Fondation Paris-Dauphine

governance & regulation

DAUPHINE UNIVERSITÉ PARIS

Les activités de surveillance : entre méthodes classiques d'observation et Big Data

Synthèse de séminaire

Séminaire du Club des Régulateurs

Université Paris-Dauphine, 30 septembre 2016

CLUB

Table des matières

Les activités de surveillance: entre méthodes classiques d'observation et Big Data

Séminaire du Club des Régulateurs,
30 septembre 2016

Les activités de surveillance des marchés ou de contrôle des comportements des opérateurs, communes à la majorité des régulateurs, reposent sur une variété de types de données, de collecte et de traitement. Ce séminaire vise à confronter les expériences de différentes autorités en la matière, mais également à identifier les nouveaux enjeux du traitement des données de surveillance et de contrôle.

Introduction

Jean-Yves Ollier
Directeur général de la CRE

Cette session s'inscrit dans la continuité de l'atelier sur les régulateurs sectoriels que nous avions organisé il y a un an, au cours duquel nous avions évoqué la transformation des activités de surveillance et de contrôle.

En effet, l'incidence sur nos missions de l'évolution des industries que nous régulons (automatisation du traitement des transactions sur certains marchés, jeux en ligne, etc.) est réelle : nous sommes de plus en plus largement chargés d'exercer des activités de surveillance nécessitant un recueil massif de données et l'intervention de méthodes de traitement automatisées permettant de filtrer ou pré-analyser ces données.

1ère table ronde : Les expériences de différentes autorités de régulation

Les autorités sectorielles

L'Autorité de régulation des jeux en ligne (Arjel)

Nicolas Boulanger
Responsable du traitement des données de contrôle

Créée en 2010, l'ARJEL (Autorité de régulation des jeux en ligne) s'est vu assignée plusieurs objectifs :

- délivrer des agréments et s'assurer du respect des obligations par les opérateurs ;
- protéger les populations vulnérables, lutter contre l'addiction ;
- s'assurer de la sécurité et de la sincérité des opérations de jeu ;
- lutter contre les sites illégaux ;
- lutter contre la fraude et le blanchiment d'argent.

Tous les logiciels de jeu sont audités avant leur mise sur le marché. L'ensemble du SI de l'opérateur est analysé avant le début de l'activité et lors des évolutions majeures du SI. Il est par ailleurs certifié chaque année.

Les opérateurs doivent transmettre à l'ARJEL les données qu'ils collectent sous le format défini par l'ARJEL, afin de permettre au régulateur d'effectuer, notamment via des tris, sélections et requêtes des activités de contrôle des obligations réglementaires des opérateurs. . Les rencontres et les compétitions sportives font également l'objet de contrôles automatisés, via des requêtes prédéfinies et des systèmes d'alerte déclenchés

en cas de suspicion de fraude sportive. Si une alerte est déclenchée, une analyse plus poussée sera effectuée. Enfin, des contrôles des joueurs seront prochainement mis en place pour améliorer la lutte contre le blanchiment et l'addiction - faisant suite à une évolution du cadre réglementaire.

L'Arjel procède également à des contrôles partiellement automatisés. Ainsi, un programme scanne toutes les minutes l'ensemble des cotes disponibles (recueillies par le coffre de l'Arjel à partir du frontal des opérateurs). En cas d'importantes variations avant ou pendant un événement sportif, une analyse de l'évolution suspecte de la cote est effectuée.

Tous ces contrôles automatisés (entièrement ou partiellement) sont rendus possible par le recours à un outil de collecte des données sous un format défini par l'ARJEL et commun à l'ensemble des opérateurs. Cet outil a été mis en place dès la création de l'ARJEL en 2010.

