L'inscription, le masque et la donnée : Datafication du web et conflits d'interprétation autour des données dans un laboratoire invisible des sciences sociales

Gilles Bastin, Jean-Marc Francony

L'INSCRIPTION, LE MASQUE ET LA DONNÉE

Datafication du web et conflits d'interprétation autour des données dans un laboratoire invisible des sciences sociales

Gilles Bastin, Jean-Marc Francony

S.A.C. | « Revue d'anthropologie des connaissances »

2016/4 Vol. 10, n° 4 | pages 505 à 530

L'inscription, le masque et la donnée

Datafication du web et conflits d'interprétation autour des données dans un laboratoire invisible des sciences sociales

Gilles BASTIN
Jean-Marc FRANCONY

RÉSUMÉ

Dans cet article, nous examinons les différentes interactions qui caractérisent un projet d'enquête en sciences sociales fondé sur l'exploitation des données du réseau social professionnel LinkedIn. Ce retour réflexif sur ce que nous appelons le « laboratoire invisible » de l'enquête nous permet de mettre en évidence les conflits d'interprétation qui naissent autour de la définition de ce qu'est une « donnée » tirée du web. Dès lors, la constitution d'une base de données ne doit pas apparaître comme une opération technique de la recherche mais comme un processus de transformation des « inscriptions » individuelles sur le réseau en « données », processus que nous appelons « datafication ». Ce processus passe par la confrontation de « masques » différents que posent sur ces inscriptions les acteurs du « laboratoire invisible » de la recherche : le sociologue et le spécialiste de l'information et du web mais aussi l'utilisateur du réseau, la plateforme et le régulateur public.

Mots clés : Datafication, LinkedIn, Enquête, Bases de données.

INTRODUCTION

Dans cet article, nous revenons sur une expérience récente de collaboration entre un sociologue et un spécialiste du web pour la constitution d'une base de données de carrières professionnelles à partir du réseau social professionnel LinkedIn. Notre objectif est de contribuer à la réflexion engagée depuis quelques années dans les sciences sociales sur les effets méthodologiques, épistémologiques et politiques du « déluge de données » qui a accompagné le développement des pratiques de documentation et de numérisation de pans entiers de l'expérience sociale des individus sur le web (Abbott, 2000 ; Hey et Trefethen, 2003). Cette mutation très rapide des modes de documentation de soi a particulièrement touché les réseaux sociaux dont le nombre d'usagers a très rapidement progressé, tout comme la quantité d'informations qu'ils mettent en ligne sur des sites à vocation personnelle (comme Facebook par exemple) ou professionnelle (comme LinkedIn qui nous occupe ici)[1].

Les chercheurs en sciences sociales se doivent de « regarder » et d'analyser ces nouveaux matériaux qui leur ont longtemps paru suspects (Cardon, 2012). D'autre part ils ont aussi tout intérêt à rendre publics les écueils qu'ils rencontrent en chemin et à rendre visible la façon dont ces données nouvelles mettent en tension les « assemblages » techniques, méthodologiques et déontologiques qui forment leur appareil de preuve (Rupert, Law, & Savage, 2013.). L'enjeu auquel sont confrontées les sciences sociales contemporaines face à ces sources d'informations n'est en effet qu'en partie susceptible d'être réduit à un problème technique ou méthodologique. C'est plus fondamentalement la question de la « juridiction » des sciences sociales – donc de leur légitimité à produire et à interpréter des données – qui se pose lorsqu'elles s'aventurent dans ce qui fait la matière de la sociabilité numérique.

De nombreux travaux ont été consacrés dans les sciences sociales à élucider les conditions de la *factualisation* du social, c'est-à-dire à la production d'énoncés factuels par les sociologues. Une longue tradition épistémologique a en effet remis en question l'évidence selon laquelle il existerait des « faits » sociaux accessibles par simple observation. La difficulté qu'il y a à se prémunir des « prénotions » lorsque l'on énonce des faits sociaux a été placée au cœur de la sociologie à sa naissance (Durkheim, 1895), tout comme « l'illusion de la transparence » produite par la sociologie spontanée (Bourdieu, Chamboredon, & Passeron, 1969) ou encore la difficulté à produire des « effets d'intelligibilité » à partir des informations recueillies lors de l'enquête (Passeron, 1991).

D'un autre côté, la production des agrégats statistiques qui permettent d'étayer ces faits – que l'on appellera ici la *quantification* du social – a aussi donné

1 D'autres sources massives de données ont émergé ces dernières années du fait de la numérisation d'un côté des bases de données de traces laissées par les individus lors de leur passage sur certains sites ou, de l'autre, des contenus produits par les médias. Sur les implications de la numérisation des contenus médiatiques pour la recherche sur les médias, cf. Bastin et Bouchet-Valat (2014).

lieu à de nombreux travaux dans la foulée d'une histoire sociale de la statistique (Desrosières, 1993, 2014) et de l'« histoire concrète de l'abstraction » (Perrot, 1992). Les techniques de quantification des comportements sociaux ne peuvent plus aujourd'hui être considérées séparément des conditions de leur production. Il est devenu incontournable de chercher dans les méthodes statistiques « en même temps leurs apports de connaissance et les circuits sociaux de leurs mises en forme et de leurs usages » (Desrosières, 2005).

Mais le processus que l'on appellera ici la *datafication* du social, c'est-à-dire la fabrication des matériaux à partir desquels sont produits statistiques et jugements de faits, a nettement moins été discuté. On peut suspecter que son invisibilisation dans l'enquête sociologique (Flichy, 2013) est liée pour partie à la faible reconnaissance dont jouissent ceux qui l'assurent (Dagiral & Peerbaye, 2012). Pour autant, comme l'a montré toute une tradition de recherche en sociologie des sciences et des techniques, les « technologies invisibles » (Berry, 1983) peuvent et doivent être rendues manifestes afin que, selon la formule connue de Bruno Latour, ce qui est de l'ordre de l'« obtenu » dans les « données » ne soit pas occulté[2].

La métaphore de « l'obtenu » renvoie cependant selon nous encore à l'intervention du chercheur ou de ses assistants dans les « données ». Dans cet article, nous montrons que les « données » sont surtout « disputées » entre de nombreux acteurs sur le web. Pour décrire les processus de ces disputes, nous proposons de parler d'*inscriptions pour désigner ce que les individus laissent derrière eux sur le web*[3]. L'enquête fondée sur des données tirées du web a en effet ceci de particulier qu'elle repose sur un processus long et complexe de transformation de ces *inscriptions* sur une plateforme de publication (par exemple la saisie du nom d'un employeur, de dates de début ou de fin de certaines activités dans le cas de LinkedIn) en un ensemble de *bases de données* dans un premier temps puis en un ensemble de *tables de données* croisant de façon très classique des individus en ligne et des variables en colonne dans un second. Cette phase de production des « données » est sans doute jusqu'à aujourd'hui la partie la moins bien connue et discutée de l'enquête en sciences sociales en général.

Afin de la décrire, nous nous concentrons sur les opérations les plus élémentaires que nous avons entreprises dans cette recherche et nous posons la question des circuits sociaux de définition – plus que d'usage – de la donnée. Il nous semble en effet que la question de la définition même de ce qu'est une donnée est la première à laquelle est confronté le chercheur en

2 « La tentation de l'idéalisme vient peut-être du mot même de "données" qui décrit aussi mal que possible ce sur quoi s'appliquent les capacités cognitives ordinaires des érudits, des savants et des intellectuels. Il faudrait remplacer ce terme par celui, beaucoup plus réaliste, d'"obtenues" et parler par conséquent de "bases d'obtenues", de "*sublata*" plutôt que de "data" pour parler à la fois en latin et en anglais », in Bruno Latour, « Pensée retenue, pensée distribuée », *Lieux de savoir*, *1* (dir. Christian Jacob), Paris, Albin Michel, p. 609.

3 Nous empruntons ce terme à Madeleine Akrich (1987).

sciences sociales[4]. Notre perspective consiste à observer les relations entre tous ceux qui participent au laboratoire sociologique qu'il est nécessaire de mettre en place pour extraire des *inscriptions* du web et les transformer en *données* de la recherche. Nous mettons particulièrement en avant les conflits d'interprétation qui naissent dans ce laboratoire autour de ces données qui, avant de pouvoir être considérées comme telles, subissent un travail conflictuel de définition que nous considérons comme la production de « masques » différents entrant en compétition les uns avec les autres (Wiszniewski & Coyne, 2002). Les inscriptions laissées par les utilisateurs des réseaux sociaux sur le web sont le produit d'un actif travail de « design de la visibilité » de la part de l'utilisateur qui y voit le moyen de se construire une personnalité (Cardon, 2008) mais aussi de la plateforme de publication et d'autres acteurs. Ce travail interfère avec le travail de design scientifique de la base de données qui mobilise lui aussi une définition du rapport entre inscription et donnée.

