**Données individuelles localisées : enjeux juridiques, éthiques et techniques à l'heure du RGPD**

**Synthèse des interventions de la journée**

Plus d’un an après l’entrée en application du RGPD (Règlement général sur la protection des données), **Lionel Maurel** a décrypté plusieurs articles du texte afin d’examiner les principales mesures portant sur le traitement des données localisées et d’en apprécier les conséquences pour les activités de recherche. Si dans le RGPD, les données de géolocalisation sont appréhendées comme des « informations à caractère personnel » comme les autres, le texte permet d’encadrer les usages de ces données à très forts pouvoirs identifiants (notamment par croisement avec d’autres informations), dans un souci de protection de la vie privée. Mais le traitement de données personnelles à des fins de recherche n’étant pas interdit (il est considéré même comme légitime), le texte laisse aussi une importante marge de manœuvre aux chercheur·se·s, *via* l’existence de nombreuses dérogations aux principes de licéité énoncés dans le règlement : il est ainsi possible de déroger à la demande de consentement systématique, à l’annonce d’une finalité précise, au respect de la durée de conservation, etc. Il n’y a pas en revanche de dérogation au principe de minimisation des données, la collecte « au chalut » n’étant pas possible. De plus, il faut pouvoir prouver l’utilité des données pour la finalité du projet, cette finalité pouvant cependant rester large. La discussion a amené L. Maurel à évoquer plusieurs points particuliers d’application du RGPD, par exemple sur le périmètre géographique d’application du texte : s’applique-t-il à des données lointaines, comment encadre-t-il les activités de recherche menées dans le cadre de partenariats avec des équipes externes à l’UE ? Il a aussi été question du périmètre des activités de recherche concernées par les dérogations : s’appliquent-elles aussi à la recherche privée ? Il semblerait que oui, Facebook ayant par exemple d’ores et déjà justifié certains traitements en les assimilant à une activité de recherche, conduisant la CNIL à afficher sa vigilance.

Après avoir rappelé l’architecture d’internet et les principales composantes d’un site web, dont la compréhension est un préalable indispensable à l’extraction des données du web, **Sébastien Rey-Coyrehourcq et Alexandre Cebeillac** ont présenté différents protocoles de collecte par *webscraping*, en se basant sur leurs expériences dans différents projets de recherche. Ils ont abordé cette question en partant des API, assimilables à une sorte de contrat avec un producteur de données, jusqu’aux requêtes *http*, en passant par les procédures plus récentes et plus complexes liées au renforcement des protections *anti-bot*. Parmi ces données de natures diverses (images raster sur la congestion du trafic routier, flux de trafic aérien, etc.), certaines sont des données personnelles (localisation croisée des hébergeur·se·s et des visiteur·se·s d’*Airbnb*, estimation des déplacements individuels *via* *Twitter*, etc.) qui s’inscrivent dans un nouveau contexte de circulation des données fortement dissymétrique entre trois types d’acteurs : les usager·ère·s du web, les propriétaires des plates-formes et les utilisateur·trice·s tiers pratiquant le *webscraping*, dont font partie les chercheur·se·s mais aussi les entreprises privées et les Etats. Au-delà de la question du cadre légal de la collecte, comment les « tiers » chercheur·se·s définissent-ils leur positionnement éthique vis-à-vis de la protection des données personnelles extraites du web ? Si les situations problématiques concernant l’identification des personnes peuvent être résolues grâce à la mise en place de divers garde-fous (anonymisation des données, brouillage de leur précision temporelle, travail sur des fichiers encryptés et stockés sur des serveurs protégés, etc.), quelle est la part d’appréciation individuelle qui doit être laissée aux scientifiques pratiquant le *webscraping* ? Faut-il encadrer ces décisions par des équivalents de comités éthiques qui seraient chargés de ces questions au sein des universités et des institutions de recherche ?

La présentation de **Julie Fen-Chong, Françoise Lucchini et Bernard Elissalde** s’est inspirée d’une dizaine d’années d’expériences de recherche sur les données géolocalisées individuelles issues de la téléphonie mobile ou du réseau social *Twitter*. Les intervenant·e·s ont tout d’abord partagé une réflexion sur l’évolution des conditions d’accès à ces données, au sein de différents contextes (du contrat CIFRE aux partenariats publics-privés), tout en décrivant la nature des données diffusées, caractérisées par différentes granularités spatiales et temporelles. Après plusieurs années de partenariats plutôt fructueux, marqués par des difficultés de circulation des données avant tout internes aux entreprises (étanchéité des données techniques et des données de facturation pour la téléphonie mobile), cet accès est aujourd’hui d’autant plus restreint que la valeur commerciale des données s’est accrue, tandis que la finesse des données auxquelles on a maintenant accès s’est dégradée. L’autre volet de la présentation proposait un tour d’horizon des différentes méthodes d’analyse spatiale permettant de dépasser l’échelon individuel pour révéler des fonctionnements spatio-temporels territoriaux (lieux de concentration de la population, mouvements entre antennes, graphes par *clustering* d’agrégation spatiale ou temporelle…), afin de mettre en valeur les contextes géographiques favorables au mouvement ainsi que les conséquences du mouvement sur l’organisation des lieux.

