Internet et les données à caractère personnel : traitement, enjeux et gouvernance

Julien Pierre

Internet et les données à caractère personnel : traitement, enjeux et gouvernance

Le concept de *Big Society* faisait craindre que son avènement soit accompagné de celui de la *Big Machine*, longtemps incarnée – en informatique en tout cas – par IBM. Même si la première est avérée aujourd'hui sous le nom de mondialisation, la seconde par contre semble plus difficile à cerner : toutefois la commercialisation d'une structure comme le *cloud computing* pourrait s'en rapprocher. Proposant d'externaliser à la fois les services (webOS, bureautique en ligne, eCRM, etc.), les processus (facturation, impression, sauvegarde) et les contenus (la plateforme de partage de photos Flickr est en le parfait exemple), l'informatique dans les nuages (ou infonuagique) semble une solution appréciée autant par les entreprises que par les particuliers. Pour autant, son appellation même ne doit pas faire oublier qu'en termes d'abstraction, cela confine parfois au nébuleux : ainsi, les avantages de mobilité et d'accessibilité laissent dans l'ombre de nombreux inconvénients, principalement éthiques : impact environnemental tout d'abord, mais aussi sécurité des données conservées dans le nuage. Nous ne traiterons pas ici des données de gestion ou de statistiques (stock, bilan social, comptabilité, résultats économiques) mais exclusivement des données à caractère personnel (DCP), qui plus est collectées en ligne, à travers les réseaux socionumériques (RSN).

Qu'est-ce qu'une donnée à caractère personnel ? Plusieurs textes de lois s'accordent à définir cette notion[1] : pour la Convention 108, il s'agit de « toute information concernant une personne physique identifiée ou identifiable ». La loi Informatique et Libertés est un peu plus précise puisque – article 2 – elle définit une DCP comme « toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. » C'est pourquoi on parle aussi de données nominatives (*versus* données anonymisées) ou, aux États-Unis, de PII (*Personal Identifiable Information*) Ces éléments sont généralement l'identité civile, les informations génétiques et médicales, les coordonnées des résidences, les immatriculations de véhicule, les données de connexion et n'importe quelle donnée (y compris l'adresse IP[2]) qui permettrait d'identifier – même indirectement – une personne (que nous nommerons dorénavant titulaire des données à caractère personnel). Même anonymisées, des données corrélées permettent de savoir qui est qui (ex. : en août 2006, AOL laisse échapper des données anonymes mais dont le recoupement de requêtes et de localisation permet d'identifier précisément des individus[3]). Où se nichent aujourd'hui les données à

---

[1] Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Convention européenne n°108 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, http://conventions.coe.int/Treaty/FR/Treaties/Html/108.htm, 1981.

[2] Il y a débat pour savoir s'il faut inclure cette information dans le champ des DCP. En effet, la Cour d'appel de Paris a rendu deux arrêts excluant l'IP (CA Paris, 27 avril 2007 et 15 mai 2007), cependant la tendance transnationale penche pour l'inclusion (CNIL, loi Informatique et libertés, le Groupe de travail de l'article 29 – équivalent européen de la CNIL – et l'ensemble des autorités de protection des données des Etats membres de l'Union européenne).

[3] Voir l'article du New-York Times dévoilant l'identité de l'abonné n°4417749, http://www.nytimes.com/2006/08/09/technology/09aol.html. Ces données sont encore accessibles à l'adresse suivante : http://www.gregsadetsky.com/aol-data/

caractère personnel ? Est-ce que toutes les données – y compris les descripteurs de notre vie sociale et affective – peuvent être considérées comme personnelles ?

Nous proposons d'étudier comment s'opère le traitement des DCP entre un niveau macro, ou industriel – avec les entreprises dont la rentabilité est assurée par le traitement des DCP, un niveau méso – avec l'interface par laquelle sont échangées les DCP, autrement dit le navigateur web, et un niveau micro – chez l'utilisateur et titulaire de DCP. Enfin, nous verrons que, face à ces dispositifs de ciblage comportemental en ligne (*Online Behavioral Advertising*) semble se constituer une opposition avec l'institutionnalisation et l'industrialisation de la mouvance Do Not Track Us (« "ne nous traquez pas »).

I, Des processus convergents vers le microsocial

Les SIC sont bicéphales, relevant pour certains de l'hybridation, pour d'autres de la confluence entre les champs de l'information et de la communication. Pour articuler le lien entre les deux, nous prendrons la définition qu'en donne Jean Meyriat : « La communication est un processus dont l'information est le contenu ; l'une ne peut être comprise sans l'autre.» [4] Il n'y a pas lieu ici de discuter la validité de cette proposition, elle a pour nous le mérite de contenir et conduire à deux notions que nous voulons rassembler pour construire notre approche : plus exactement, nous proposons d'observer à quels processus communicationnels participent les données à caractère personnel.