Ces masques concurrents sont ceux que les différents intervenants du « laboratoire invisible » de la *datafication* posent sur les données : le sociologue et le spécialiste des techniques de collecte de l'information sur le web mais aussi l'utilisateur du réseau social et premier producteur des inscriptions qui intéressent le sociologue, la plateforme de publication LinkedIn dont les modes de gestion de ces inscriptions influent sur la capacité que nous avons de les interpréter et de les utiliser et enfin le régulateur public chargé de faire respecter le droit et qui qualifie pour sa part juridiquement les inscriptions des individus en les appelant des « données personnelles ».

LA PLATEFORME : LES ENJEUX DE MISE EN VISIBILITÉ DES INSCRIPTIONS INDIVIDUELLES SUR LE WEB

L'enquête qui est analysée ici vise à décrire à grande échelle les trajectoires professionnelles des journalistes à partir des profils qu'ils publient sur le réseau LinkedIn. Cette enquête peut être qualifiée d'« indirecte » par comparaison avec l'enquête directe par questionnaire. En effet, elle utilise des données disponibles sur le web et ne suppose donc pas nécessairement une démarche de l'enquêteur auprès des individus enquêtés. Pour reprendre une métaphore

4 La réflexion sur les questions de causalité qui ont souvent été mises en avant comme caractéristiques de l'émergence des *big data* ne nous semblent pas devoir occulter les aspects techniques, sociaux, éthiques et politiques (au sens de l'économie politique des données) des conflits de définitions qui naissent aujourd'hui autour de ce « masque » qu'est la donnée sur les réseaux sociaux.

usuelle, elle repose sur l'interprétation de « traces » d'activités laissées par les individus[5]. Pour le sociologue, la démarche vise avant tout à solutionner des problèmes méthodologiques qui se posent aujourd'hui dans l'étude de ce groupe professionnel. La plupart des données disponibles actuellement en France sont en effet issues des organismes professionnels de journalistes. Ces données reposent sur une définition restrictive des acteurs du monde des médias (les titulaires de la carte de presse) et ne permettent pas de décrire des carrières complètes. Elles n'intègrent en effet pas les activités non journalistiques et ne permettent pas de suivre les individus après une éventuelle sortie de la profession. De plus elles sont strictement nationales et posent des problèmes de comparaison. Le développement des réseaux sociaux professionnels comme LinkedIn, sur lesquels les individus rendent visible leur *curriculum vitae*, est donc vite apparu comme une opportunité d'accéder à de nouvelles données en grand nombre, fondées sur de l'autodéclaration, permettant d'envisager une démarche d'enquête longitudinale, et comparables internationalement[6].

Les journalistes utilisent de nombreux réseaux sociaux pour leur activité professionnelle. La perméabilité entre médias et réseaux sociaux (présence des réseaux sur les sites des médias et structuration des données sur les réseaux comme des flux d'information) a massivement introduit ces plateformes techniques dans l'environnement de travail immédiat des journalistes qui y trouvent des informations ou des sources. Ces réseaux jouent aussi un rôle important de mise en visibilité personnelle dans une profession traversée de tendances à l'individualisation et à la personnalisation (marquée notamment par la signature des contenus produits) tout en reposant sur des contraintes collectives très fortes. Enfin, ils permettent de s'assurer – par le suivi des trajectoires des journalistes appartenant à son réseau – une forme de veille sur le marché du travail, semblable au mécanisme de « quasi-recherche » d'emploi qui a été bien décrit par Mark Granovetter pour les cols blancs[7].

5 La démarche s'apparente de ce fait au paradigme de l'indice théorisé par Ginzburg dans la mesure où elle repose sur une interprétation de données qui n'ont pas été intentionnellement constituées pour le chercheur. Voir Ginsburg (1989), « Traces. Racines d'un paradigme indiciaire ». Ce paradigme fait aujourd'hui l'objet d'un regain d'attention dans les études sur le numérique. Voir notamment Antonio Casili (2010), *Les liaisons numériques. Vers une nouvelle sociabilité ?*, Paris, Seuil et Louise Merzeau (2009), « Du signe à la trace : l'information sur mesure », *Hermès, La Revue* (1). La référence aux *big data* comme relevant de sciences sociales fondées sur le modèle de la police scientifique (« forensic social sciences », Goldberg, *Big Data and Society*, 2015) relève aussi de ce retour en force du paradigme de l'indice.

6 Pour plus d'informations sur le cadre de l'enquête, voir Bastin (2015).

7 Voir Mark S. Granovetter (1995), *Getting a Job: A Study of Contacts and Careers*, Chicago, The University of Chicago Press, chapitre 1. On peut voir une confirmation de l'importance des réseaux pour les journalistes dans la création du réseau JOL qui présente les trois caractéristiques décrites plus haut : JOL-Group créé en 2011 héberge un site d'information (JOL-Press), le réseau JOL-Social depuis 2013 et une « place de marché d'articles, de contenus éditoriaux, à l'unité », JOL-Store qui se définit comme « la plus grande rédaction au monde ».

L'intérêt d'enquêter sur un réseau social comme LinkedIn apparaît donc nettement. Par opposition à d'autres réseaux sociaux comme Facebook, Twitter ou Flickr, LinkedIn a une vocation professionnelle très marquée. L'objectif des utilisateurs est de présenter leur « profil » professionnel sous la forme d'un CV et de le faire connaître dans des cercles de plus en plus larges de personnes, sur la base des recommandations faites par le site lui-même, de recherche par mots clés ou encore de l'appartenance à des groupes communs[8]. Ce réseau revendique près de 400 millions d'utilisateurs dans le monde en 2015, dont plus de 8 millions en France, ce qui fait de ce pays le sixième plus important. De très nombreux journalistes sont présents sur LinkedIn en France : au moment où cette enquête a été lancée en janvier 2011 (le site ne comptait alors « que » 90 millions d'utilisateurs dans le monde), une recherche sur l'ensemble des profils du site avec le mot clé « journaliste » (donc restreinte à l'espace francophone) donnait 11.956 résultats[9]. De nombreux groupes rassemblent des journalistes comme « Journalistes francophones », « LinkedIn for Journalists », « Linked Journalists », « Media Jobs », « Journalists and Journalism » ou « Media Professionals Worldwide ». Par ailleurs, LinkedIn a lui-même produit des arguments sur son utilisation par des journalistes, pas seulement pour promouvoir et rendre visible leur carrière mais aussi à des fins d'enquête journalistique[10]. Des salariés de LinkedIn proposent d'ailleurs des formations spécifiques pour les journalistes afin de les inciter à utiliser le réseau[11].

La mise en visibilité des trajectoires professionnelle sur LinkedIn se fait au moyen de plusieurs champs remplis par l'utilisateur pour chacune des activités qu'il rend publiques (figure 1). Les formations suivies sont documentées au moyen de trois champs : le diplôme obtenu, l'institution et la matière étudiée. Les emplois au moyen de trois champs aussi : l'activité, l'employeur et un résumé laissé à la liberté de l'utilisateur. Pour chaque séquence – formation ou emploi –, le profil contient la date de début et la date de fin éventuelle. Enfin, il contient aussi un résumé global, l'indication de la localisation géographique de la personne, des mots clés décrivant ses compétences et la liste des groupes auxquels elle participe.

8 Voir Zizi Papacharissi (2009), « The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld », *New Media & Society*, 11.

9 Sur ces 11.956, 7.189 mentionnent une résidence en France. Une recherche sur le mot clé « journalist » donne 150.702 résultats à la même date, dont 3.842 en France. Ces chiffres ont augmenté depuis. Le 4 février 2013 le moteur de recherche renvoyait 39.741 « journalistes » et 317.407 « journalists ».

10 Cf. http://press.linkedin.com/about et http://press.linkedin.com/linkedin-for-journalists.

11 Cf. http://www.poynter.org/latest-news/top-stories/253445/spreading-the-gospel-of-linkedin-for-journalists/: et aussi pour d'autres exemples : http://ijnet.org/stories/five-mistakes-journalists-make-linkedin; http://www.journalism.co.uk/news-features/10-linkedin-tips-for-journalists/s5/a547539/; http://www.journalism.co.uk/news-features/10-linkedin-tips-for-journalists/s5/a547539/.

