Concevoir des dispositifs qui associent des traces numériques individuelles à des données d'enquête

Thomas Louail

Concevoir des dispositifs qui associent des traces numériques individuelles à des données d'enquête : le cas de l'écoute de musique enregistrée et le projet RECORDS

Depuis 2020, le laboratoire Géographie-cités (UMR8504, CNRS / EHESS / Université Paris 1 Panthéon Sorbonne / Université Paris Cité) pilote RECORDS, un projet collaboratif sur l'écoute de musique en streaming*, en partenariat avec le Centre de recherche sur les inégalités sociales (CRIS, UMR7049, CNRS / Sciences Po), l'équipe de sciences sociales computationnelles du Centre Marc Bloch (CMB, UAR3130, CNRS / MEAE / MESR / BMBF), ainsi que les départements recherche des entreprises Orange et Deezer. Dans ce projet, les chercheurs et chercheuses ont développé un dispositif qui associe* big data *et enquêtes par questionnaires et entretiens, pour étudier la diversité des consommations et des pratiques d'écoute sur les plateformes de* streaming.

Le *streaming*[1] est devenu le mode principal d'écoute de musique enregistrée dans de nombreux pays, dont la France[2]. Les plateformes de *streaming*, comme Spotify, Apple Music ou Deezer, enregistrent les historiques individuels de *stream* de leurs utilisateurs et utilisatrices afin de calculer la rémunération des artistes. Ces historiques individuels de *stream* listent de façon détaillée les écoutes de musique des utilisateurs, et permettent d'entraîner la recommandation algorithmique et personnalisée de musique. Ces données de *stream* constituent aussi une ressource très utile pour les recherches sur différents sujets, comme la circulation des œuvres ou l'influence des systèmes de recommandation sur la diversité musicale à laquelle sont exposés les individus qui les utilisent. Les plateformes elles-mêmes mènent des recherches sur ces sujets et les publient dans les actes de grandes conférences en informatique. Toutefois, en l'absence d'informations sociodémographiques précises sur les personnes à l'origine des *streams*, sur les contextes dans lesquels elles écoutent, ou encore sur leurs préférences musicales et culturelles, ces seules données se révèlent limitées pour situer socialement les écoutes et les usages, et soutenir des recherches utiles à l'élaboration de politiques culturelles en régime numérique. En effet, c'est lorsqu'elles sont associées à des données d'enquête que les traces numériques d'usage prennent tout leur intérêt pour l'analyse sociale des pratiques numériques et des consommations de contenus sur internet.

Au-delà du *streaming* musical, une part croissante de la population utilise des services numériques pour un grand nombre d'activités du quotidien : faire ses courses, prendre des rendez-vous médicaux, s'informer, se divertir, etc. Un enjeu important pour les sciences sociales est donc de concevoir des dispositifs d'enquête qui associent des informations déclaratives, collectées dans des questionnaires ou en entretien, à des données d'usage de ces services, si possible collectées dans des conditions naturelles de pratique, et ceci en garantissant l'anonymat des enquêtées et la sécurité de leurs données personnelles. L'évolution de la législation européenne (Règlement général sur la protection des données - RGPD) a constitué un progrès majeur dans la protection de la vie privée, tout en aménageant un cadre pour la recherche, qui doit notamment respecter un principe de proportionnalité de l'information collectée — on collecte ce qui est nécessaire mais pas plus. La conception de dispositifs d'enquête « mixtes », articulant données d'enquête et données observationnelles de pratique numérique, est donc un sujet d'actualité en méthodologie d'enquête. Plusieurs approches font l'objet de recherches. Des panels internet, comme le GESIS Panel. dbd en Allemagne, expérimentent des protocoles de collecte de traces digitales en demandant à leurs panélistes d'installer sur leur téléphone des *trackers* qui collectent le temps d'utilisation de différentes applications. D'autres proposent d'enrichir les enquêtes en ligne en donnant la possibilité aux répondantes de téléverser des contenus multimédia (voir, par exemple, les travaux conduits dans le projet ERC *Web Data Opp*). Une autre approche consiste à demander à des personnes volontaires de faire don de leurs données personnelles d'utilisation de plateformes très populaires, comme Facebook ou WhatsApp, en garantissant leur anonymisation et leur utilisation à des fins de recherche scientifique uniquement[3]. Enfin une autre façon de procéder consiste à collaborer avec une entreprise qui déploie un service numérique utilisé par une part importante de la population, et de conduire des enquêtes auprès de ses usagers consentants.