L'Autorité de régulation des communications électroniques et des postes (ARCEP)

Nicolas Desmons
Chef de l'unité Régulation par la donnée - Direction Internet et Utilisateurs

L'Arcep régule les opérateurs de communications électroniques et des postes. 1 900 opérateurs télécoms sont aujourd'hui déclarés, mais seuls 150 à 250 ont une activité significative. Sa principale mission consiste à assurer une concurrence effective et loyale au bénéfice de l'utilisateur. Dans ce cadre, l'Arcep dispose d'un pouvoir de sanction à l'égard des opérateurs et de règlement des différends en cas de problèmes techniques et de connexion entre les réseaux. Elle utilise les données en tant que telles, à froid et non en temps réel, comme outils de régulation.

Les données collectées proviennent de trois sources : l'Arcep (par exemple via l'enquête annuelle de qualité de service du réseau mobile), les opérateurs et les utilisateurs (remontées spontanées de dysfonctionnements). Les données collectées auprès des opérateurs sont de plusieurs types :

- des données agrégées et statistiques, le plus souvent à la maille trimestrielle sont publiées dans des observatoires sur le site de l'Arcep à des fins d'information de l'ensemble du secteur ;
- des données techniques à usage interne, dans l'optique de mieux comprendre le fonctionnement des opérateurs et les rapports de force ;
- des données sur le déploiement des réseaux mobiles, des données sur la qualité de service fixe et des cartes de couverture au format numérique, pour améliorer l'information des utilisateurs pour les guider dans leur choix.

Les données sont collectées via un extranet pour ce qui est des opérateurs et trois canaux pour ce qui est des utilisateurs (hotline téléphonique, mail et courrier). L'information est publiée sous la forme d'observatoires de l'Arcep et diffusée au format open data (données brutes).

A terme, l'Arcep souhaite davantage favoriser les remontées des utilisateurs, en mettant en place une plateforme de signalement sur son site Internet, pour pleinement appliquer des méthodes d'analyse de la donnée - notamment dans le cadre de sa nouvelle compétence de gardien de la neutralité d'Internet. Elle envisage également de développer le crowd sourcing (avec des outils permettant aux utilisateurs d'effectuer des tests et d'en remonter les résultats, comme cela existe déjà dans quelques pays en Europe), pour donner sa place au citoyen dans la régulation des télécoms.

Compte tenu de l'importance des enjeux, les outils doivent être solides et inspirer confiance. Aussi les méthodes de collecte de l'information doivent-elles être certifiées.

La Commission de régulation de l'énergie (CRE)

Matthieu Morin
Chef du département de la surveillance des marchés de gros

La CRE a pour principales missions de surveiller le respect des obligations qui incombent aux gestionnaires de réseau mais également aux opérateurs du marché de détail et du marché de gros. Sur le marché de gros, il s'agit notamment de surveiller les transactions et les ordres sur les marchés organisés d'électricité et de gaz naturel et des échanges aux frontières, ainsi que la cohérence entre le marché de CO2 et celui d'électricité et de gaz ou encore certains instruments financiers (dans le cadre d'un protocole d'accord avec l'AMF pour le partage de données). Le régulateur publie des observatoires trimestriels (données quantitatives de marché), un rapport annuel, des analyses et des enquêtes. Il dispose également d'un pouvoir de sanction.

Concernant le recueil des données, la CRE constate la règle « des trois V » : variété, volume et vélocité (ou fréquence). Sur le marché de détail, par exemple, les types de documents sont très variés (contrats, statuts d'entreprise, comptes sociaux, etc.) mais les acteurs sont peu nombreux. Le volume dépend de la nature des données collectées : de 100 à 20 000 lignes par flux, avec un volume global de 100 000 lignes pour les données quantitatives, mais une centaine de pages pour les données qualitatives. La collecte des données se fait au pas mensuel dans la majeure partie des cas. Pour les données qualitatives, le traitement est nécessairement manuel. Pour les données quantitatives, le traitement algorithmique peut s'avérer très lourd et complexe lorsqu'il s'agit de re-pricer les offres des fournisseurs, par exemple. Sur le marché de gros, la variété des données collectées est très élevée également, avec une cinquantaine de sources différentes et une centaine de flux selon le contenu métier (transactions, ordres, fondamentaux, publications). La même variété s'observe concernant les formats et les protocoles d'échanges. Par ailleurs, en rythme de croisière, les volumes attendus sont autour de 1Go/semaine, contre 7 à 10 To/jour pour Twitter ou Facebook - la CRE est donc encore loin du Big Data ! Enfin, les fréquences de réception diffèrent selon les flux, du journalier (J + 1) au trimestriel ou annuel. L'analyse s'effectue ex post.