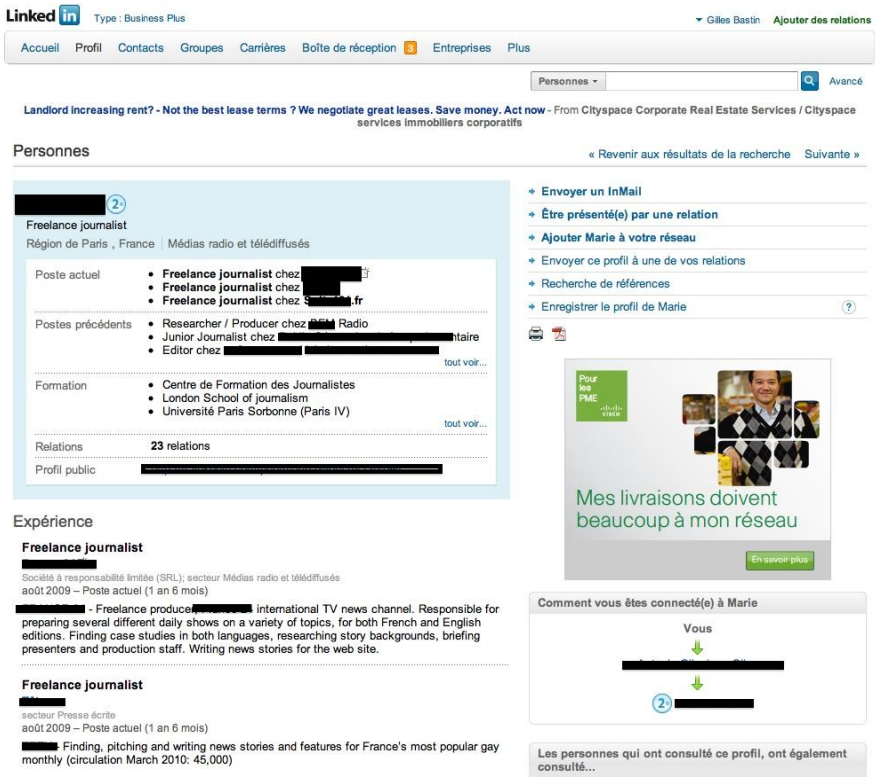

Figure 1. Un profil de journaliste utilisateur de LinkedIn

Pour ce qui est de la mise en visibilité de ces inscriptions, et donc de leur accessibilité pour la recherche, le site obéit à une triple logique.

La première est une logique d'innovation permanente. Les plateformes comme LinkedIn doivent en effet encourager le développement d'écosystèmes applicatifs en favorisant une ouverture de leurs données aux développeurs du web. Les interfaces applicatives (API) assurent cette fonction de portail avec des restrictions qui ont un effet sur l'enquête. Celles-ci portent sur l'accessibilité en volume et en nature afin d'établir un compromis entre attractivité des données et contrôle de la valeur. La solution communément adoptée – et qui prévaut chez LinkedIn – consiste à organiser un accès égocentrique, c'est-à-dire restreint aux données accessibles depuis un compte authentifié pour une profondeur ou un volume d'information limités. Ce bornage au voisinage immédiat de l'abonné (*ego*) empêche l'aspiration récursive des données tout en encourageant le développement de services aux abonnés (pour mieux organiser leurs contacts…). Les plateformes se réservent de cette manière le potentiel de valeur associé à l'analyse globale des données et à la prédiction qui suppose de traiter de grandes masses de données. Le masque égocentrique appliqué aux inscriptions par les outils fournis par LinkedIn (comme le moteur de recherche interne de la plateforme) ou par les développeurs d'API dont la marge de manœuvre est limitée par LinkedIn ne permet donc pas au sociologue de « voir » ces inscriptions comme il est habitué à le faire avec d'autres types de données,

c'est-à-dire sans biais d'échantillonnage. L'échantillonnage de ces profils est en effet rendu compliqué par le phénomène de dépendance au point d'entrée dans le réseau : le moteur de recherche renvoie d'abord les profils les plus proches (en termes de connexions) de l'utilisateur qui fait la recherche et interdit une segmentation de l'ensemble des utilisateurs selon des variables sociologiques.

La seconde logique est celle de la monétisation des inscriptions individuelles sur la plateforme. Puisqu'elle assure une source de revenus, la commercialisation de données est une stratégie que les plateformes envisagent comme moyen de consolider leur modèle économique qui reste le plus souvent fragile. Dans le cas de LinkedIn, le service fourni par la plateforme est fourni gratuitement dans une version dégradée et des classes d'abonnement de niveau croissant permettent d'étendre la visibilité de l'offre personnelle (en donnant à l'utilisateur plus de moyens de rendre son propre profil visible) et d'augmenter les volumes des résultats de recherche sur la plateforme (accès à davantage de profils d'autres utilisateurs). Cette seconde logique se combine avec la première pour limiter les possibilités d'accès du sociologue à une grande masse de données : celui-ci ne peut accéder qu'à un « petit » nombre de résultats pour chaque requête qu'il soumet au moteur de recherche de LinkedIn et les profils renvoyés sont toujours les mêmes du fait du masque égocentrique appliqué par la plateforme.

Une troisième logique permet cependant d'envisager une solution à cette difficulté d'accès aux inscriptions des utilisateurs. C'est celle de la viralité. Les plateformes comme LinkedIn organisent en effet une « porosité » contrôlée des inscriptions de leurs utilisateurs afin d'assurer leur visibilité et leur promotion, notamment auprès des moteurs de recherche. En fonction des réglementations, cette nécessité a amené les opérateurs à adapter leurs chartes de confidentialité et à proposer la personnalisation de pages publiques visibles sur le web pour tous. Afin de limiter les possibilités de détournement ou d'aspiration des données ainsi rendues publiques, les plateformes utilisent pour ce faire des technologies dynamiques du web qui renforcent « mécaniquement » la sécurisation des données en rendant difficile l'exploitation automatique des contenus des pages diffusées. Le sociologue peut donc se replier sur des moteurs de recherche généralistes du web pour identifier les pages LinkedIn à partir de mots clés. Il est alors soumis aux limitations de ces moteurs qui sont moindres que celles imposées par le masque égocentrique de la plateforme mais bien réelles aussi. On peut en citer deux : la dégradation des équations de recherche en fin de pile qui rend l'échantillonnage assez aléatoire ; une recherche d'information holistique dans toute la page d'un individu, ce qui produit beaucoup de bruit. La recherche de page LinkedIn contenant le mot « journaliste » sur un moteur généraliste renvoie par exemple de nombreuses pages dans lesquelles ce mot est employé ailleurs que dans la description de la carrière personnelle d'un individu, par exemple dans le profil de ses quelques relations qui s'affichent sur sa page.

D'un point de vue technique, la conduite de l'enquête dans le cas particulier de LinkedIn conduit à traiter des masques publics qui sont par nature fortement

structurés. Cette structuration est guidée par le schéma général offert par la plateforme et utilisé par les abonnés. Ce formatage correspond aux grandes rubriques d'une description individuelle (carrière, parcours de formation, compétences, etc.) et assure ainsi l'unité globale des présentations personnelles. Cependant les règles de compositions associées à ces schémas sont souples. Elles autorisent de nombreuses variations, permettant aux abonnés de s'approprier une forme graphique et d'exprimer les singularités de leurs profils. Cette diversité, peu sensible visuellement, est en fait très perturbante lors de l'extraction automatique d'information. La sémantique des rubriques est fluctuante et les attributs, par exemple l'intitulé de formation, le niveau d'étude, etc., ne sont pas toujours interprétés de manière univoque. Par ailleurs, l'expression des dates et des durées conduit également à des formes très diverses (notations abrégées, codes anglo-saxons, etc.). Enfin, la rédaction des entités référentielles (telles que les noms des institutions) a été laissée libre par LinkedIn, ce qui donne lieu à une très grande diversité de formes lexicales supposées équivalentes et plus ou moins bien orthographiées[12]. Ces écarts sont peu préjudiciables pour le lecteur humain mais redoutables dans une perspective computationnelle pour le sociologue.

LA RELATION D'ENQUÊTE ENTRE SOCIOLOGUE ET UTILISATEUR SUR LE WEB

Par bien des aspects, les univers numériques dans lesquels interagissent les individus aujourd'hui ressemblent aux mondes sociaux que Shibutani proposa en 1955 de qualifier de « relâchés » (*loosely connected*) pour les opposer aux mondes très solidaires d'un côté (comme l'*underworld* ou les élites sociales) et aux réseaux d'association volontaire de l'autre (comme les syndicats ou les groupes d'intérêt). À l'image des amateurs de sport de plein air ou des collectionneurs de timbres, qui peuvent échanger par l'intermédiaire de médias spécialisés, les utilisateurs de LinkedIn n'entretiennent pas de relations régulières les uns avec les autres et ne partagent pas nécessairement les mêmes intérêts[13]. Leurs interactions sont souvent fugaces (un exemple typique est le fait de cliquer sur un bouton pour accepter une mise en relation avec un autre membre du réseau). Lorsqu'ils se connectent à LinkedIn, la probabilité qu'ils croisent d'autres membres qu'ils connaissent de façon récurrente dans les flux d'actualité qu'ils parcourent est assez faible dès lors que leur réseau s'étend.