I,1. La (re)documentarisation

Lors de la constitution d'un fonds documentaire, chaque élément (article, revue, livre, etc.) se voit signaler par des informations relatives à son contenu (mots-clés, résumés), ses droits (propriété intellectuelle) et son classement afin d'en faciliter l'accès au public (cotation, URL). Ce processus de documentarisation (Pédauque, 2006) est à son tour complété par les informations que les lecteurs estiment devoir associer au document pour leur usage (commentaires et annotations, nouveaux mots-clés ou informations de signalement, etc.). Un exemple de cette redocumentarisation (Pédauque, 2007) se trouve dans l'indexation sociale, ou folksonomies. « Redocumentariser, c'est documentariser à nouveau un document ou une collection en permettant à un bénéficiaire de réarticuler les contenus sémiotiques selon son interprétation et ses usages » (Zacklad, 2007). L'un des corpus documentaires le plus soumis à ces processus serait l'humain, nous dit O. Ertzscheid, faisant de « l'homme un document comme les autres » (Ertzscheid, 2009). Ainsi, l'internaute principalement, mais aussi chaque consommateur, chaque citoyen d'une nation pourrait être considéré comme un document, car des informations lui seraient sans cesse accolées (livret de famille, *curriculum vitae*, profil Facebook, fichiers administratifs ou commerciaux, etc.).

Si l'on regarde historiquement ce qu'il en a été des données à caractère personnel, on pourrait estimer que les photographies anthropométriques et les premières fiches de police relèvent de la documentarisation. Les précurseurs de l'anthropologie et de la statistique judiciaire vont adopter comme critères la morphologie crânienne, l'origine ethnique, puis sociale pour établir une prédisposition au crime et à l'insurrection (Mattelart, 2007, pp.15-41).

---

[4] Meyriat J., Entretiens avec les fondateurs de la SFSIC, SFSIC, reprographié, Paris, 1993, 16 p.

On sait d'ailleurs que les premiers systèmes automatiques de classement de la population ont été ré-exploités lors de génocides (Seltzer et Anderson, 2001). Les logiciels modernes de reconnaissance comportementale embarqués dans les caméras de surveillance signalent aux forces de l'ordre tout individu dont les agissements correspondent à une interprétation de ce que serait un *habitus* criminel. Tous les agents ont choisi arbitrairement des critères opérants selon la finalité qui était la leur (ici, la sécurité intérieure). Partant de là, nous pourrions considérer *primo*, que l'agent – concepteur et utilisateur du dispositif de documentarisation – est un bénéficiaire au même titre que le titulaire ; *deuxio*, que l'interprétation a lieu dès la conception du dispositif de documentarisation, faisant d'elle un processus « structurant-structuré ». C'est parce que l'on ne peut se débarrasser de ce biais que nous proposons de distinguer, au sein de la (re)documentarisation, entre un processus de saisie endogène, où le titulaire est l'agent de sa propre (re)documentarisation, et un processus exogène, où c'est un tiers qui s'en charge (ami ou proche, fonctionnaire ou commerçant, humain ou automate). Se pose alors la question de la performativité de ces informations, et notamment celles relevant du premier processus où la subjectivité et la théâtralisation de soi peuvent conduire à un profil tout à fait trivial, faussant *de facto* l'analyse mercatique ; il n'empêche que les données exogènes ne sont pas non plus exemptes de mésinterprétation (comme avec les algorithmes de reconnaissance comportementale embarqués dans les dispositifs de vidéosurveillance).

Mais de quelle performance parle-ton ? Et en quoi la trivialité du profil peut-elle être négative ?

I,2. L'industrialisation des DCP

Dans le champ de la communication, nous empruntons à Bernard Miège la définition qu'il donne de l'informationnalisation : « il s'agit d'un procès ou d'une logique sociale de la communication qui se caractérise par la circulation croissante et accélérée des flux d'information éditée ou non, autant dans la sphère privative, dans celle du travail que dans l'espace public » (Miège, 2007, p.66). Ainsi, les données préalablement formatées lors de la (re)documentarisation sont mises en mouvement : il y a dans un premier temps élargissement de l'offre informationnelle, dans le sens où tout peut devenir information (le savoir comme les affects) ; dans un second temps, il y a médiatisation de la communication dans la mesure où la circulation des flux est opérée par une industrie, inscrivant dès lors l'information à la fois dans un processus de production (on reboucle ici sur la documentarisation) et sur une logique de rentabilité économique (formatant les contenus en un consensus sensé correspondre aux perceptions que les agents économiques se font du marché). Et dans un dernier temps, il y a transnationalisation des activités info-communicationnelles, les internautes français exploitant des services web et des infrastructures principalement originaires des États-Unis, et soumis au droit de cette juridiction ; Facebook par exemple stocke ses données sur des serveurs en Californie, pourtant son siège social se trouve en Irlande, donc soumis au droit européen. Quelle juridiction est compétente en cas de plainte d'un utilisateur ?