C'est cette approche qui a été suivie dans le projet RECORDS, une collaboration entre trois unités de CNRS Sciences humaines & sociales et les entreprises Orange et Deezer, financé par l'ANR sur la période 2020-2024. Ce projet a permis de concevoir et tester un dispositif qui associe pour plusieurs milliers d'enquêtées, utilisatrices et utilisateurs de Deezer, leurs historiques pluriannuels d'écoute de musique à leurs réponses à des enquêtes par questionnaire et par entretien. La conception d'un tel dispositif soulève des questions méthodologiques, relatives à l'échantillonnage des personnes sollicitées, les modalités de sollicitation, les incitations mises en place, ou encore la création d'indicateurs permettant de comparer les réponses des enquêtées à leurs traces d'écoute en *streaming*. Plusieurs vagues d'enquête d'ampleur croissante ont permis de faire la preuve de la pertinence de cette approche[4], et ces travaux doivent être poursuivis pour produire des données ouvertes et anonymisées, et plus représentatives, au sens des critères de la statistique publique, de l'ensemble des personnes qui utilisent ces plateformes.

Le principe de la collecte des données mixtes est le suivant : on diffuse à des centaines de milliers d'usagers de la plateforme de *streaming* Deezer une invitation à participer à une enquête en ligne. Les personnes sollicitées sont échantillonnées parmi toutes celles qui ont explicitement consenti (*opt-in*) à recevoir des offres de partenaires de Deezer — ici le CNRS. Ces invitations sont envoyées par mail et via des notifications directement dans l'application mobile Deezer. L'enquête en ligne porte à la fois sur les préférences musicales et culturelles, les habitudes d'écoute de musique en général (hors plateforme), les sorties

1. Le *streaming* est une technique informatique qui permet de visionner des contenus multimédia sans avoir à télécharger des fichiers au préalable.
2. Voir les rapports annuels du SNEP et de l'IFPI.
3. Un symposium européen rassemblant des équipes travaillant sur cette approche a eu lieu en mai dernier.
4. Voir Renisio Y., Beaumont A., Beuscart J-S., Coavoux S., Coulangeon P. & al. 2024, Integrating digital traces into mixed methods designs: An application to the study of online music listening using survey, interview and stream history data collected from the same people

culturelles, etc. et comporte un volet socio-démographique qui permet de mieux saisir qui sont les personnes. Les données sociodémographiques sont collectées en utilisant des indicateurs standards — comme la taille et la composition du ménage, la profession, le niveau d'études, etc. — et l'enquête profite des outils conçus par l'Insee et ses partenaires lors de la refonte de la nomenclature socioprofessionnelle (PCS2020). Pour les enquêtées consentantes, les chercheurs et chercheuses apparient ensuite les données d'écoute de musique en *streaming* aux informations sociodémographiques anonymisées, pour qualifier socialement les écoutes.

Cette qualification sociale de traces numériques d'écoute constitue une avancée méthodologique importante dans le champ de l'analyse socio-spatiale des consommations culturelles, et ce à plusieurs titres. Tout d'abord les historiques de *stream* collectés par les plateformes renseignent sur des événements d'écoute, là où les enquêtes de référence collectent des déclarations d'habitudes d'écoute et de préférences musicales. Cette différence de nature (écoutes effectives *vs* déclarations d'habitudes et de préférences) s'accompagne d'une seconde différence qui concerne la résolution de l'information collectée : alors que les enquêtes doivent nécessairement interroger à un niveau assez agrégé de description de la musique, comme les genres musicaux ou quelques artistes — parce qu'il serait inenvisageable de demander aux enquêtés leur avis sur de longues listes d'artistes — les données de *stream* renseignent sur les écoutes au niveau le plus fin : quels morceaux, albums et artistes ont été écoutés, quand, combien de temps et combien de fois ? Une troisième différence tient au caractère longitudinal des traces d'écoute: là où les données d'enquête donnent une photographie des habitudes et préférences à un instant de la vie, les historiques d'écoute ont été collectés pour chaque utilisateur et utilisatrice depuis la création de son compte (depuis 2018 dans cette enquête), donc possiblement pendant plusieurs années. Enfin, une autre différence notable est que les données d'écoute sont standardisées par défaut : elles sont collectées à l'identique dans tous les pays, tandis que les opérations d'alignement et d'harmonisation des nomenclatures utilisées dans les enquêtes nationales sont souvent difficiles et coûteuses.

