Ouvrir le dédale des données des recherches myriadisées

Lisa Chupin (1), Karën Fort (2)

(1) - Maîtresse de conférences - Université Paris-Descartes, laboratoire Dicen-IDF, le Cnam, 292 rue St Martin, 75003 Paris, France
(2) - Maîtresse de conférences - STIH - EA 4509, Sorbonne Université, 28 rue Serpente, 75006 Paris, France
lisa.chupin@parisdescartes.fr, karen.fort@sorbonne-universite.fr

Chapô :
A la diversité des formes de participation proposées par les projets du réseau ParticipArc correspond une grande variété de modalités d'enregistrement et de mise en valeur des contributions. Nous nous intéressons aux projets du réseau qui fédèrent des collectifs par des plateformes en ligne. Nous montrons que seule une partie des données conservées est visible des contributeur·rice·s et librement téléchargeable.

Texte :
En favorisant une approche transverse des démarches participatives quelles que soient les disciplines de référence des porteur·se·s de projet, le réseau Particip-Arc a fait émerger une préoccupation commune de valoriser et d'ouvrir les données issues des démarches participatives, en particulier pour mettre en conformité les recherches participatives avec les exigences réglementaires et éthiques d'ouverture des données de recherche, mais aussi de protection des données personnelles.

Notre étude est donc centrée sur les projets de ce réseau, constitués autour d'une plateforme proposant aux internautes la réalisation de tâches d'annotation, de catégorisation ou de saisie d'observations sous une forme prescrite par un protocole, mais avec l'objectif de montrer la différence entre les données enregistrées par les plateformes et le jeu de données scientifiques réutilisables, différence qu'on pouvait aussi observer entre les synthèses issues des ateliers participatifs et les proposition des participant·e·s.

Les projets représentés dans le réseau nous ont permis d'étudier la production et l'ouverture de données dans différents champs disciplinaires : la biodiversité, avec les observatoires participatifs du Muséum National d'Histoire Naturelle (MNHN), l'enrichissement de corpus de documents patrimoniaux numérisés (transcription de *Testaments de Poilus*[1], transcription d'étiquettes d'herbiers, *Les Herbonautes*[2]) et l'annotation de corpus linguistiques (avec un jeu reposant sur la ludification de tâches d'annotation grammaticale, *Zombilingo*[3], et un site de partage et d'annotation de recettes rédigées dans différentes variantes de l'alsacien, *Recettes de grammaire*[4]). Nous avons ajouté aux projets du réseau deux plateformes contributives rassemblant un très grand nombre de contributeur·rice·s, *Zooniverse*[5] (portail de projets de catégorisation et d'annotation d'images multidisciplinaires), et *Tela Botanica*[6], réseau de botanique francophone.

Nous avons réalisé une double recension des données collectées sur les différentes plateformes

1 Archives nationales, https://testaments-de-poilus.huma-num.fr/#!/
2 http://herbonautes.mnhn.fr/
3 Fort, K. *et al*, https://zombilingo.org/
4 Millour, A. *et al*, https://bisame.paris-sorbonne.fr/recettes/info
5 Zooniverse, http://zooniverse.org/
6 Réseau Tela Botanica, https://www.tela-botanica.org/

d'une part, et de celles qui sont effectivement ouvertes et mises à disposition d'autre part, tenant compte des interfaces offrant un accès aux données, des formats d'export proposés ainsi que du statut juridique des données et de son explicitation.

Le dédale des données de recherche myriadisées

Un premier ensemble de données collectées concerne les contributeur·rice·s. Outre un identifiant et un mot de passe nécessaires à l'inscription, d'autres données les concernant peuvent être enregistrées, et apparaître sur le site. Le profil peut être enrichi automatiquement d'indicateurs de participation (par exemple les badges obtenus ou le nombre de contributions réalisées), la rétribution symbolique reposant notamment sur la mise en évidence des profils des contributeur·rice·s les plus impliqué·e·s. Si les contributions sont ainsi associées à des données personnelles, la donnée scientifique s'en démarque toutefois du fait qu'elle est produite par la synthèse de contributions de plusieurs personnes en interaction. L'enregistrement des contributions produit un volume de données d'autant plus imposant que les versions des contributions sont parfois enregistrées, et s'ajoutent aux actions de correction et validation également conservées. Ces données, que nous qualifions d'intermédiaires, au sens où elles contribuent à produire une unique donnée validée, sont d'autant plus abondantes que le fonctionnement du site suppose des interactions entre contributeur·rice·s. En outre, chaque contribution, ou chaque objet auquel elle se rapporte, peut lui-même être commenté.