Les données à caractère personnel peuvent être considérées comme un contenu équivalent à ceux opérés par les industries culturelles. Les logiques informationnelles – on l'a vu – et les logiques marchandes – on va le voir – sont les mêmes. « Dans les industries de la culture et de l'information, note Franck Rebillard, la valorisation des biens présente un

caractère aléatoire, beaucoup plus fortement que dans d'autres secteurs industriels. Le devenir économique d'un bien informationnel ou culturel (…) est difficile à anticiper. Le marketing aval (…) permet de limiter cet aléa, mais le marketing amont – études de marché permettant d'identifier les attentes de la demande – est en revanche largement inefficient. » (Rebillard, 2007, p.32). Or la réduction de cet aléa pourrait se résorber par l'exploitation des données à caractère personnel, c'est en tout cas ce que revendiquent les auteurs de l'École de Chicago (R. Posner, G. Stigler) : la protection de la vie privée provoque une asymétrie informationnelle au profit du titulaire des DCP et au détriment de l'agent économique qui cherche à les exploiter. Cette rétention (socialement instituée ou individuellement actée) produirait ce que les économistes appellent des externalités négatives. Les entreprises devraient conséquemment supporter soit le coût d'une collecte contraire aux lois ou aux mœurs (amendes, diminution de la réputation) ; soit le risque d'une production inadéquate avec les attentes du marché (Rochelandet, 2010). De plus, la doctrine de l'École de Chicago s'inscrit aussi dans une éthique de l'humain, résumée ainsi par Posner : « Pourquoi quelqu'un voudrait-il dissimuler un fait, si ce n'est pour induire en erreur les autres afin de faire des transactions avec eux ? » (Posner, 1981 ; p. 408) ; et actualisée par Eric Schmidt, quand il était CEO de Google : « Si vous voulez que personne ne le sache, c'est peut-être déjà que vous n'auriez pas dû le faire »[5]. La protection de la vie privée serait ainsi sujette à suspicion, et devrait être levée par le dévoilement des DCP. Cette injonction de transparence, qui touche déjà la gouvernance de la chose publique (libération des données publiques et sousveillance) donne aussi lieu, en ce qui concerne la sphère privative, à toute une industrie des données et de la réputation. Les plus grands groupes publicitaires, les groupes de télécommunication, les entreprises dominantes sur Internet disposent tous de filières verticales ou de partenariats dans le traitement des données : construction et hébergement de serveurs, déploiement et gérance de réseaux filaires, équipes de statisticiens, d'ingénieurs informatiques spécialisés dans le requêtage, la certification de données, l'agrégation de données, la production d'outils de publication assistée incitant à la saisie endogène ; il existe également des structures commerciales dans l'achat ou l'échange d'espaces publicitaires, dans la diffusion sectorielle et le ciblage d'audience (du fait de l'identification des niches), dans la relation publique et le lobbying, etc[6]. De fait, ces entreprises s'intercalent entre des annonceurs et des internautes. Dans le lexique de la stratégie mercatique, avant qu'une entité (individu ou entreprise) devienne cliente, elle se nomme prospect (auprès de qui l'on pourra diffuser une publicité ciblée), mais avant cela encore, elle se nomme suspect : dans la logique commerciale (et encore plus dans la logique sécuritaire), tout le monde est suspect, c'est pourquoi en ligne, la très grande majorité des sites web embarque des dispositifs surveillant la consultation par les internautes, et cela à des fins de prospection commerciale. Dans l'absolu, si les données personnelles n'étaient pas protégées, tout le travail de fouille et de requêtage de données pourrait inéluctablement engendrer un acte d'achat. Il y a loin encore pour que cela soit systématique.

[5] Interview accordé à la chaine CNBC le 03/12/2009.

[6] Un panorama (non exhaustif) de ce complexe info-industriel a été réalisé par par Terence Kawaja, créateur du fonds d'investissement LUMA Partners LLC, en 2010. Il est accessible à l'adresse : http://www.adexchanger.com/pdf/Display-Advertising-Technology-Landscape-2010-05-03.pdf.

I.3, L'informatisation

Comme le signalait T. Lamarche, les SIC sont disposées à occuper une posture scientifique critique à l'égard du procès d'informatisation de la société (Lamarche, 2005). L'auteur relève que, d'une informatique lourde et centralisée, opérée dans les grandes organisations (firmes et États au cours des années 1970), nous sommes passés à une micro-informatique domestique (1984, McIntosh), puis en réseau (1991, World Wide Web), et dorénavant mobile (3G, wifi, etc.). Nous pourrions parler même d'un procès de micro-informatisation de la société : notre hypothèse est de renverser cette proposition et de parler d'une informatisation du microsocial. Ce faisant, l'informationnalisation aurait pour objet ce qui se déroule dans l'ordre de l'interaction, et les phénomènes affectifs seraient l'objet d'une documentarisation, la résultante répondant à l'acception large d'une donnée à caractère personnel. Son exploitation est déjà considérée comme stratégique pour les industries culturelles et mériterait dès lors d'être inscrite dans ce champ de recherche.

En informatique, mais également dans les sciences cognitives, l'identification est un processus par lequel un individu délivre son identité par l'entremise d'un identifiant (un code, un geste, un son ; bref une DCP). L'authentification est le processus suivant par lequel le bénéficiaire reconnait le titulaire en comparant son identifiant à un catalogue : cela marche pour un mot de passe comme avec la voix ou la démarche de nos amis. Si l'authentification réussit, le bénéficiaire autorise le titulaire à accéder aux ressources qu'il détient, ou au service qu'il propose. C'est pourquoi l'identification, et derrière elle l'identité et les données à caractère personnel, peuvent être considérées comme une condition d'existence dans l'espace commun, et comme une condition d'accès à l'espace public (pour la différence entre les deux espaces : Tassin, 1992, pp.24-25 ; pour les conditions d'entrée : Dodier, 1999, p. 109)[7].

