motive la proposition d'un nouveau service de crowdsourcing à destination des programmes de recherche en SHS, Arts et Lettres, qui puisse être également un terrain d'expérimentation pour des recherches en science et ingénieurie des données.

Néanmoins, le développement du crowdsourcing pour les humanités présente de nombreux défis : mise en œuvre technique (matérielle et logicielle), choix de conception, gestion de communauté, évaluation automatique de la fiabilité, post-traitements et restitution des données recueillies, etc. Il s'agit, selon nous, d'un service qui requiert toute l'attention et le savoir-faire de la communauté de recherche en Informatique, et donc d'un sujet au cœur des humanités numériques.

Le Crowdsourcing pour les Humanités

Summary

Crowdsourcing is a key concept of citizen science. Despite numerous success stories such as the *Galaxy Zoo* platform or the *Foldit* serious game, it seems to have difficulties entering Humanities. Based on our own 4-years experience of a fully-featured crowdsourcing platform for History project, we are going to discuss about challenges of such a development in the broad area of Digital Humanities.