12 Il semble du reste que les derniers développements de LinkedIn tendent à régler cette absence de contrôle, allant ainsi dans le sens des *linked data*.

13 Shibutani (1955) « Reference groups as perspectives », p. 566.

Le point de vue de l'observateur des traces

De ce fait, le sociologue qui parcourt ces mondes peut se percevoir comme relativement étranger à leur fonctionnement tout en ayant accès à un très grand nombre d'informations. Il peut concevoir sa présence comme une forme d'observation participante dans laquelle les exigences de participation seraient minimales et les possibilités d'observation maximales. Les outils de collecte des données – une fois passée la phase d'observation préliminaire – peuvent eux aussi conforter ce sentiment par leur nature mécanique, invisible et anonyme pour l'utilisateur lambda, à l'opposé des outils traditionnels de collecte des données manipulés par les sociologues comme l'entretien ou la passation de questionnaire. Les tensions habituelles auxquelles sont confrontés les praticiens de ces deux méthodes (comme les problèmes de compréhension des questions et des réponses, d'imposition de problématique, de cadrage de l'interaction ou de neutralisation des distances sociales entre enquêteur et enquêté) ne semblent pas de mise lorsque la collecte est assurée par un robot informatique.

Notre expérience remet cependant en question ce type de raisonnement *a priori*. De nombreuses tensions entre enquêteur et enquêté doivent aussi être surmontées pendant l'enquête afin que celle-ci produise des données exploitables. Nous en isolerons ici deux types principaux : des tensions sur les inscriptions des individus d'un côté ; des tensions sur les intentions du sociologue de l'autre.

Le premier effet de la collecte d'informations sur un réseau social comme LinkedIn est un effet de décalage entre inscription et donnée. Le matériau qui est rendu visible sur ce site s'apparente en effet à une inscription de l'utilisateur qui est en décalage aussi bien temporel qu'en termes de finalité avec la production d'une donnée de l'enquête par le sociologue. L'inscription a été faite antérieurement à l'enquête et elle n'a pas été faite pour l'enquête. La première tension qui émane de ce double décalage concerne la possibilité de transformer facilement cette inscription en donnée. L'utilisateur peut simplement s'être trompé dans son inscription, mal utiliser les zones de saisies qui lui sont proposées (par exemple les trois zones différentes décrivant chaque étape de la formation) ou encore introduire dans cette inscription des caractères qui mettent en échec le robot censé collecter les données qui les interprète non pas comme des inscriptions mais comme du code informatique (guillemets, point-virgule…). La conséquence immédiatement observable pour l'enquêteur est la multiplication des inscriptions singulières qu'il lui faudra réduire en catégories pour les transformer en données. Si l'on se limite au cas du niveau de diplôme obtenu par exemple, le premier échantillon de 10.573 individus enquêtés (soit 26.556 inscriptions de diplômes) a conduit à collecter 8.141 inscriptions de diplômes différentes alors que l'enquête s'appuie *in fine* sur une variable à 6 modalités décrivant le niveau de diplôme.

On pourrait être tenté d'attribuer ces tensions à l'utilisateur lui-même (selon l'adage d'informaticiens : « *never trust a user entry* ») mais c'est bien l'écart entre le moment et la finalité de l'inscription et de sa transformation

en donnée qui est en cause. Le partage des inscriptions n'a en effet pas pour finalité l'enquête sociologique mais la création de réseaux relationnels et d'une audience. L'absence de certaines informations essentielles à l'enquête sociologique sur LinkedIn en atteste, comme le cas du sexe de la personne ou de son âge. Mais plus fondamentalement, l'enquête doit prendre en compte le fait que ce *design* des données conduit aussi à omettre certaines informations (comme une activité qui introduirait un doute sur l'identité professionnelle revendiquée dans le profil) ou à des variations dans le niveau d'agrégation de ces données (ainsi les journalistes *freelance* peuvent être tentés de résumer en une seule ligne un grand nombre de collaborations pour montrer la diversité de leur portefeuille de piges).

Le phénomène de « double herméneutique »

Ce travail des utilisateurs intègre aussi des représentations de la société et une relation au sociologue. On peut le montrer à partir des réactions des utilisateurs de LinkedIn postées sur le « groupe » que nous avons créé en 2011 lors du lancement de cette enquête. Afin de rendre publique l'enquête et de recueillir le consentement des utilisateurs à ce que nous exploitions leurs données notre première démarche a en effet consisté à créer un groupe d'utilisateurs du réseau social intitulé « Profil : Journaliste » et de poster sur ce groupe un message présentant notre démarche. Ce groupe a connu un succès considérable qui a largement dépassé le cadre de notre enquête. Il compte actuellement plus de 11.500 membres qui échangent régulièrement sur le journalisme et ont sans doute – pour la plupart – perdu de vue sa finalité première.

Mais dès son lancement il nous est apparu clairement que nous ne pouvions pas nous considérer comme extérieurs à l'objet et simples polygraphes mécaniques des inscriptions individuelles. Le phénomène de « double herméneutique » qui réduit les sciences sociales à être des interprétations d'interprétations (Giddens, 1987) joue en effet aussi sur les réseaux sociaux. Un grand nombre des 125 commentaires suscités par la création du groupe sur LinkedIn peut s'apparenter à une forme d'offre de service en général étayée par un bref résumé de carrière. Des formules de délégation renforcent parfois l'idée selon laquelle les sujets de l'enquête s'en remettent aveuglément au sociologue et à son laboratoire. Ainsi de cette formule que l'on trouve dans un des commentaires : « le CNRS est un organisme sérieux » (5 janvier 2014) ou de formules d'encouragement comme « courage pour l'enquête » (22 août 2012).

Cependant, si l'on y regarde de plus près, on voit très nettement apparaître dans ces commentaires des formulations qui relèvent de trois types d'intervention dans l'enquête. Le premier type correspond à l'énoncé de théories sur le sujet de l'enquête qui peuvent facilement être considérées comme des prénotions par le sociologue (« pour commencer dans ce métier, conseils et expériences valent parfois mieux que de longs discours », 9 avril

2012 ; « il faut commencer à travailler dans les médias locaux », 20 août 2012 ; « Je pense qu'il faille distinguer le métier de la vocation et de l'opportunité »). Le second correspond à des conseils sur l'enquête, sa méthode (« Je vous invite à contacter directement les écoles », 13 février 2013), voire des critiques (« une enquête n'est rien d'autre qu'un sondage supplémentaire ! », 13 février 2013) ou des demandes de précisions sur les objectifs scientifiques (« À qui ça va servir ? », 11 juin 2014). Un des intervenants dans cette discussion résume bien cette volonté de participer au débat sur le cadre de l'enquête : « Bonjour à toutes et tous. Quelles sont les catégories ? On peut être journaliste carté, journaliste sans carte, journaliste ayant perdu sa carte, journaliste faisant de la communication et du journalisme, communicant se qualifiant de journaliste. Et peut-on définir une profession à partir de l'utilisation subjective du mot qui la définit ? »

Pas plus sur le web que dans les modes plus traditionnels de collecte des données on ne peut séparer le consentement à faire partie du champ d'une enquête de l'expression d'un point de vue – éventuellement de dissensions – sur le sujet de l'enquête d'un côté et sur ses principes méthodologiques de l'autre[14].

La négociation des obtenus

Enfin, on trouve un troisième type très fréquent qui paraît de prime abord relever du malentendu. De nombreux commentaires ont pour objectif d'entamer une relation d'enquête avec le sociologue (« Je suis prêt à faire profiter de mon savoir », 19 novembre 2011 ; « Si vous souhaitez l'éclairage d'un homme de communication, je suis à votre disposition », 7 avril 2012 ; « Je répondrai volontiers à vos questions », 25 mai 2013).

Il était pourtant clair pour nous que le fait de rendre publique l'enquête ne visait qu'à faire état de la collecte des inscriptions contenues dans les profils, pas à solliciter de nouvelles données (sauf sous la forme d'un envoi de CV que nous mentionnions comme une possibilité). Pour beaucoup d'individus, une enquête passe nécessairement par une interaction par interview ou questionnaire. Dès lors, la publicisation de l'enquête au moment où les inscriptions sont recueillies passe pour une sollicitation à répondre à des questions.

Ce malentendu révèle le fait que les individus dont les sciences sociales convoitent les « données » connaissent les modes habituels de recueil de l'information de ces disciplines et s'attendent à le voir appliqué. La constitution d'une base de données tirée d'inscriptions sur le web doit donc s'accompagner d'une forme de pédagogie pour légitimer la méthode au risque de la voir rejetée non pas pour des raisons déontologiques que l'on aborde dans la

14 Il faut évidemment ajouter ici que les « enquêtés » ont cette particularité d'être eux-mêmes dans leur propre activité des « enquêteurs ». Sur les effets de la concurrence entre journaliste et sociologue autour des formes d'enquête comme l'interview, voir Gilles Bastin (2012).

section suivante (jamais évoquées dans les commentaires) mais simplement parce que les individus ne considèrent pas ces inscriptions comme des données intéressantes en soi.