Le projet d'autonomie individuelle, qui soutient l'existence de et dans l'espace public, conduit à la multiplicité des faces exigée par les rôles à tenir dans les différents cercles sociaux que nous fréquentons et qui composent l'espace commun. Ce jeu avec l'identité, ce *facework* (Goffman, 1973) est acté aujourd'hui dans les réseaux socionumériques.

Par une analyse sémiopragmatique (Georges, 2010), on peut découvrir les stratégies de présences en ligne orchestrées par les internautes, mais en réalité guidée par les formulaires et l'ensemble du dispositif d'écran mis en place par les entreprises du web social. Nous redonnons ici brièvement les concepts proposés par F. Georges :

[7] Dans le cadre de cet article, nous ne pouvons développer plus en détail les liens entre identité et espace public. Cette articulation sera présente dans la thèse de doctorat que prépare l'auteur (soutenance été 2012). Voir aussi Pierre J., « Génétique de l'identité numérique », Les Cahiers du numérique, n°59, à paraître (début 2012).

- Identité déclarative : le sujet utilise un ligateur (autonyme) et un ensemble de signes (qualifiants, sociatifs et possessifs) pour se représenter via des attributs[8] ;
- Identité agissante : le système affiche l'activité du sujet (en ligne, en relation avec X, en place IRL)[9] ;
- Identité calculée : le système affiche un score de cette activité (nombre d'amis, d'endroits visités, de contenus consultés ou proposés à la communauté)[10].

« Ainsi, le cycle de vie d'une information sur une page de profil commence par une notification agissante (l'action s'est produite), puis simultanément elle fait l'objet d'un stockage dans la zone déclarative et est comptabilisée numériquement » (Georges, 2010, p.194). Mais l'on peut se demander alors si l'identité est performative ? Si tout discours est descriptif, et si l'identité comme signe est un discours sur le Moi, alors tous les signes projetés – y compris les trois items précédents provenant tous d'une saisie endogène – sur les réseaux socionumériques relèvent du descriptif. En quoi cette identité est-elle un acte de langage ? À quelle performance peut-on dès lors s'attendre quand elle est traitée par un système informatique ?

II, Les technologies de commensuration

II.1, La commensuration

Nous proposons de rapprocher cette performance numérique du processus cognitif de commensuration qui « implique l'utilisation de nombres pour créer des relations entre des choses. La commensuration transforme des différences qualitatives en différences quantitatives, la différence s'exprimant précisément en termes de grandeur par rapport à un paramètre commun » (Espeland, 2002). Pour la sociologue, cette quantification objective les émotions et, dans le cadre d'une psychologisation accrue, aurait des vertus thérapeutiques. Or ce processus est l'avatar de la rationalisation appliquée à la vie affective. C'est dans ce cadre que nous considérons les données à caractère personnel comme résultant d'une documentarisation et produisant une commensuration actée en ligne.

Ces marqueurs quantitatifs font que l'identité est intrinsèquement numérique. Non pas tant qu'elle est réduite par numérisation à un code binaire (on parlerait dans ce cas d'identité numérisée), mais parce que l'identité renvoie *in extenso* à sa dimension mathématique : de l'égalité contenue dans le principe d'identité à l'anthropométrie et la biométrie, l'identité d'un individu s'exprime par des valeurs ou des opérateurs algébriques. Pour les premières, on peut citer les numéros de sécurité sociale, de passeport ou de permis de conduire, les numéros de compte et de téléphonie, les adresses IP, les mensurations, les scores ; pour les seconds, relevant de l'algèbre relationnelle des bases de données (Codd, 1970), on peut citer les requêtes de sélection, les opérateurs booléens, les jointures, etc.

---

[8] Sexe, date de naissance, intéressé par, situation amoureuse, opinion politique et religieuse, orientation sexuelle, informations personnelles.

[9] Mise à jour de profil, demande d'amis, participation à un événement ou à un groupe, création d'événement ou de groupe, a commenté ou tagué ou envoyé un cadeau, a envoyé un billet collectif, a été tagué par un ami, a utilisé une application.

[10] Nombre d'amis, nombre de groupes, nombre d'événements visibles dans le mini-historique, nombre d'événements par connexion, taux de présence.

Concrètement, la très grande majorité des services en ligne requiert l'inscription et l'identification des internautes (Google, Facebook, mais aussi d'autres organisations se proposent comme tiers de confiance dans ce processus) ; la très grande majorité procède à l'enregistrement de leurs parcours et de leurs habitudes de navigation (notamment via l'outil Google Analytics) ; les réseaux sociaux établissent des jointures pour savoir qui est l'ami de qui ; les sites de rencontre font de même pour proposer qui pourrait être l'amant de qui ; et les sites de commerce réalisent des statistiques sur le panier d'achat, leur permettant de proposer aux internautes des publicités ciblées (*Online Behavioural Advertising*, OBA, publicité comportementale en ligne).

Deux dispositifs techniques rendent possible cette pratique de traitement des DCP : les bases de données et les cookies[11].