Les traitements sont algorithmiques (détection d'alertes automatiques) ou relèvent de l'informatique décisionnelle automatisée (outils d'analyse, reporting) ou spécifique (enquêtes). L'une des principales questions actuellement posée est celle de l'articulation entre l'automatisation et l'analyse humaine. Plusieurs pistes de réflexion sont identifiées. Tout d'abord, plus le marché est mature ou de masse, plus les offres sont standardisées et plus la donnée peut être structurée, donc requêtée, ce qui permet de basculer d'une analyse qualitative vers une analyse quantitative. Ensuite, lorsque les données sont structurées, systématiques et exhaustives, une surveillance automatisée peut être envisagée. Pour autant, cette automatisation présente des difficultés techniques. Notamment, plus les données de marché s'approchent du marché de masse, plus l'on s'oriente vers une complexité de type Big Data, avec une problématique de stockage, de durée d'intégration, de rapidité d'analyse et de traitement ou encore d'indexation. Des contraintes existent sur le plan de la sécurité également, étant en possession de données commercialement sensibles et relevant du secret des affaires, le régulateur européen impose de définir une politique de

sécurité de l'information représentant une vingtaine de documents. La validation de ces documents s'effectue par une procédure de revue par les pairs (entre régulateurs nationaux). Enfin, l'automatisation pose des questions RH : le cœur de métier des régulateurs n'étant pas l'informatique, comment attirer, recruter et fidéliser des spécialistes?

Il importe également de s'interroger sur les perspectives et les limites de l'automatisation de la surveillance. Est-il possible d'utiliser de nouvelles technologies (Machine Learning, Big Data ou Big Analaytics si le marché se développe significativement) avec des modèles prédictifs ? Comment croiser les données de différents marchés pour détecter des abus de plus grande ampleur ?

L'Autorité de régulation des activités ferroviaires et routières (Arafer)

Nicolas Quinones-Gil
Responsable du département des études et de l'observation des marchés

Dans le cadre de la loi « Macron » du 6 août 2015 qui lui confère des pouvoirs de collecte avec une possibilité de sanction, l'Arafer a pris un nouveau tournant en matière de régulation par la data. C'est en octobre 2015 (avec la création de l'ARAFER) qu'a été créé le département des études et de l'observation des marchés, lequel est chargé d'implémenter les dispositifs de collecte et poursuit deux objectifs-clés :

- en interne, le passage d'une vision « macro » liée aux problématiques d'accès au réseau (marché amont) à une vision « micro » et intermodale en intégrant l'analyse des données du marché aval (trafics à la maille de l'Origine/Destination, comportement des utilisateurs finaux) ;
- en externe, la diffusion d'informations fiables et régulières via un Observatoire des marchés.

Cette mission a conduit à plusieurs décisions motivées de collecte régulière de données, dans tous les secteurs (conventionné ou non, routier, ferroviaire). Par rapport aux autres régulateurs sectoriels, l'Arafer en est encore au stade de la mise en place des dispositifs de collecte, dans un secteur des transports qui est moins mature en termes de transmissions régulières d'informations, avec des réticences à lever plus ou moins fortes en fonction des marchés et des acteurs. Les données collectées sont essentiellement quantitatives (offre, demande, résultats économiques).

Une unité spécifique de la direction du transport de voyageurs et des autoroutes est dédiée au contrôle des marchés publics passés par les sociétés concessionnaires d'autoroutes, ce qui va l'amener à traiter des gros volumes de données quantitatives et qualitatives (possibilités éventuelles d'analyses « Big Data » dans le futur).

Enfin, outre les décisions de collecte visant les secteurs qu'elle régule, l'Arafer conduit des enquêtes sur les pratiques de mobilité des utilisateurs finaux prenant en compte tous les modes de transport.