Le dispositif expérimental : la donnée comme collection et mise en corpus des inscriptions individuelles

La problématique des grandes masses de données a rapidement fait partie de notre projet d'étude. Cette orientation s'est imposée naturellement dans un contexte où les discours sur le *big data* sont très liés au devenir des technologies de l'information et du web. Sans être à une échelle suffisante pour revendiquer le label *big data*, nos travaux concernent en effet des volumes suffisamment importants pour que ni l'activité de collecte ni celle de la préparation des données ne puissent être envisagées sans le recours à un dispositif informatique automatisant le traitement de ces données. La mise en œuvre de ce dispositif crée des tensions entre le sociologue et le spécialiste de l'information sur le web qui sont à la fois de nature technologique et conceptuelle.

Sur un plan technologique, la constitution d'une base de données à partir des inscriptions des individus crée un effet de boîte noire pour le sociologue car elle introduit une distance technique entre lui et les données. Les techniques de la fouille du web (*web mining*) mobilisent en effet des connaissances informatiques comme la programmation ou l'écriture de requêtes ainsi que la maîtrise des technologies du web. Le dispositif de collecte des données devient de ce fait plus opaque pour le sociologue dans la mesure où il repose sur une traduction technique de ses interrogations en requêtes informatiques. Par exemple, une question très simple comme celle de l'identification du sexe des personnes entrant dans le champ de l'enquête ne peut pas passer par la formulation d'une question comme dans un questionnaire ou par l'évidence de la relation de face-à-face comme dans un entretien. Elle ne peut être résolue que par la combinaison de requêtes informatique. Dans notre cas : trouver les balises identifiant le prénom dans la page web, collecter ce prénom, le comparer systématiquement avec la base fournie par l'INSEE qui donne pour chaque prénom une probabilité de genre et finalement régler les problèmes d'ambiguïté qui se posent nécessairement lors de l'utilisation d'une routine statistique.

Ces tensions sont aussi de nature conceptuelle. Dans une enquête de ce type, le cycle de vie des données est plus long que dans une enquête traditionnelle. De nombreuses étapes de traitement des inscriptions s'interposent entre ce que l'on pourrait appeler leur « découverte » (lorsque le sociologue explore la plateforme pour la première fois et constate le nombre très élevé de réponses à une requête dans son moteur de recherche) et leur « collection » dans une base de données. La donnée du web nécessite un processus de maturation qui passe par des traitements et une lecture suivant différents niveaux d'analyse

et d'abstraction. L'absence d'interaction avec l'utilisateur rend par exemple impossible de demander à celui-ci des précisions sur ce qu'il a voulu dire (comme dans un entretien) ou de lui préciser le sens des questions qui lui sont posées (comme dans un questionnaire). Nous travaillons avec un matériau qu'il faut en permanence interpréter sans l'aide de ceux qui l'ont produit. De là naît un processus itératif dans lequel une première requête produit des données qui révèlent des problèmes de formulation et rendent nécessaire une seconde requête, puis une troisième, etc. Les deux chercheurs qui mènent cette enquête doivent donc collaborer en permanence pour améliorer la qualité des données produites. De manière générale, la fouille du web nécessite une approche par étapes, chacune d'elles correspondant à une activité dont l'aboutissement permet d'orienter la suite du processus. Cette logique de développement, inspirée des méthodes agiles a imposé une architecture fonctionnelle organisée autour de trois cycles de production et de raffinement de données. Ces cycles sont constitués de phases ou étapes de traitements caractérisées par un objectif de production (ci-après inscrit dans des rectangles). Ces objectifs se concrétisent techniquement dans l'élaboration de base de données (cylindres) et produisent simultanément une connaissance empirique associée à la manipulation des informations enregistrées. Cette connaissance nourrit à différents niveaux d'ordre technique ou méthodologique le processus expérimental et amène des régulations dans la collecte ou des révisions plus globales dans la définition du dispositif ou des objectifs de l'expérimentation.

Les cycles de la donnée

Pour le premier de ces cycles (figure 2), la première étape est celle de l'identification des ressources (documents) à partir desquelles l'étude sera conduite. Dans le cas présent, il s'agit des inscriptions individuelles (page web) correspondant à un critère d'éligibilité (être ou avoir été journaliste à un moment de sa carrière). L'indexation mise en œuvre par les moteurs de recherche, que ces derniers soient internes et spécialisés (*e.g.* LinkedIn) ou externes et généralistes (*e.g.* Bing), est un moyen d'établir un premier filtrage afin de constituer une base de références, c'est-à-dire d'URL pointant vers ces pages. On constate cependant que les résultats obtenus ne valident qu'imparfaitement les hypothèses de filtrage. En effet, les techniques d'indexation et les logiques de l'offre de service qui président au fonctionnement des réseaux sociaux et des moteurs de recherche (qui produisent *toujours* des résultats en dégradant au besoin la qualité de leur réponse) ne permettent pas d'accorder un crédit trop important à la sélection d'URL proposée. C'est la raison pour laquelle il est préférable d'adopter une approche faiblement sélective des URL et de reporter à une étape ultérieure l'établissement de la validité des contenus des pages qui leur sont associées. En revanche, il est possible dès cette phase d'identification d'établir si une URL est ou non associée à une page web d'une inscription LinkedIn.

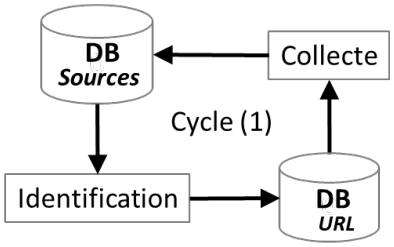

Figure 2. Cycle d'extraction des données du web

La seconde étape est celle de la collecte des contenus des pages liées aux URL identifiées. Deux types de pages sont produits selon que l'on s'adresse au site comme utilisateur authentifié ou non. Les pages individuelles sont produites à la volée, à partir d'informations contenues dans le système d'information de LinkedIn. La technologie Ajax mise en œuvre ne permet pas d'en saisir les contenus. Ceux-ci ne sont en effet assemblés dans une forme définitive qu'au moment de la visualisation de la page sur un navigateur. Il est de ce fait nécessaire d'exploiter le moteur graphique d'un navigateur afin de simuler la restitution et d'accéder aux représentations internes et complètes des pages. Cette étape longue (plusieurs dizaines d'heures), du fait de la sollicitation des serveurs, ne permet pas d'assurer un contrôle fin et continu sur son déroulement ni sur les contenus extraits. Les représentations associées à ces données (pages) sont susceptibles de variations difficilement maîtrisables. Par exemple, ne pas fixer la langue d'interprétation du navigateur virtuel entraîne un choix de langue aléatoire par LinkedIn, ce qui a des conséquences sur la représentation interne des données. Par ailleurs, les réponses du serveur à une sollicitation varient en fonction de la fréquence des requêtes qu'on lui soumet, de leur nature authentifiée ou non, du statut humain/robot qu'il calcule, ainsi que de la charge de requêtes qu'il gère. Ainsi, la réponse à une requête peut être la page attendue mais aussi une page de groupes homonymes ou professionnels, une proposition d'abonnement individuel ou de rattachement à un groupe, un message publicitaire ou une page d'authentification. La plateforme déjoue en quelque sorte en permanence la volonté des chercheurs de transformer les inscriptions individuelles en données. Il en résulte une fragilité du processus qui tient à sa nature et à sa mise en œuvre massive.

Pour ces raisons, les représentations des contenus de pages sont mémorisées afin de constituer un point de reprise dans le processus d'analyse et d'en réduire la durée. La constitution de cette base de données repose sur un processus qui conduit, soit à boucler sur un premier cycle (1) que nous associons à celui d'une collection de données sources, soit à embrayer sur un second cycle (2) que nous associons alors à celui d'une collection de données premières dérivées de la précédente.

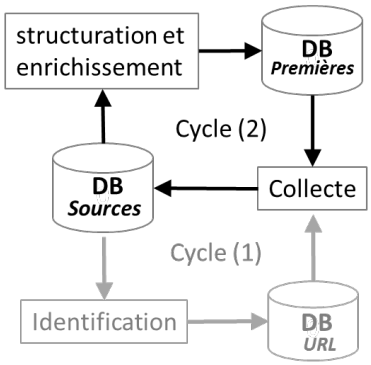

Figure 3. Cycle d'enrichissement et de consolidation des données

Le second cycle (figure 3) joue un rôle intermédiaire dans la recherche. Sa justification repose sur un double argumentaire. Il s'agit premièrement d'améliorer les articulations entre les différents niveaux de traitements afin de faciliter et d'organiser les développements informatiques. Ces articulations techniques s'imposent dans une perspective méthodologique allant au-delà de l'expérience en cours. Elles conditionnent la réutilisation des éléments logiciels produits dans un projet plus vaste d'atelier logiciel dans lequel les modules répondent à d'autres contraintes, notamment de réutilisation pour le spécialiste du web.