II.2, Les cookies d'Heidi

Nous avons mené l'expérience de naviguer en ligne avec une machine vierge, et d'étudier les cookies consécutifs à cette navigation[12]. Sachant les griefs portés à l'encontre de Facebook, nous avons créé un faux profil, et jouer avec les paramètres de confidentialité et les publicités[13]. Puis nous avons compté les cookies, et regarder l'information qu'ils portaient. Avant d'en venir aux résultats, il faut savoir comment fonctionne un cookie : il s'agit d'un fichier texte déposé par le serveur du site web sur le client de navigation (InternetExplorer, Firefox, Chrome, etc. : c'est-à-dire sur l'ordinateur de l'internaute). Un cookie contient un certain nombre d'informations, voici le cookie d'identification d'Heidi sur Facebook[14] :

- `Domaine: .facebook.com`
- `Nom: c_user`
- `Contenu: 100001982354948`
- `HTTP Only: TRUE`
- `Path: /`
- `Date d'expiration: 1298987168 (en temps Unix, autrement 22/06/2011)`

On voit que l'information contenue dans ce fichier est minime, mais corrélée aux autres cookies du domaine, il est possible de savoir quand Heidi s'est connectée et quelle a été la durée de cette connexion (cookie de session, à la fin de laquelle est écrite une nouvelle date d'expiration du cookie d'identification). À partir de cet objet documentaire *a priori* simple, deux usages peuvent en être fait : le cookie peut être embarqué (encapsulé) sur un autre domaine (cookie tierce partie). Ainsi, le cookie de Facebook peut se retrouver encapsulé sur

---

[11] Une précision : nous nous focalisons ici sur les cookies, et les cookies tierce partie, mais il existe pléthore d'autres dispositifs de traçage en ligne (evercookie, flash cookie).

[12] Expérience réalisée à partir du 17 janvier 2011 en créant un nouveau profil utilisateur sur Windows Vista et Firefox 4.0. Nous avons utilisé l'extension Export Cookie pour consulter les cookies.

[13] Heidi Numsberger est un personnage de fiction que nous avons scénarisé pour les besoins de la mise en relation : c'est une jeune alsacienne en quête d'un nouveau travail en agence de communication. Elle a quelques amis sur Facebook et Twitter.

[14] L'hôte désigne le domaine qui a écrit, et de facto qui a le droit exclusif de lire le cookie. Des sites web sur d'autres domaines n'y ont pas accès. `c_user` est le nom du cookie (nous avons trouvé jusqu'à 13 cookies différents provenant de Facebook). Le numéro `100001982354948` correspond à l'identifiant unique d'Heidi. La date d'expiration désigne quand ce fichier ne sera plus autorisé à la lecture (en général, Facebook accorde soit 1 mois, soit 2 ans à ses cookies ; nous en avons trouvé d'autres domaines qui expiraient en 2078).

d'autres sites (via le bouton J'aime[15]), soit le domaine de Facebook peut placer sur la machine de l'internaute des cookies d'applications-tierces (publicités, jeux, quizz, etc.). En conséquence, il est possible de se retrouver sur un autre site où le profil de l'internaute serait déjà alimenté par des données saisies et collectées via Facebook : la photo du profil par exemple. Or il s'avère, même en consacrant toute une session de navigation à la consultation de publicités, et après avoir générés plus de 500 cookies en 30 minutes de présence sur Facebook, que l'identifiant unique d'Heidi ne s'est retrouvé nulle part ailleurs ; il s'avère que les données d'Heidi se sont retrouvées sur d'autres sites uniquement par consentement (Facebook Connect[16]) ; enfin, aucune autre donnée d'Heidi n'a été repérée ailleurs, y compris au niveau du spam, y compris en saisissant le nom d'Heidi dans les moteurs de recherche de personne. La circonscription des données par Facebook est effective, et il semble que seul le consentement du titulaire génère l'exfiltration des DCP.

Le deuxième usage du cookie est à relier à l'autre dispositif technique annoncé précédemment : les bases de données[17]. Sans revenir sur les spécifications du modèle relationnel adopté pour modéliser les associations entre les différentes entités contenues dans une base de données (Codd, 1970), nous tenons simplement à rappeler que les différents enregistrements d'une table se distinguent par un identifiant unique : dans le cas de Facebook, comme de tous les sites à inscription, c'est la même valeur que celle contenue dans le cookie d'identification. Pour Heidi, c'est le `100001982354948`. Mark Zuckerberg est le n°4[18]. Cette combinaison de données relevant d'une documentarisation côté client (le cookie) et côté serveur (la base de données) produit ainsi un graphe social : « At Facebook's core is the social graph; people and the connections they have to everything they care about. The Graph API presents a simple, consistent view of the Facebook social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags). »[19]

Ainsi, il est possible de savoir comment les données personnelles sont collectées par les réseaux socionumériques ; il est possible de connaitre le premier niveau de traitement de

---

[15] Pour une explication du mécanisme de traçabilité induit par le bouton J'aime de Facebook, lire http://online.wsj.com/article/SB10001424052748704281504576329441432995616.html?mod=WSJ_Tech_LEFTTopNews

[16] Il s'agit d'un service proposé aux membres de Facebook pour s'authentifier sur d'autres sites web. C'est avec Facebook Connect qu'Heidi s'est créée un compte sur TripAdvisor, ce dernier associant à son profil la photo présente sur Facebook. Heidi a aussi utilisé ce service pour s'authentifier sur Twitter. C'est uniquement dans les cookies de ces deux sites que nous avons retrouvé trace de l'identifiant unique.