La Commission nationale de l'informatique et des libertés (Cnil)

Richard Montbeyre
Chef du service des contrôles

Le Big Data pose de nombreux défis à l'activité de contrôle de la Cnil - qui n'est pas compétente si les data ne permettent pas d'identifier les individus. Et pour cause, première autorité administrative indépendante créée par la loi, en 1978, la Cnil a une mission de protection des données personnelles et de la vie privée. Elle compte 200 agents, dont 23 au service des contrôles (16 contrôleurs, 8 juristes et 8 auditeurs des systèmes d'information). En 2015, elle a reçu 8 000 plaintes et procédé à 510 contrôles, 90 mises en demeure et 10 procédures de sanction.

La Cnil peut effectuer des contrôles suivant quatre modalités : sur place, sur pièces (y compris des bases de données comme les fichiers clients ou RH), sur audition et en ligne (sites Internet, fonctionnement d'applications). Par ailleurs, aucun contrôle ne se déclenche sur la base d'analyse de type Big Data. Les sources de contrôle sont de trois ordres : le programme annuel, les plaintes et l'initiative propre de la Cnil (via les alertes dans la presse, par exemple). Par ailleurs, même si la Cnil a un pouvoir de copie de données, celles-ci sont exclusivement utilisées dans le cadre des procédures concernées, sans mutualisation possible des documents recueillis. Il convient de noter, à cet égard, qu'il s'agit de données de procédure et non de données de surveillance.

Les principaux défis portent sur la réception des pièces (les plus volumineuses ne passant pas par mail, un projet de plateforme sécurisée de collecte est en cours), la conservation sécurisée et intègre des pièces (calcul d'empreinte numérique) et l'exploitation des données. L'analyse des zones « commentaires », par exemple, via un outil de reconnaissance de mots-clés dans une base de données fait remonter des centaines voire des milliers de lignes : il s'agit alors davantage de Medium Data que de Big Data. Et face à des textes de millions de lignes, il est très compliqué d'avoir une analyse automatisée : « arabe » est légitime lorsqu'il s'agit d'un cours d'arabe littéraire, mais beaucoup moins dans un autre contexte. L'analyse humaine reste indispensable. Finalement, la Cnil touche parfois du Big Data, mais toujours de l'extérieur et en spectateur. Aussi développe-t-elle surtout des outils qui lui permettent de réceptionner des bases et de les analyser en garantissant leur intégrité.

Enfin, le règlement européen applicable en mai 2018 imposera à la Cnil de travailler en coordination voire de manière intégrée avec d'autres autorités. Dès lors, les difficultés actuelles à recevoir une seule base de données en provenance d'un seul organisme deviendront critiques. D'où l'importance de continuer à s'outiller davantage.

L'Autorité de la concurrence

Etienne Pfister
Chef du service économique

Contrairement aux autorités de régulation sectorielle, l'Autorité de la concurrence n'a pas pour mission de réguler de manière ex ante les conditions de marché d'un secteur particulier. Dès lors, compte tenu de la diversité des champs d'investigation de l'Autorité de la concurrence, la collecte massive et automatisée de données, qu'il s'agisse de prix, de marge ou de toute autre variable économique, n'est pas nécessaire et cela exigerait d'ailleurs des outils de collecte et de traitement extrêmement puissants.

Il peut cependant être relevé qu'avec la loi Macron, l'Autorité de la concurrence s'est vu confier une mission de conseil du ministre de l'Economie et de la Chancellerie sur les tarifs et le nombre de professionnels dans les secteurs des professions réglementées du droit. Pour y procéder, des informations comptables, classiques mais également analytiques, sont indispensables. Aussi des processus de remontée d'informations sont mis en place. A cet égard, cette mission se rapproche de celles des autorités sectorielles.