Le second argumentaire est celui du cycle de vie des représentations numériques d'un projet guidé par les données. C'est ainsi que nous distinguons d'un côté les collections (cycles 1 et 2) sur lesquelles reposent les enjeux de conservation dans un rapport quasi « patrimonial » à la donnée source ; et de l'autre, les corpus (cycle 3) dont la production est nécessitée par des « urgences » ponctuelles ou des problématiques circonscrites auxquelles la collection doit répondre. Le temps de la collection est celui de l'archive numérique, de la coopération scientifique et de la preuve. Le temps du corpus est davantage celui de la résolution de problèmes et de la communication scientifique.

D'un point de vue technique, cette structuration intermédiaire, a favorisé le processus de développement informatique en amenant plus de lisibilité et d'intelligibilité des données. Ce point s'est avéré crucial pour la compréhension globale du processus expérimental et de l'instrumentation de la collecte. Dans une analogie mécanique, l'articulation ainsi produite, apporte un degré de liberté supplémentaire. Cela se traduit par une granularité plus fine des données favorisant une meilleure liaison dans l'organisation des modules fonctionnels et ainsi une réponse plus facile aux besoins non anticipés. Cette articulation permet en outre de mettre en place une gestion de versions de données propre à supporter les reprises et les aléas de traitements mécaniques ou encore des incertitudes de l'analyse empirique.

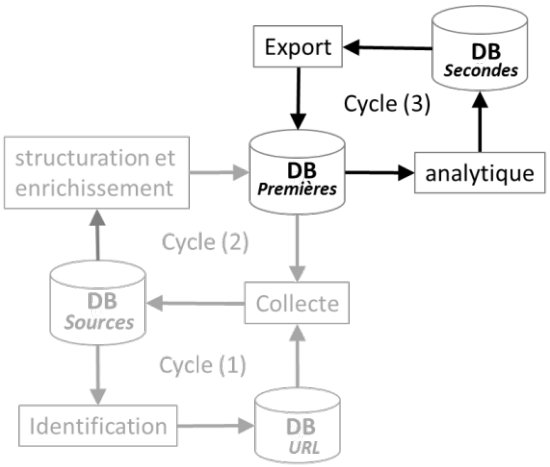

Figure 4. Cycle analytique et production de corpus (export)

Le troisième cycle (3) (figure 4) est celui de la production de résultats ou d'exportation des données utilisables dans un cycle externalisé de traitements pouvant conduire à la réimportation de corpus de données dans des univers plus classiques et familiers pour le sociologue (tableurs, suite logicielle R, etc.).

La réintroduction éventuelle des résultats de traitements comme enrichissement des données administrées en base de données est importante mais s'avère délicate à organiser. Elle suppose un effort important pour maintenir une interopérabilité sur les données qui passe par la définition stricte d'une interface contrôlant l'encodage, les formats et les structures de données externes et internes. L'harmonisation est également d'ordre méthodologique dans la mise en œuvre de métadonnées descriptives assurant la traçabilité des étapes externalisées. La manipulation des données entre différents environnements appelle une très grande vigilance de la part des chercheurs opérant sur les données, vigilance à laquelle ceux-ci ne sont pas préparés.

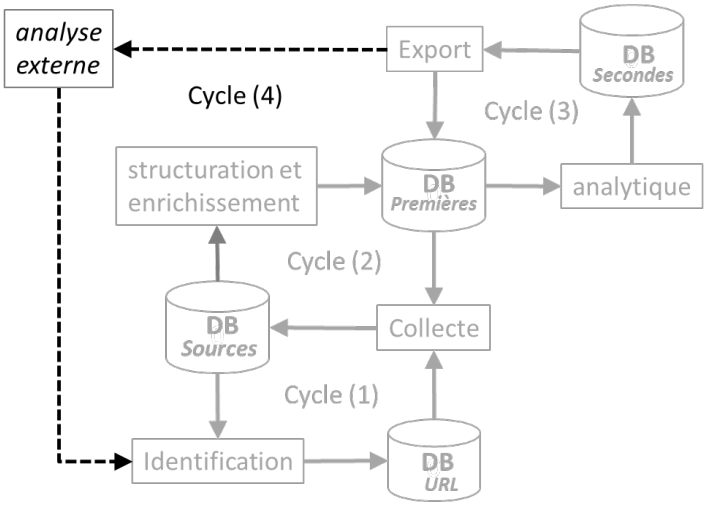

Figure 5. Cycle d'analyse externe

Enfin le dernier cycle (4) (figure 5) traduit la circularité de l'évolution au fil de l'eau du projet. Celle-ci peut influencer la nature des données attendues, les choix méthodologiques de leur raffinement progressif, voire la définition des collections appelant alors une nouvelle collecte. Nous lions formellement ce méta-cycle à l'exportation des données, étape à laquelle sont associés les outils de l'intelligibilité des données (dont la visualisation).

Interfaces disciplinaires

La contribution de chacun des acteurs à cette chaîne de traitement suppose une attribution des rôles qui s'est opérée naturellement au sein du projet. Elle a accordé *a priori* au sociologue le cycle (4) qui gouverne le projet d'étude et oriente la constitution des collections et des corpus. Ce cycle maintient la cohérence du projet dans son ensemble, sur le plan scientifique des objectifs de l'analyse et, en conséquence, sur la nature des données d'origine ou produites par le processus expérimental.

De son côté, le spécialiste de l'information et du web a pris en charge la mécanisation des traitements. Il a ainsi été investi d'une double autorité, celle de la création du dispositif et celle de sa mise en œuvre. Dissocier la conception de l'utilisation n'a pas été envisagé dans un régime de développement permanent afin de ne pas freiner la production de résultats. Compte tenu d'un principe de conception centré sur les données, cette absence de distinction reporte l'autorité du spécialiste de l'information et du web sur les données associées aux cycles (1) à (3).

Cette répartition séquentielle et alternée de plages d'interventions dans le processus d'élaboration des données fait apparaître deux interfaces (4-1) et (3-4) de synchronisation entre les perspectives respectives des deux chercheurs (figure 6).

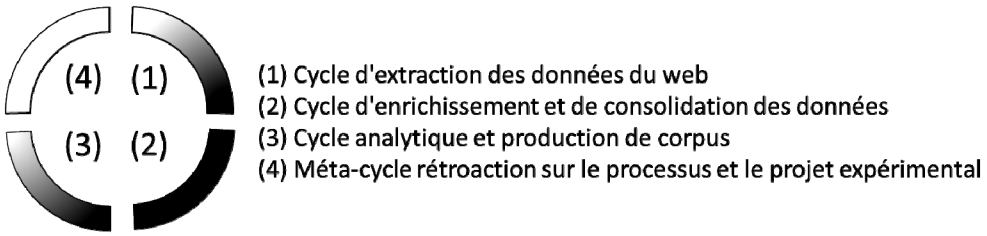

Figure 6. Attribution des rôles/autorités relatives aux cycles de traitements. Les nuances de couleur traduisent la « spécificité » des représentations associées aux données. Le cycle (2) correspond à la plus grande spécificité puisque répondant aux logiques « internes » du projet.

Les difficultés rencontrées dans l'interface (4-1) sont mineures. Elles concernent la compréhension des données du projet. Elles tiennent à la différence des points de vue portés sur la page web qui peut être considérée au travers de sa réception (lecture) ou de sa représentation (données). Ces écarts

se sont comblés rapidement dans l'élaboration d'un langage et d'un discours technique commun.

L'interface (3-4) constitue un espace de négociation beaucoup plus sensible dans lequel s'expriment des tensions méthodologiques et épistémologiques provenant des deux objectifs distincts : la définition d'un environnement stable pour les données *vs* la fourniture de jeux de données pour une étude ciblée. Le dialogue qui se noue alors sur la donnée mobilise deux définitions conceptuelles différentes. La première définition est issue d'une approche cybernétique qui privilégie la complétude et la cohérence interne de la représentation de la donnée. La seconde est associée à un usage heuristique et informationnel de celle-ci.

Donnée « variable » et donnée « trace »

Ce qui se décide à cette étape peut impliquer une rétro-ingénierie susceptible de transformer les structures de données comme les traitements. Les adaptations qui en découlent peuvent s'inscrire dans la continuité des travaux ou être en rupture par rapport aux choix techniques et aux développements déjà effectués. L'enrichissement des données – par exemple l'ajout d'une typologie de profils individuels calculée à partir de critères multiples – peut, selon les cas, amener à redéfinir la structuration des enregistrements et en conséquence à réviser l'ensemble des fonctions qui s'appuyaient sur la structure précédente. Ces adaptations techniques se justifient soit comme une réponse aux enjeux et à l'élaboration empirique du processus analytique soit, de manière orthogonale, comme la prise en compte de considérations architecturales qui peuvent être éloignées du projet initial.