**Antoine Courmont** s’est intéressé aux recompositions de pouvoir liées à la mise en circulation des données entre acteurs publics et privés. Ces échanges de données (open data, partenariats avec des plateformes) participent à l'émergence de nouveaux modes de gouvernance des territoires par la donnée, mais ils ne sont pas sans susciter un certain nombre de tensions. Les interactions observées s’inscrivent dans un double contexte de fin du quasi-monopole de la mise en données du monde par les institutions publiques et de montée en puissance d’entreprises privées produisant des données massives (Waze, Uber, Airbnb…), ces données étant de plus en plus envisagées par les acteurs publics comme un levier d’action pour accompagner ou contrer certaines pratiques urbaines. À travers la question particulière des données ouvertes, A. Courmont s’est demandé dans quelle mesure les acteurs publics pouvaient se placer au centre de la circulation des données et éviter des utilisations contraires à l’intérêt général, ce contrôle passant notamment par la maîtrise de la propriété des données et par le contrôle des formats et standards des bases. Plus généralement, la « privacy » tient souvent une place centrale dans ces négociations et peut parfois être utilisée comme un argument pour refuser la diffusion de données d’opérateurs privés, au motif qu’il s’agirait d’une enfreinte à la protection des données clients. L’accès aux données des plates-formes fait ainsi tout autant émerger des situations d’entente et de partenariat entre collectivités et entreprises (par exemple avec Waze) que des situations conflictuelles (par exemple avec Airbnb ou Uber), qui soulignent les enjeux plus généraux de gouvernance urbaine associés aux questions d’accès aux données localisées.

L’accès aux données individuelles localisées de source administrative ainsi que le traitement de ces données sont au cœur du travail de thèse de **Luc Guibard**, qui s'appuie sur les données sensibles de la CAF pour suivre les mobilités résidentielles au niveau individuel et analyser le processus d’éviction résidentielle dans la région parisienne. Il a d’abord présenté la manière dont s’était organisé concrètement son travail dans les locaux de la CAF, et notamment la longue période nécessaire à la rédaction d’une convention de mise à disposition de données administratives sensibles, dont le modèle n’existait pas encore. Puis il est revenu sur la structure et le traitement de ces données, l’enjeu principal étant de savoir comment tirer parti de la géolocalisation tout en assurant l’anonymat de données soumises au secret statistique. Ces données présentent en effet de nombreux avantages en termes de liberté d’agrégation spatiale par rapport au problème du MAUP. Il a ainsi présenté plusieurs types d’expérimentation d’anonymisation, reposant sur différents choix d’agrégation spatiale respectant le seuil de confidentialité. Cet exemple a aussi été l’occasion d’ouvrir la discussion sur l'utilisation des données individuelles par la CAF à des fins d’optimisation du service, notamment pour la détection des cas de fraude potentielle (scoring). Ces pratiques interrogent quant aux règles de protection des données individuelles s’appliquant en interne.

La présentation de **Malika Madelin** a clos la journée en proposant un retour d’expérience sur une recherche participative de mesure de la qualité de l’air conduite à l’aide de « capteurs citoyens » (projet *AirCitizen*). Plusieurs types de données individuelles sont collectés et manipulés dans le cadre de ce projet (données de localisation passive ou déclarée, données d’observation du contexte rapportées par les participant·e·s, données sur l’identité même des participant·e·s). M. Madelin a exposé les nombreuses questions soulevées par la mise en œuvre d’une démarche conforme au RGPD, en insistant sur les flous qui avaient accompagné l’application des principes du règlement à leurs travaux : quels risques d’identification des personnes liés aux données de localisation, même lorsque les noms ont été supprimés ? comment s’assurer de la pertinence des données collectées avant même de les avoir traitées ? jusqu’où limiter la conservation des données ? … Les chercheur·se·s du collectif ont ainsi adapté pas à pas leur protocole pour les phases de collecte comme de traitement (non communication des coordonnées personnelles, non conservation des chroniques d’activités, brouillage des repères temporels lors de la diffusion des cartes d’itinéraires individuels, …), au risque que le RGPD bloque certains volets de l’animation. Au final, les échanges mis en place avec les différents interlocuteurs du projet (capteurs-citoyens, collectivités et associations) ont souvent joué un rôle de sensibilisation à la protection des données personnelles, qui apparaît rarement comme une priorité dans le cadre de projets à fort engagement individuel.