Quatre modèles de mise à disposition des données de recherches myriadisées

Tela Botanica illustre un premier modèle correspondant à la déclinaison de critères de l'*open data* aux données d'observation de la biodiversité, et une ouverture de toutes les données disponibles téléchargeables sous licence libre. Le·la contributeur·rice peut saisir ses données en choisissant de les garder privées, sans les publier. Dans le cas où l'utilisateur·rice accepte de publier ses données, celles-ci le sont alors sous licence libre et sous condition de citation de l'auteur·e de la contribution.

La solution trouvée à l'équilibre entre le droit à l'oubli et la préservation de l'intégrité des jeux de données constitués à partir des contributions est la suivante : le·la contributeur·rice peut supprimer ses données même après avoir autorisé leur publication, « sans toutefois pouvoir revenir sur les droits cédés antérieurement lors de leur publication initiale, dans le cas où ces données auraient déjà été utilisées par des tiers. »[7] Ce modèle suppose l'acceptation par le·la contributeur·rice d'ouvrir ses données d'observations mais également celles qui concernent son profil et son activité sur la plateforme.

L'accessibilité sur demande aux données de recherche myriadisées sous licence libre

Le second modèle d'ouverture, qu'on retrouve dans de multiples sites (les portails *Zooniverse*, différents sites d'observation de la biodiversité du MNHN), est celui qui s'en tient à un encadrement juridique de la participation sans fournir d'interface de récupération des jeux de données produites (seule la consultation d'une partie limitée des données – les dernières produites – est possible), quand bien même elles sont placées sous licence libre. Cette solution permet de fournir des jeux anonymes, expurgés des données qu'il serait risqué de communiquer sans en contrôler les utilisations – cas des données concernant des espèces protégées en biodiversité. L'accès aux données est conditionné par la disponibilité de l'équipe du projet à échanger avec les personnes souhaitant les récupérer, avec un risque de fermeture de fait des données. Cela n'est toutefois pas un obstacle à l'utilisation scientifique de celles-ci, comme en atteste le grand nombre de publications alimentées par les données de plateformes qui ne rendent pas accessibles à tous les données par un espace de téléchargement en format ouvert et sous licence libre.

---

7 https://www.tela-botanica.org/thematiques/flora-data/#comment-participer

Ouverture des données de recherche myriadisées sous licence libre par détachement de leur contexte de production

D’autres projets ouvrent une partie des données consignées dans la plateformes, qui constituent des données de recherches, détachées de leur contexte de production participatif. La plateforme de transcription participative *Les Herbonautes* exporte ainsi vers la base de données des collections d'histoire naturelle *Recolnat* la description standardisée de chaque document d'herbier, sans mentionner les enrichissements apportés en commentaire par les contributeur·rice·s. La même forme de sélection d'une partie des contributions pour composer le jeu de données proposé au téléchargement s'observe aussi dans le jeu *Zombilingo*, comme dans d'autres jeux permettant de produire des corpus annotés pour les recherches linguistiques : le corpus annoté, issu des parties jouées par de multiples contributeur·rice·s, est téléchargeable, sans qu'il soit possible de remonter facilement d'une annotation à la partie qui l'a générée, ni aux interactions des joueur·se·s dans les forums. De fait, un travail de recomposition de l’ensemble du texte, qui a été segmenté dans les tâches d’annotation proposées aux contributeur·rice·s, est nécessaire à son exploitation et l’export est basé sur un algorithme créé par les chercheur·se·s.

La sélection de certaines des données par rapport à d'autres qui leur sont associées participe à la construction d'un jeu de données considéré comme scientifique qui se distingue de ce que serait un simple export (rendu anonyme) de l'ensemble des contributions. Comme dans d'autres cas observés de constitution de jeux de données ayant vocation à intégrer des espaces de dépôt internationaux (Heaton et Proulx, 2012), le caractère standardisé de la donnée scientifique, garant de sa comparabilité, est obtenu par une mise en invisibilité du contexte de sa production et des informations qui le décrivent. Le détachement de la contribution par rapport aux informations concernant son auteur·e (profil, activité sur la plateforme par exemple) garantit aussi la possibilité de faire circuler la donnée dans les écrits scientifiques en réglant le problème de la confidentialité de certaines des données contextuelles.

Enfin, les solutions de consultation et de visualisation directement sur la plateforme de contribution des données indisponibles au téléchargement peuvent pallier leur accessibilité réduite. Nous qualifions ce dispositif de communication des données comme une solution de consultation des données ancrées dans le contexte de leur production. Elle ne correspond pas aux critères de l'ouverture des données de l'*Open Knowledge Fondation*, et correspond à un public et usage particulier, celui des contributeur·rice·s et de leur lecture des documents en relation avec les tâches qui leur sont proposées.

Références

Heaton, L., Proulx, S. (2012), « La construction locale d’une base transnationale de données en botanique », *Revue d’anthropologie des connaissances*, 6-1, p. 141-162.