Les données de *stream* présentent aussi des limites. Mises à part la date et l'heure, ainsi que quelques informations sur le dispositif matériel et logiciel supports de l'écoute, on sait peu de choses sur le contexte d'écoute : quel est le niveau d'attention de la personne à la musique *streamée* ? A-t-elle choisi la musique (écoute en groupe), et est-elle bien en train d'écouter (plusieurs personnes distinctes peuvent utiliser le même profil utilisateur) ? Si sur plusieurs mois ou années un grand nombre d'écoutes peut permettre de « lisser » les biais et offrir une bonne vue d'ensemble des préférences d'une personne, les enquêtes qualitatives menées dans RECORDS ont contribué à éclairer les différences entre écoutes et goûts. Un grand nombre d'écoutes d'un artiste n'est pas nécessairement un bon prédicteur du goût déclaré par l'auditeur pour cet artiste ; à l'inverse, un petit nombre d'écoutes ou des *skips* répétés d'un même morceau ne sont pas nécessairement un bon indicateur de dégoût, le contexte jouant un rôle important dans la sélection et l'appréciation de la musique. Une autre limite des données collectées dans ce dispositif, en comparaison des enquêtes nationales (en particulier l'enquête de référence Pratiques Culturelles conduite par le Département des études, de la prospective, des statistiques et de la documentation - DEPS), tient à ce que les personnes enquêtées ne relèvent pas de la « population générale ». Ici, la population de référence est l'ensemble des personnes qui utilisent la plateforme de *streaming* Deezer, où les jeunes et les hommes sont sur-représentés en comparaison de la population nationale. De plus, les répondants aux enquêtes ne sont pas représentatifs de l'ensemble des usagers (biais de sélection), et il est difficile de redresser les données d'enquête *a posteriori*, parce qu'on ne sait pas exactement quelle est la distribution des groupes sociodémographiques parmi l'ensemble des usagers d'une plateforme. Malgré cela, les données mixtes ainsi constituées permettent d'évaluer avec un matériau empirique inédit les modèles statistiques qui relient origines et positions sociales d'une part et consommations musicales de l'autre.

En amont de l'enquête, un travail avec les déléguées à la protection des données des organismes partenaires de la recherche est nécessaire pour garantir la licéité du traitement. Toute collecte et traitement de données personnelles doit avoir une base légale et une finalité. Le RGPD oblige les responsables de traitement à informer les personnes sur le traitement de leurs données, et impose un certain nombre de principes. Un traitement de données personnelles dans le cadre d'un projet de recherche public-privé ne peut pas avoir comme base légale une mission d'intérêt public, et il faut alors avoir recours à une autre base légale, le consentement des personnes ou l'intérêt légitime. Fonder légalement le traitement sur le consentement des personnes implique de pouvoir gérer ces consentements individuels (et en particulier d'éventuelles demandes de retrait) des dizaines de milliers de répondantes à l'enquête. Pour des raisons pratiques, à cette échelle il est difficilement envisageable de procéder ainsi, et la base légale de l'intérêt légitime est alors privilégiée. Aucune donnée personnelle considérée comme sensible par la CNIL n'est collectée dans l'enquête, ni aucune donnée directement ou indirectement identifiante. Un des enjeux de l'anonymisation des données, avant leur ouverture à des fins de recherche, est de s'assurer, par l'agrégation notamment, que les données ne permettent pas une identification par croisement d'informations (par exemple une femme médecin déclarant travailler dans une commune où n'exerce qu'une médecin). L'enjeu actuel est d'appliquer une procédure d'anonymisation qui permette d'ouvrir les données, en garantissant la non-identifiabilité des personnes tout en préservant un niveau raisonnable d'agrégation et de bruitage de l'information, dans l'intérêt des recherches sur la diversité socio-spatiale des consommations culturelles en régime numérique.

contact&info
▶Thomas Louail,
Géographie-cités
records@parisgeo.cnrs.fr
▶Pour en savoir plus
https://records.huma-num.fr

RECORDS