S'agissant des autres missions de l'Autorité de la concurrence, la détection de cartels à partir de données de prix est une tâche délicate. Les données de prix et plus largement les éléments de preuve économique suffisant rarement à établir l'existence d'un cartel. En revanche, ces informations sont utiles pour, par exemple, justifier d'enquêter sur un secteur donné. En tout état de cause, fréquemment, la donnée analysée reste relativement « traditionnelle ». Même s'il est désormais possible de passer par Internet, il ne s'agit pas encore vraiment de Big Data.

En revanche, le développement d'Internet entraîne une montée en puissance et un enrichissement de certains dossiers contenant un volume de données de plus en plus important. Longtemps, il s'est agi par exemple des données des cartes de fidélité des distributeurs, utilisées notamment pour délimiter la zone de chalandise d'un magasin. Désormais apparaissent de plus en plus fréquemment des données de prix de vente recueillies sur Internet au travers de robots, qui, par leur nombre et l'automatisation de la collecte, se rapprochent du Big Data.

Réguler les acteurs d'un marché par la data : enjeux et perspectives

Henri Isaac
Vice-Président « Transformation numérique », Université Paris-Dauphine

La numérisation que l'on connaît depuis 25 ans correspond à une extension de la mise en données du monde (débutée avec la mathématisation). Le phénomène de datafication est largement étudié, d'autant qu'il s'étend désormais au monde physique, aux objets (objets connectés), au vivant (les puces RFID sur les animaux, capteurs sur les sols) et à l'environnement (au-delà des transactions économiques).

L'économie de la trace

De l'économie de l'information puis de la captation, nous serions ainsi passés à une « économie de la trace » (Kessous, 2011) - traces individuelles et collectives laissées lors d'un usage et produites en continu, tandis que les transactions économiques s'inscrivaient dans un modèle discret. Cette explosion des traces est au cœur des modèles de plateforme, qui reposent sur trois piliers de construction de valeur : réseau, infrastructure, données (capture des traces et transformation en monnaie).

Par ailleurs, le numérique permet de générer des données de nature très différente : les données personnelles (dont le régime juridique tend à étendre la définition, incluant les *shadow* data ou traces), les données contextuelles (heure et lieu d'une transaction, par exemple), les données collaboratives (interactions sociales sur les réseaux numériques), les données automatiques (générées par des dispositifs techniques, dans les voitures récentes par exemple) et les données d'environnement (données du vivant et capteurs). En général, ces données sont agrégées et traitées par des acteurs économiques. D'autres données peuvent s'y ajouter. Sur le marché publicitaire ou digital, par exemple, il existe des Second Party Data (apportées par l'éditeur), des Third Party Data (apportées par des Data Brokers qui commercialisent des données de tiers comme Carrefour ou Amazon, qui revendent des fichiers de données permettant de cibler des segments publicitaires) et des données ouvertes (Open Data).

Toutefois, tant qu'elle reste brute, la valeur d'une donnée est nulle ou quasi nulle. En effet, sans les méta-données (données qui décrivent les données, comme l'heure et les données GPS d'une photographie par exemple), les données ne peuvent souvent pas vraiment être traitées.

Une nouvelle ingénierie de la valeur

C'est de la combinaison Data + Meta Data que naît la valeur. Qui plus est, la valeur de la donnée est liée à la capacité algorithmique déployée, et à la capacité à la restituer de façon simple et instantanée à l'utilisateur final. Ces questions de représentation de la donnée dans la restitution sont centrales dans l'économie de l'attention dans laquelle nous sommes entrés. La chaîne de valeur de la donnée nécessite donc de combiner de la donnée, des méta-données, des algorithmes et la médiation de la valeur créée par de la data-visualisation.

Dans le Big Data, les caractéristiques propres sont certes les trois V, mais la question déterminante est celle de l'algorithmique appliquée. En l'occurrence, deux ruptures sont à noter : le Machine Learning (par rapport à de la statistique classique, on ne spécifie pas un modèle avec des variables, mais ce sont les données qui définissent les variables par découverte) et le Deep Learning (intelligence artificielle, en multipliant les couches d'apprentissage pour arriver à des résultats plus fins).