[17] Nous reviendrons sur cette logique de classification dans un article à paraitre : Solutions (techniques) pour une dissolution (sociale) : la modélisation de la vie affective dans les bases de données du web.

[18] Toute allusion à un quelconque n°6, et a fortiori à la possibilité qu'il existe un numéro 1, serait complètement fortuite. Il n'empêche que la tentation est grande de comparer Facebook au village du Prisonnier (série britannique des années 60, portée essentiellement par l'acteur Patrick McGoohan), village duquel on ne peut s'échapper qu'en criant « – Je ne suis pas un numéro ». Mais comme l'épilogue – et le générique – le laissent entendre : « – Who is number one ? – You are, number six ». Les membres du Village Facebook sont leurs propres tyrans.

[19] « Au cœur de Facebook se trouve le graphe social ; les gens et toutes les connections qu'ils ont avec tout ce qui leur tient à cœur. L'interface de programmation du graphe présente une vue simple et consistante du graphe social de Facebook, représentant uniformément les objets du graphe (comme les gens, les photos, les événements et les pages) et les connexions entre eux (relations amicales, contenu partagé, photos taguées) » (traduction libre).

ces données, mais en surface uniquement. Il faut interroger Facebook et les RSN comme une boîte noire, où sont seulement identifiables les flux entrants provenant de la saisie endogène, parfois redocumentarisés par les autres membres du graphe social. Mais quels résultats en sortie ? *Quid* des flux sortants ? Le dévoilement des internautes ne conduit-il qu'au spam ? Ce phénomène est-il avéré ? La base de données des intentions[20] est-elle achevée, et opérationnelle ? Comme pour Google organisant notre accès à la connaissance sur la base d'un algorithme que tout le monde ignore, faut-il s'inquiéter aussi de que nos interactions sociales soient archivées selon une formule secrète, et avec des finalités encore plus obscures ?

C'est sur ce dernier point qu'il faut être vigilant : depuis sa création, Facebook a modifié à de nombreuses reprises – six fois en cinq ans – ses conditions d'utilisation et sa politique de confidentialité (traduites en français depuis peu[21]). La société reste propriétaire de tout le contenu mis en ligne par ses membres, et peut l'exploiter ou le céder à des tiers selon ses besoins[22]. Cette notion semble s'ancrer chez les utilisateurs mais, dans un article précédent, nous notions une certaine fatalité par rapport à cette surveillance commerciale (Martin-Juchat et Pierre, 2011). Nous constatons aujourd'hui, à travers une nouvelle série d'entretiens, mais surtout au vu des acteurs socioéconomiques, des propositions de réappropriation des données à caractère personnel.

III, Modalités de gouvernance des DCP

III.1, Propositions d'acteurs économiques

Le navigateur est l'outil exclusif de consultation des pages web, et c'est par son entremise qu'à la fois les internautes ont accès à du contenu et que les régies publicitaires y placent leurs cookies. Les principaux éditeurs – et le premier d'entre eux, Microsoft avec son logiciel InternetExplorer, ont longtemps été accusés d'être les vecteurs du ciblage comportemental, souffrant de surcroit de défauts de conception qui permettaient à des tiers malintentionnés de placer des logiciels espions. La pression commerciale a effectivement accéléré les cycles de développement des logiciels, multipliant ainsi les failles de sécurité, les risques d'attaque logicielle et pour l'utilisateur final la probabilité de récupérer des collecteurs de données (pourriciel ou *malware*).

---

[20] Nous empruntons cette expression à l'un des gourous du web, John Battelle, entrepreneur, journaliste et intervenant à l'Université de Berkeley. L'idée générale est d'anticiper le comportement d'achat des internautes en croisant les traces de leurs consommations, y compris culturelles. L'article d'origine est consultable à cette adresse : http://battellemedia.com/archives/2003/11/the_database_of_intentions

[21] Pour les conditions d'utilisation, lire http://www.facebook.com/terms.php?ref=pf. Pour les règles de confidentialité, lire http://www.facebook.com/privacy/explanation.php, consultés le 01/05/2011.

[22] Art.2.1 des conditions d'utilisation : « Vous nous accordez une licence non-exclusive, transférable, sous-licenciable, sans redevance et mondiale pour l'utilisation des contenus de propriété intellectuelle que vous publiez sur Facebook ou en relation à Facebook ».
Art.10.1 : « Vous pouvez utiliser vos paramètres de confidentialité pour limiter la façon dont votre nom et votre photo de profil peuvent être associés à du contenu commercial, du contenu parrainé ou d'autres contenus (tels qu'une marque que vous indiquez aimer) que nous diffusons. Vous nous donnez la permission d'utiliser votre nom et votre photo de profil en association avec ce contenu, conformément aux limites que vous avez établies.
Art.10.2 : « Nous ne donnons pas votre contenu ou vos informations aux annonceurs sans votre accord. »