Dans le cas présent, les négociations ne portent que sur la perspective commune du projet global d'analyse. Elles conduisent à des postures mettant en tension la production de résultats (le sociologue) et les conditions de celle-ci (le spécialiste de l'information). Cette différence d'approche provient des compétences et des rôles respectifs de chacun tels qu'ils se sont établis et confortés tout au long du projet. Ces tensions sont nécessaires au processus créatif de la démarche empirique. Elles contribuent en outre à l'élaboration d'un discours commun et à l'acculturation interdisciplinaire.

Les décisions prises dans ce contexte sont prescriptives. Elles suspendent l'avancée du projet à la capacité du spécialiste de l'information sur le web à produire une réponse dans des délais compatibles avec ceux du sociologue. Elles interfèrent également dans l'organisation de l'activité du développeur. Conjointement, si les délais sont clairement assumés par le sociologue, son absence de maîtrise sur l'ingénierie des données lui donne le sentiment de ne pas avoir prise sur le protocole expérimental dans son intégralité. L'articulation des cycles (2) et (3) révèle cette tension sur les données : entre production nécessaire et sophistication superflue.

La perception des processus en œuvre est de fait très différente pour chacun des protagonistes : elle est associée au processus d'élaboration des données plutôt linéaires pour le sociologue ; elle s'apparente à celui de la mise au point non linéaire pour le spécialiste de l'information sur le web. Cette différence de perception rejoint une autre expression de la durée se rapportant au cycle (2) opposant au temps long de l'archive le temps court du corpus. Le temps long est celui d'une mémoire qui peut être institutionnelle ou celui de la mise à disposition des éléments de la preuve. Le temps court est celui de la production et de la valorisation immédiate des résultats. Cette question met en évidence la différence du rapport à la « donnée » du sociologue et du spécialiste des technologies de l'information mobilisés ensemble sur ce projet. D'un côté, la donnée est définie dans sa capacité à pouvoir être transformée en « variable » ; de l'autre, elle apparaît comme une énonciation singulière dont il faut conserver la trace et les spécificités[15].

LE RÉGULATEUR PUBLIC ET LES DONNÉES *PERSONNELLES*

Un quatrième acteur intervient dans ce projet de datafication : le régulateur public, terme par lequel nous désignerons à la fois le législateur qui définit le cadre d'utilisation des données informatiques en France depuis la loi dite « Informatique et liberté » de 1978 et la CNIL (Commission Nationale Informatique et Liberté) qui veille à l'application de cette loi. En l'absence de régulations internes aux organismes finançant la recherche publique – à la différence des *review boards* américains par exemple (Feeley, 2007) –, des recherches peuvent être menées sur le web sans préoccupation particulière de nature déontologique ni prise en compte *a priori* de leur adéquation au cadre légal. Il est pourtant important de considérer le régulateur public comme un participant invisible au laboratoire de la *datafication* sur le web.

Celui-ci peut en effet se rendre visible et influer sur le déroulement de l'enquête soit de sa propre initiative soit, comme ce fut le cas ici, à l'initiative des chercheurs souhaitant faire valider leur processus de création de base de données.

Le masque que le régulateur public pose sur les inscriptions individuelles consiste à les qualifier de « données personnelles ». C'est en effet le premier

15 Les frictions précédentes apparaissent à l'intérieur d'une bulle collaborative qui s'est constituée afin d'assumer une autonomie productive et une perspective méthodologique commune. Elles se résolvent dans un équilibre relationnel qui se fonde sur un fonds commun, issu d'une pratique de la recherche dans un laboratoire en Sciences sociales au sein duquel s'hybrident les disciplines depuis plusieurs années. Dans le contexte de cette bulle, la sujétion disciplinaire vécue par chacun des chercheurs est atténuée mais elle demeure notamment dans l'évaluation des travaux. Celle-ci impose une traduction différente de l'activité collaborative et des résultats produits.

critère qui justifie leur protection au sens de la loi de 1978 dont le deuxième article définit la donnée personnelle 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 »[16]. Ce masque est extrêmement large dans la mesure où il recouvre aussi bien des données directement identifiantes (comme un nom par exemple) que des données indirectement identifiantes (comme un numéro de compte) ou des recoupements d'information susceptibles d'amener à l'identification de personnes. Dans le cas de notre enquête, il est évident que les inscriptions des individus sont vues par le régulateur à travers le masque des « données personnelles ». Elles contiennent des éléments d'identification directe d'une part. D'autre part, même si le chercheur s'engage à ne pas collecter cette inscription, la combinaison des autres inscriptions peut conduire théoriquement à une identification.

Cette définition entre elle aussi en tension avec le masque appliqué par le chercheur qui tend à considérer ces inscriptions comme des données d'enquête produites et rendues visibles par les individus eux-mêmes. Les inscriptions individuelles peuvent en effet – si l'on considère qu'elles sont produites pour être visibles – êtres vues à la fois comme des *artefacts* (et pas des données qui, comme le nom ou un numéro de carte d'identité sont imposés aux individus) et des informations *publiques* (chaque individu sait que les inscriptions qu'il dépose sur la plateforme seront visibles par tous sur le web et il peut s'il le souhaite limiter cette visibilité dans les paramètres de son compte). Mais le régulateur pour sa part ne prend aucunement en compte l'intention des individus et considère que son devoir de protection couvre aussi les informations volontairement rendues publiques par les intéressés. Ce devoir de protection est donc fondé sur une définition des données concurrente aussi bien de celle du chercheur que de celle de l'individu.

Il se fonde aussi sur un principe dit de pertinence qui prescrit à la fois de ne pas collecter plus de données que ce qui est nécessaire (ce qui nous amènera par exemple à ne pas collecter le nom des individus) et à prendre un soin particulier des données considérées comme plus sensibles que d'autres. Ce point provoque à nouveau une tension car il est difficile de définir matériellement la « sensibilité » des données. Dans le cas de cette enquête, notre discussion avec le régulateur a notamment porté sur un point que nous n'avions à aucun moment envisagé : la question de savoir si nos données permettaient d'identifier

16 La loi définit aussi dans cet article le « traitement de données à caractère personnel » et le « fichier de données à caractère personnel » : « Constitue un traitement de données à caractère personnel toute opération ou tout ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction.
Constitue un fichier de données à caractère personnel tout ensemble structuré et stable de données à caractère personnel accessibles selon des critères déterminés. »

les préférences politiques des individus. Là encore c'est donc une nouvelle forme d'ouverture du laboratoire de la recherche dans la mesure où on peut traduire cette interrogation sous la forme d'une hypothèse scientifique. Pour le régulateur, la trajectoire d'un individu peut dénoter des opinions (par exemple on peut inférer du fait qu'un individu a été journaliste à *L'Humanité* une sensibilité politique proche du parti communiste). Pour le sociologue, cette hypothèse est assez fragile du fait des évolutions du marché du travail journalistique et ne concerne que très peu de situations. Mais l'hypothèse scientifique du régulateur a pour elle la force du droit, ce qui nous conduira à la prendre en compte en entamant des démarches auprès des syndicats de journalistes pour vérifier que l'enquête ne soulevait pas d'objections de principe sur ce type d'inférences d'opinion[17].

CONCLUSION

Les sciences sociales sont encore mal équipées pour réfléchir aux implications épistémologiques de la recherche sur le web. Elles se sont largement constituées dans un rapport conflictuel aux données administratives dans les années 1960. L'étymologie même du terme « statistique » en atteste. Elle renvoie autant à l'exercice de l'État qu'à la science mathématique. De même, la représentation statistique du monde social par « enquête directe » s'est longtemps opposée à l'utilisation de registres administratifs issus d'une activité administrative de tenue de fichiers individuels (Desrosières, 2005). L'exemple que nous avons utilisé dans cet article montre bien qu'il est aujourd'hui nécessaire d'intégrer dans la réflexion en sciences sociales d'autres acteurs, au premier rang desquels les entreprises du « *knowing capitalism* » (Thrift, 2005) et, par ricochet, les usagers de leurs services et les organes de régulation qui tentent de réglementer leur activité comme celle du sociologue par la même occasion.