Le Machine Learning consiste, à partir d'un jeu de données, à entraîner un algorithme jusqu'à ce que sa performance prescriptive soit satisfaisante. Et à chaque fois que l'on ajoute une donnée, on teste à nouveau l'algorithme. C'est très auto-réalisateur, finalement. Dans certains cas, la capacité de prédiction est très élevée. Plus l'on a de données, plus l'algorithme - donc le modèle - est fiable. Pour optimiser le taux de réservation Airbnb, par exemple, les algorithmes classiques de yield management ne sont pas adaptés. Airbnb a donc choisi de définir un modèle à partir des données de réservation constatées. Ce faisant, trois variables ont pu être déterminées - variables qui n'auraient jamais été retenues par un modèle statistique spécifié par un humain : la présence du wifi, la taille et la qualité des photographies de l'appartement et l'existence d'une salle de bains plutôt que d'une douche. Ce faisant, Airbnb est passé d'une business intelligence classique a du prédictif et du prescriptif. De la même façon, l'algorithme d'Uber est capable de prédire et de prescrire à un chauffeur où se placer pour augmenter ses chances d'effectuer une course, à partir de données non structurées et recueillies en temps réel.

Le Deep Learning, pour sa part, consiste à multiplier les couches d'apprentissage pour obtenir des résultats encore plus fiables.

Les nouveaux mécanismes d'exploitation de la donnée et création de valeur

Airbnb et Uber ont largement déployé des mécanismes de Machine Learning. En effet, dans la mesure où les capacités de production du service (chambres disponibles ou chauffeurs disponibles) sont dynamiques, le modèle de pricing d'Airnbn ne repose pas sur la maîtrise de la fixation du prix. Dans le modèle d'Uber, les capacités (chauffeurs) sont dynamiques mais le pricing est maîtrisé. Le modèle vise donc à minimiser le temps d'attente de l'utilisateur et à maximiser le taux de remplissage donc le chiffre d'affaires du chauffeur. Et pour ces deux acteurs, les données servent à valider les transactions et à créer un modèle de confiance, indispensable pour réguler la qualité de service et maximiser la réputation de l'entreprise. Et tous les algorithmes (matching, pricing) reposent sur le data mining et la data-visualisation, avec la mise à disposition d'une courbe de prix et d'une courbe de demande à disposition du loueur, chez Airbnb, via une interface très simple et dynamique permettant à tout un chacun de comprendre le mécanisme de pricing. Plusieurs études montrent que l'algorithme prédictif d'Uber parvient à augmenter les capacités de façon significative, en incitant les chauffeurs à rester sur cette plateforme afin de réduire les temps d'attente et les distances à parcourir. Aujourd'hui, cet algorithme parvient à prédire 74 % des destinations avant même qu'elles n'aient été déclarées.

Les effets sont intéressants en termes de transformation du marché également. Une étude conduite par Uber à Los Angeles montre que le temps d'attente maximum acceptable par un client est passé de huit à quatre minutes, donc que l'effet des

algorithmes sur la demande modifie la caractéristique de la demande. Et ce, uniquement à partir des données collectées (et pas par sondage).

Cela étant, ces modèles imposant de disposer d'un jeu de données suffisant avant de construire un algorithme de Machine Learning, ils génèrent une course à la taille des données, qui pose une question de concurrence pour le régulateur. L'ordre d'arrivée sur le marché est devenu un avantage déterminant : un nouvel entrant qui n'aurait pas de capacité algorithmique équivalente ne pourrait pas proposer une qualité de service équivalente. On peut toutefois nuancer cette analyse en considérant que ce phénomène s'apparente au final à un effet d'expérience automatisé qui se matérialise rapidement par les algorithmes auto-apprenants.

Le régulateur et la donnée

L'accès à des données sur de nouveaux marchés (transport urbain) est une véritable opportunité. Jusqu'ici, les données générées par les taxis n'étaient ni utilisées, ni accessibles. Désormais, il est possible de les étudier voire de les superviser en temps réel. Par ailleurs, dès lors que la transparence est accrue grâce à des données, il est possible de laisser la place à l'expérimentation règlementaire.