Face au *leadership* d'InternetExplorer, mais aussi dans le souci de respecter les recommandations techniques du W3C et de l'IETF[23], et peut-être aussi après avoir identifié les besoins des internautes, plusieurs logiciels ont intégré des dispositifs de navigation privée. L'une des premières propositions de navigation confidentielle – *traceless* – fut le Porn Mode du logiciel Safari (sur machine Apple) : en fait, il s'agissait simplement d'effacer l'historique en fin de consultation. Mais cette précaution ne visait qu'à protéger les usages au sein du domicile ou du lieu de travail : d'autres utilisateurs du logiciel ne pouvaient découvrir les sites précédemment consultés. Ce mode n'effaçait en rien l'échange de données entre les sites (à contenu pornographique ou non) et les régies publicitaires. Tous les navigateurs vont intégrer cette fonctionnalité, à laquelle s'ajouteront des greffons que les utilisateurs pourront rajouter afin de supprimer définitivement les cookies et autres dispositifs de traçage[24]. Petit à petit, l'arsenal de fonctionnalités visant à protéger la navigation et la collecte de DCP va s'étoffer, mais nous pouvons remarquer qu'en termes d'ergonomie, ces outils sont peu opérants. La connaissance de ces solutions et la maitrise de leur installation ne sont pas égales chez tous les utilisateurs[25], de telle sorte qu'aujourd'hui de très nombreux navigateurs sont utilisés dans leur configuration d'origine, laissant grande ouverte la porte aux traqueurs.

Cela dit, les développeurs de logiciels réfléchissent aussi à d'autres solutions : protocoles sécurisant l'échange de données, démocratisation de la cryptographie. L'une des solutions consisterait aussi à associer des métadonnées aux règles de confidentialité, permettant d'afficher de manière synthétique les données collectées, leur finalité et la durée de leur conservation. Ces labels pourraient être exportés sur toutes les pages du site, y compris dans les applications pour téléphone mobile[26]. Il est clair qu'une telle initiative demanderait d'abord une adhésion de la part de tous les éditeurs (le W3C et l'ISOC peuvent jouer le rôle de médiateur à ce sujet). Ensuite, cela exigerait une sensibilisation et une éducation du côté des utilisateurs. L'autonomie individuelle ne peut suffire ici, et un tel effort relève plus, nous semble-t-il, des États et des organismes transnationaux.

III.2, Propositions de régulation

Ainsi, les hasards du calendrier font qu'ont été célébrées récemment les 30 ans des *privacy guidelines* publiées par l'OCDE en 1980. À cette occasion, un bilan a été dressé, et des préconisations ont été délivrées[27] : même si elles ne concernent pas directement les DCP collectées sur les réseaux socionumériques, elles invitent à généraliser les cartes blanches[28].

---

[23] World Wide Web Consortium, Internet Engineering Electronic Task Force

[24] AdBlock, Abine, TACO, etc..

[25] Il faudrait étudier une population d'utilisateurs pour voir dans quelles mesures, et pour quels contenus, ces fonctionnalités sont exploitées.

[26] Voir la proposition d'Aza Raskin : http://www.azarask.in/blog/post/privacy-icons/

[27] Voir les publications du Working Party for Information Security and Privacy (WPISP), et notamment Acquisti (2010), à l'adresse suivante : http://www.oecd.org/sti/privacyanniversary

[28] Une carte blanche ou smart card est un support électronique délivrant seulement l'information requise, sans dévoiler d'autres DCP. Un exemple : lors d'un contrôle de police sur la route, la smart card dévoile que vous êtes titulaire du permis de conduire, et que votre nom n'est pas inscrit sur le fichier des personnes recherchées. La carte ne délivre ni votre nom, ni votre adresse, ni votre date de naissance, etc. Voir les travaux de Deswarte Y. et Gambs S., 2010.

Autre coïncidence, la directive européenne 95/46 est en cours de révision. L'Internet Society a aussi engagé une réflexion sur la gestion des identités (programme Trust and Identity).

Ailleurs, en Californie notamment, la tendance au Do Not Track Us (ne nous traquez pas) semble s'institutionnaliser : il s'agit de légiférer sur le mécanisme des options d'adhésion. Actuellement, il y a deux doctrines sur ce mécanisme : l'*opt-out* aux États-Unis, l'*opt-in* en Europe (directive 2002/58). On appelle *opt-in* le consentement explicite du titulaire à l'exploitation de ses DCP : en cochant la case, il autorise le service à collecter, traiter et conserver les données. Dans l'*opt-out*, le consentement relève du refus d'exploitation : il faut décocher la case pour que les données ne soient pas exploitées. Techniquement, sur un navigateur web par exemple, il est possible de signifier le refus de certains cookies réputés associés à des entreprises de prospection commerciale, tout en conservant le bénéfice des autres cookies, et des services en ligne auxquels on souhaite adhérer. Ce mécanisme est renforcé en Europe depuis l'adoption du « paquet télécom ». Il semblerait que cette tendance ait traversé l'Atlantique, même si au niveau fédéral cela semble plus compliqué, et plus modeste : le *Online Privacy Bill of Rights* présenté par les sénateurs McKain et Kerry n'intègre pas cette proposition. Dans le même genre, le USA Patriot Act a été reconduit, favorisant ainsi le décloisonnement complet des données à caractère personnel.