Si l'on adopte le point de vue réflexif de Burrows et Savage dans leur fameux article sur la « *coming crisis* » de la sociologie empirique, celle-ci serait en train de perdre sa « juridiction » sur tout un pan de la connaissance de la société. L'entretien et l'enquête par questionnaire qui lui ont longtemps assuré cette juridiction sont dépassés par de nouveaux modes de représentation de la société fondés sur la corrélation plus que la causalité, explorant la dimension prédictive du *big data* et fondées sur la commodification des données personnelles. Les *datasets* se multiplient mais ils sont le « *digital by-product of the routine operations of a large capitalist institution* » (Savage & Burrows, 2007).

17 La démarche permettait aussi de rendre à nouveau publique l'enquête. Cet impératif de publicité s'imposait du fait de l'impossibilité d'informer chaque individu dont les inscriptions étaient collectées à la fois de cette collecte et de ses droits d'opposition ou de rectification. Le courrier envoyé aux principaux syndicats de journalistes est resté lettre morte…

Cette révolution en cours – et les frictions qui l'accompagnent du fait de la perte de la juridiction des sciences sociales sur la définition du bon masque à poser sur les inscriptions des individus pour les rendre lisibles comme des données – met évidemment l'accent sur le fait que les données ne sont pas plus « données » sur le web qu'ailleurs. Il nous semble que dans un monde dans lequel « *digital devices and the data they generate are both the material of social lives and form part of many of the apparatuses for knowing those lives* » (Rupert, Law, & Savage (2013), notre travail consiste finalement à essayer de dégager un assemblage de ces définitions contradictoires des « données » qui a) soit acceptable – techniquement et socialement – par toutes les parties prenantes à cette recherche (y compris celles qui se sont invitées tardivement) et b) produise *in fine* quelque chose que nous pouvons (mais *ex post* seulement) appeler une « base de données ».

Remerciements

La recherche mentionnée dans cet article a été financée par une subvention de la Commission scientifique de Sciences Po Grenoble. Les auteurs remercient les coordinateurs du numéro, la Revue d'anthropologie des connaissances et les participants au workshop de Lausanne en octobre 2015 pour la discussion critique de la première version de l'article.

RÉFÉRENCES

Abbott, A. (2000). Reflections on the Future of Sociology. *Contemporary Sociology*, *29*, 296-300.

Akrich, M. (1987). Comment décrire les objets techniques. *Techniques et culture*, *9*, 49-64.

Bastin, G. (2015). Analyser les carrières de journalistes dans les mondes de l'information : propositions pour une enquête indirecte sur le réseau LinkedIn, in Christine Leteinturier & Cégolène Frisque (dir.), *Les espaces professionnels des journalistes. Des corpus quantitatifs aux analyses qualitatives.* Paris : Éditions Panthéon-Assas, 203-228.

Bastin, G. (2012). Le cas Mathieu ou l'entretien renversé, *Sur le Journalisme / On Journalism / Sobre Jornalismo*, *1*, 40-51.

Bastin, G. & Bouchet-Valat, M. (2014), Media corpora, text mining, and the sociological imagination – A free software text mining approach to the framing of Julian Assange by three news agencies using R.TeMiS, *Bulletin de méthodologie sociologique*, *121*(1), 5-25.

Berry, M. (1983). *Une technologie invisible ? L'impact des instruments de gestion sur l'évolution des systèmes humains.* Paris : École Polytechnique.

Bourdieu, P., Chamboredon, J.-C., & Passeron, J.-C. (1969). *Le métier de sociologue.* La Haye : Mouton.

Cardon, D. (2008). Le design de la visibilité : un essai de typologie du web 2.0. *InternetActu*, février.

Cardon, D. (2012). Regarder les données. *Multitudes*, *49*(2), 138-142.

Cukier, K. & Mayer-Schönberger, V. (2014). *Big data. la révolution des données est en marche.* Paris : Robert Laffont.

Dagiral, E. & Peerbaye, A. (2012). Les mains dans les bases de données : connaître et faire reconnaître le travail invisible, *Revue d'anthropologie des connaissances*, *6*(1), 191-216.

Desrosières, A. (1993). *La politique des grands nombres. Histoire de la raison statistique*. Paris : La Découverte.

Desrosières, A. (2014). *Prouver et gouverner. Une analyse politique des statistiques publiques.* Paris : La Découverte.

Desrosières, A. (2005), Décrire l'État ou explorer la société : les deux sources de la statistique publique. *Genèses*, *58*, 4-27.

Durkheim, É. (1895). *Les règles de la méthode sociologique.* Paris : Presses universitaires de France.

Feeley, M. M. (2007). Legality, social research, and the challenge of institutional review boards. *Law & Society Review*, *41*(4), 757-776.

Flichy, P. (2013). Rendre visible l'information. Une analyse sociotechnique du traitement des données. *Réseaux*, *178-179*, 55-89.

Giddens, A. (1987). *Social Theory and Modern Sociology*. Cambridge: Cambridge, Polity Press.

Hey, T. & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox, & T. Hey (eds.), *Grid Computing: Making the Global Infrastructure a Reality*. New York: Wiley & Sons, 809-824.

Passeron, J.-C. (1991). *Le raisonnement sociologique. L'espace non popperien du raisonnement naturel.* Paris : Nathan.

Perrot, J.-C. (1992). *Une histoire intellectuelle de l'économie politique, XVII^e^-XVIII^e^ siècle*. Paris : Éditions de l'Ecole des hautes études en sciences sociales.

Law, R. & Savage, M. (2013). Reassembling Social Science Methods: The Challenge of Digital Devices. *Theory, Culture & Society*, *30*(4), 22-46.

Savage, M. & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885-899.

Star, S., (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, *43*(3), 377-391.

Thrift, N. (2005). *Knowing Capitalism*. London: Sage.

Wiszniewski, D. & Coyne, R. (2002). Mask and Identity: The Hermeneutics of Self-Construction in the Information Age, in K. A. Renninger & W. Shumar (eds.), *Building Virtual Communities. Learning and Change in Cyberspace* (pp. 191-214). Cambridge UP.

Gilles BASTIN est Maître de conférences habilité à diriger des recherches à Sciences Po Grenoble et membre du laboratoire Pacte (CNRS / Université Grenoble Alpes / Sciences Po Grenoble). Ses recherches portent sur la sociologie des trajectoires d'individus dans les mondes de l'information et la morphologie de ces mondes, la théorie sociologique et l'histoire des sciences sociales, en particulier dans leur relation avec l'analyse des médias, les aspects méthodologiques, politiques et éthiques de l'usage des données dans les sciences sociales, notamment des données tirées des réseaux sociaux et du web.

Adresse	Sciences Po Grenoble PACTE F-38000 Grenoble (France)
Courriel	gilles.bastin@sciencespo-grenoble.fr

Jean-Marc FRANCONY est Maître de conférences en Sciences de l'Information et de la Communication à l'Université Grenoble Alpes et membre du laboratoire Pacte (CNRS / Université Grenoble Alpes / Sciences Po Grenoble). Ses recherches portent sur l'analyse des dispositifs info-communicationnels et des pratiques sociales à l'occasion d'événements médiatiques majeurs. Il s'est particulièrement investi dans les problématiques de l'observation expérimentale et de la constitution de collections de traces numériques d'usages.

Adresse	Université Grenoble Alpes PACTE F-38000 Grenoble, France
Courriel	jean-marc.francony@umrpacte.fr

Abstract: Inscriptions, Masks and Data. Datafication of the web and interpretation conflicts around data in the invisible laboratory of social sciences

This paper deals with the interactions that occur during a research project based on the exploitation of social media data gathered on LinkedIn. Our reflexive approach to what we call the invisible laboratory of the research project allows us to describe the conflicts of interpretations that characterizes those data and the definition of data on the web. The production of our dataset is described as a social process involving the transformation of users "inscriptions" into sociological data. We call this process "datafication" and study the many actors that are involved in it: the sociologist, the IT and web specialist, the LinkedIn user, the social media and the public body in charge of data privacy regulations.

Keywords: datafication, LinkedIn, Research, Database.

Resumen: La inscripción, la máscara y los datos. *Datafication* de la web y conflictos de interpretación en torno a los datos en un laboratorio oculto de ciencias sociales

En este artículo se reflexiona sobre las diferentes interacciones que caracterizan a un proyecto de encuesta en ciencias sociales sobre la explotación de los datos de la red social profesional LinkedIn. Esta reflexión sobre lo que llamamos el "laboratorio invisible" de la encuesta permite destacar los conflictos de interpretación que surgen en torno a la definición de lo que es

un dato tomado de la web. Por lo tanto, el establecimiento de una base de datos no debe aparecer como una operación técnica de la investigación, sino como un proceso de transformación de "inscripciones" individuales en la red en "datos", un proceso que llamamos "*datafication*". Este proceso implica la confrontación de "máscaras" que se ponen los actores del laboratorio invisible por encima de aquellas inscripciones: el sociólogo y el especialista de la información y de la web, sino también el usuario de la red, la plataforma y el regulador público.

Palabras clave: datafication, linkedin, encuesta, base de datos.