Plusieurs problématiques concurrentielles se posent, dans ce contexte. Le stock de données ne peut-il pas constituer une barrière à l'entrée ? L'ordre d'entrée sur le marché est-il décisif ? Quelle régulation possible : obliger à de l'Open Data et laquelle, le cas échéant ? Le plus compliqué est d'arriver à comprendre que les données en tant que telles ne sont pas très intéressantes - étant entendu que leur valeur dépend du lien et de leur exploitation par les algorithmes.

Par ailleurs, peut-on concevoir des périodes expérimentales ? Il est déjà possible d'accéder à une partie des données d'Uber ou d'Airbnb via leurs API. Le régulateur pourrait-il avoir accès à une partie ou toutes les données par des API plutôt que par un extranet ? Comment les autorités de régulation peuvent-elles avoir une capacité de Reverse Engineering (capacité à analyser les algorithmes pour savoir s'ils sont *fair* ou *unfair*) pour les algorithmes comme celui de *dynamic pricing* d'Uber. Mais se pose alors la question de la propriété intellectuelle. Une grande partie des algorithmes produits par ces entreprises sont disponibles en Open Source, sur Github, mais sans les données : cela montre bien que la valeur vient de la combinaison entre les données et l'algorithme.

Enfin, quel est le niveau d'anonymisation et d'agrégation des données que les opérateurs devront mettre à disposition (loi Lemaire) ? La ville de Boston, par exemple, a demandé à disposer de toutes les données d'Uber. Mais rien ne s'est passé, puisque cette ville n'a pas la capacité informatique de traiter ces algorithmes et ces données.

2nde table ronde : Débat

Pourquoi et comment les données deviennent-elles un enjeu de régulation ?

Jean-Yves Ollier - Les fichiers de clients sont devenus un enjeu de régulation déterminant pour la CRE lors de la disparition des tarifs réglementés pour les professionnels - impliquant autant la CRE ex ante que l'Autorité de la concurrence et quand le sujet est devenu contentieux, ainsi que la CNIL. La problématique de l'accès aux données recueillies par les opérateurs historiques dans le cadre de leur activité de service public tend à se confondre avec celle de l'accès aux facilités essentielles.

Etienne Pfister - Concernant les concentrations avec des enjeux relatifs aux données (Facebook/WhatsApp, par exemple), l'argument initial de la Commission, il y a quatre ans, était, pour simplifier, que les données étant partout, il n'était pas gênant que Facebook mette la main sur celles de WhatsApp. Enfin, se pose la question de l'ampleur de l'avantage que confèrent les données. Les opérateurs mettent fréquemment en avant le rendement marginal décroissant des données : passé un certain volume de données, les données supplémentaires recueillies ne seraient pas d'une grande utilité. Mais en réalité, ces différents enjeux ne peuvent être appréhendés qu'au cas par cas. Les plateformes qui collectent les données ne fonctionnent pas toutes de la même manière et leur utilisation des données n'est pas la même non plus. A ce jour, les cas et les plaintes concernant exclusivement la collecte ou l'utilisation de données sur un mode « Big Data » sont encore peu nombreux. Pour vérifier par elle-même que le Big Data ne créait pas de risques concurrentiels, l'Autorité a lancé une enquête sectorielle sur l'utilisation des données sur le marché de la publicité en ligne.

Joëlle Toledano - Si l'on manque de cas, c'est sans doute parce que l'incertitude profite aux acteurs. Est-ce véritablement un problème de régulateur ? En l'occurrence, les secteurs où la chaîne de valeur est modifiée et où les règles sont mises à mal ne sont pas régulés. Ce phénomène ne s'observe pas encore dans les secteurs régulés. Les acteurs du numérique indiquent qu'ils ont acquis leurs données par leur propre mérite et non du fait d'un monopole d'origine. Ce serait donc davantage un problème politique qu'un problème technique de régulation. A court terme, du moins, c'est une question de rapport de forces politiques.