Comme de nombreuses tendances sociales, la protection de la vie privée semble animer d'un rythme ventriculaire : à la protection des années 70 (Privacy Act de 1974, Loi Informatique et Libertés de 78, Privacy Guidelines de l'OCDE en 80, Convention européenne n°108 en 1981) succède la dérégulation des années 80. La construction européenne des années 90 (directive 95/46) semble réduite par les attentats du 11-septembre, et le bouquet juridique qui s'en est suivi : USA Patriot Act, LOPSI, etc.. En ce début des années 2010, les pays émergents adoptent aussi des lois en faveur de la protection des DCP (Inde, Chine, mais aussi Philippines, Costa-Rica). Mais il faudra aussi surveiller ce que recèle le projet américain d'un écosystème identitaire (NSTIC : *National Strategy for a Trusted Identity on Cyberspace*), où convergent des intérêts de sécurité nationale et de rentabilité économique, des enjeux de confiance dans le commerce électronique et dans les technologies qui le supportent, ainsi que des reconfigurations des usages et des interactions sociales[29].

Conclusion

Pour conclure, nous nous permettons de paraphraser Jean Meyrat : l'identification est un processus ayant l'identité et les DCP pour contenu. Nous pensons que l'identité – en tant que métadocument produit de et par les données à caractère personnel – répond à la définition que Franck Rebillard donne d'une configuration sociotechnique : « une modalité évolutive d'agencement social d'une technologie résultant des relations entre groupes sociaux engagés dans sa conception, son utilisation et sa représentation, et (historiquement) structurée par ses modalités antérieures comme par des logiques macro-sociales environnant son développement » (Rebillard, 2007, p. 132). Le traitement des données à caractère personnel saisies ou collectées en ligne se situe entre technicité et imaginaire social. Les langages logico-mathématiques contenus dans les dispositifs (formulaire, base de données) agrègent

[29] http://www.nist.gov/nstic/

des données selon les finalités voulues par leurs bénéficiaires : pour la plupart (des sites web), c'est la logique marchande qui prévaut. Cette exploitation des interactions sociales et des affects individuels donne lieu à une réification sous forme numérique. Il y a loin d'ici à ce que la commensuration devienne monétarisation, à ce que le moyen devienne la fin, mais l'instrumentalisation de l'identité et des données à caractère personnel objective le rapport d'échange interindividuel (Simmel, 1987) : nous pensons cependant que, même si le web cristallise notre rapport à l'identité, l'échange des attributs identitaires a toujours été au cœur du fait social, depuis la politique athénienne et le droit romain jusqu'à l'espace public contemporain (Miège, 2010). C'est pourquoi aussi, en retour d'externalités jugées négatives par les titulaires, et certaines catégories d'acteurs socioéconomiques, les données à caractère personnel sont sujette à un regain de protection. Mais pour encore combien de temps ?

.

Bibliographie

Acquisti A., « The economics of personal data and the economics of privacy », texte de la conférence donnée en décembre 2010 à Paris, « 30 Years after the OECD Privacy Guidelines. The Economics of Personal Data and Privacy», centre de conférence de l'OCDE, accessible à l'adresse suivante : http://www.oecd.org/dataoecd/8/51/46968784.pdf

Codd E.F., «A Relational Model of Data for Large Shared Data Banks», Communications of the ACM, vol. 13, n° 6, 1970, p. 370-387

Deswarte Y., Gambs S. (2010), « A Proposal for a Privacy-preserving National Identity Card », Transactions on Data Privacy, décembre 2010, pp.253-276. http://www.tdp.cat/issues/tdp.a060a10.pdf

Dodier N., « L'espace public de la recherche médicale – Autour de l'affaire de la ciclosporine », Hermès n°95, 1999

Ertzscheid O., « L'homme, un document comme les autres ». Hermès n°53, 2009, p. 33-40

Espeland W. N., « Commensuration and Cognition », *in* Cerula K. A. (dir.), Culture in Mind : Toward a Sociology of Culture and Cognition, New York, Routledge, 2002, p. 64

George F., « Approche statistique de trois composantes de l'identité numérique dans Facebook », in Millerand F., Proulx S. et Rueff J. (dir.), Web social. Mutation de la communication », PUQ, 2010

Goffman E., « La présentation de soi », tome 1 *in* « La mise en scène de la vie quotidienne », Les éditions de minuit, 1973

Lamarche T., 2005. « Les postures critiques de l'informatisation », Terminal n°93-94, p. 101-110

Martin-Juchat F., Pierre J., « Facebook et les sites de socialisation : une surveillance librement consentie » *in* Galinon-Mélénec B. (dir.), « L'homme trace. Perspectives anthropologiques des traces contemporaines », CNRS-Éditions, 2011

Mattelart A., La globalisation de la surveillance, La découverte, 2007

Miège B., Les Tics entre innovation technique et ancrage social, in La société conquise par la communication, tome III, PUG, 2007

Miège B., L'espace public contemporain, PUG, 2010

Pédauque R.T., Le document à la lumière du numérique, C&F éditions, 2006

Rebillard F., Le web 2.0 en perspective, L'Harmattan, 2007

Rochelandet F., Économie des données personnelles et de la vie privée, La découverte, 2010

Simmel G., La philosophie de l'argent, PUF, 1987

Tassin E., « Espace commun ou espace public : l'antagonisme de la communauté et et de la publicité », Hermès, n°10, CNRS Éditions pp.23-37

Zacklad M., 2007. Réseaux et communautés d'imaginaire documédiatisées, *in* Skare, R., Lund, W. L., Varheim, A., A Document (Re)turn, Peter Lang, Frankfurt am Main : 279-297