Etienne Pfister - Certains cas sont en cours (comme celui de Google, par exemple) et il est trop tôt pour en tirer des enseignements. Par ailleurs, le problème est certes politique. Mais il est aussi juridique (quelle jurisprudence ?) et économique (quel effet sur les incitations à l'investissement ?). De manière générale, il convient de ne pas généraliser les difficultés éventuelles qui pourraient être rencontrées dans certains secteurs à d'autres. Une régulation des plateformes est fréquemment évoquée, mais ces dernières doivent nécessairement être analysées au cas par cas car leur fonctionnement et la concurrence entre elles varient selon les secteurs.

Quid des données du régulateur, c'est-à-dire de la régulation par la donnée ? Comment articuler une approche Open Data et les exigences de sécurité extrême qui peuvent s'appliquer à certains types de données manipulées par ce dernier ?

Jean-Yves Ollier - Dans nos activités de surveillance, certaines données sont statistiques, destinées à alimenter les observatoires, avec un accès naturel en Open Data, tandis que d'autres (données relatives aux transactions sur les marchés) restent naturellement couvertes par le secret en matière industrielle et commerciale avec des modalités de protection très strictes. Mais il peut exister des types de données dont le statut est moins évident au regard de ces critères.

Eric Brousseau - Un autre enjeu n'est-il pas celui du délai d'accès aux données, au-delà du degré de finesse de celles-ci ?

Richard Montbeyre - Nous sommes allés assez loin dans le passage à l'Open Data, notamment pour les contrôles (avec une recherche possible par thème, par année, par lieu et par organisme). Jusqu'ici, ces données étaient annexées aux rapports annuels, mais elles sont désormais totalement disponibles.

Stéphane Lhermitte - L'Arcep publiera des cartes de couverture en Open Data, conformément à la loi Lemaire. Des modèles numériques permettront de zoomer au niveau le plus fin ou de superposer différentes cartes.

Joëlle Toledano - Ce qui est communiqué est-il si « subversif » qu'on le craint ?

Richard Montbeyre - Je n'ai pas vraiment la vision des usages qui ont pu être faits des informations en Open Data (certaines ne présentant pas de véritable intérêt). Cette ouverture est assez récente, qui plus est. En ce moment, nous sommes un peu dans « l'ouverture pour de l'ouverture », sans nécessairement que cela entraîne un usage.

Eric Brousseau - Il existe un problème de masse critique à atteindre avant qu'un organisme de recherche se saisisse des données disponibles.

Stéphane Lhermitte - L'Arcep est très réticente à fournir des données de mesure individuelle. En outre, les données publiées en Open Data sont toujours travaillées et jamais brutes. Nous pouvons aussi contraindre les opérateurs à publier eux-mêmes certaines données.

Jean-Yves Ollier - Finalement, le sujet de la massification ou de la digitalisation des données est très présent en termes de collecte (évolution des activités et des supports, nécessité d'appréhender certaines activités sur le marché de détail), mais des limites apparaissent en termes d'analyse. Néanmoins, les outils d'analyse systématique et les algorithmes peuvent fournir des éléments de filtrage ou d'aide à la décision. Nous ne sommes encore qu'aux prémices de cette évolution, à l'exception de l'Arjel.

Joëlle Toledano - Quid des données fausses ?

Nicolas Boulanger - Elles sont transmises involontairement, par dysfonctionnement des logiciels. Nous ne détectons pas de données volontairement fausses. Par ailleurs, les outils de récupération des données sont audités.

Matthieu Morin - L'article 15 du règlement REMIT impose aux opérateurs de marché fournisseurs de données d'avoir une activité de surveillance de leurs propres données, activité que nous pouvons auditer.

Jean-Yves Ollier - Avec ces données de transaction brutes, le risque est limité. Dans d'autres domaines, nous recevons parfois des données assez approximatives. Mais nous avons la possibilité de faire contrôler des données aux frais des opérateurs.

Chaire Gouvernance et Régulation
Fondation Paris-Dauphine
Place du Maréchal de Lattre de Tassigny - 75116 Paris (France)
http://chairgovreg.fondation-dauphine.fr