Christine L. Borgman

Qu'est-ce que le travail scientifique des données ?
Big data, little data, no data

OpenEdition Press

6. Le travail scientifique des données dans les sciences sociales

DOI : 10.4000/books.oep.14762
Éditeur : OpenEdition Press
Lieu d'édition : OpenEdition Press
Année d'édition : 2020
Date de mise en ligne : 18 décembre 2020
Collection : Encyclopédie numérique
ISBN électronique : 9791036565410

6. Le travail scientifique des données dans les sciences sociales

Introduction

Les sciences sociales recouvrent la recherche sur le passé, le présent et l'avenir du comportement humain. En dépit de leur longue histoire, ces champs sont attaqués dans les universités comme au-dehors. Le débat sur les « deux cultures » des sciences et des lettres lancé par l'essai de C. P. Snow (1956) a fait rage tout au long des années 1960, mais a largement laissé les sciences sociales à la marge de cette dichotomie. Malgré la montée en puissance des collaborations et des parcours interdisciplinaires, les divisions entre domaines se sont creusées avec la transformation des structures de financement et de rémunération (Hollinger, 2013). Dernièrement, la politologie est devenue la science sociale la plus politisée, au point que le Congrès américain a suspendu le financement public de la discipline sauf pour les travaux qui « promeuvent la sécurité nationale ou les intérêts économiques des États-Unis » (Prewitt, 2013). Les réactions, tant aux États-Unis qu'à l'étranger, expriment une profonde inquiétude quant à l'immixtion de la politique dans l'attribution de subventions individuelles, quelle que soit la discipline. L'évaluation par les pairs (*peer review*) et l'expertise scientifique sont en jeu et la pression pour des résultats immédiats s'accentue au détriment de la construction théorique à long terme (P. Boyle, 2013 ; Prewitt, 2013).

Certains des problèmes rencontrés par la recherche en sciences sociales sont attribuables aux difficultés dans la gestion des données. Par exemple, en tentant de répliquer une étude économique influente, un étudiant y a découvert des erreurs de calcul. Celles-ci – et la réponse que leur ont apportée les auteurs – ont fait la une des revues scientifiques comme de la presse économique, contribuant à interroger la fiabilité de la recherche (Marcus, 2013 ; Monaghan, 2013 ; Wiesenthal, 2013). Un cas de fraude par un professeur de psychologie sociale néerlandais est passé inaperçu des années durant, soulevant des inquiétudes semblables sur la recherche et l'efficacité de l'évaluation par les pairs dans cette discipline (Enserink, 2012a ; Shea, 2011). Le taux de rétractation en économie et dans les autres champs des sciences sociales est certes nettement moindre que dans les sciences exactes, mais cela n'indique pas pour autant une plus grande intégrité. Les différences seraient plutôt liées aux manières dont les maisons d'édition de chaque discipline gèrent les propositions comportant des erreurs, des plagiats ou de la fraude (Karabag et Berggren, 2012 ; Oransky, 2012).

Les méthodes traditionnelles d'échantillonnage, comme le recours au courrier postal ou aux appels aléatoires sur des téléphones fixes, sont moins fiables depuis que les activités de communication se font en ligne. D'aucuns affirment que l'intérêt des sondages et des autres méthodes conventionnelles de recherche en sciences sociales touche à ses limites et que de nouvelles approches sont nécessaires (Savage et Burrows, 2007, 2009). Les politiques de protection des sujets de recherche sont elles aussi en pleine transformation. Un placard fermé à clé ne suffit plus à protéger les dossiers numériques recueillis sur des personnes ; les nouvelles méthodes doivent répondre à la réalité de la protection de la vie privée, du *data mining* et de la réidentification. La recherche sur des sujets sensibles tels que le terrorisme et les conflits est essentielle à la politique publique, mais la mener tout en préservant la confidentialité – et la vie – des sujets oblige à des choix difficiles (Jackson *et al.*, 2013). Étant donné la complexité du comportement humain et des institutions sociales, il est souvent très difficile d'établir une causalité dans les sciences sociales. Ces disciplines sont en quête de nouvelles méthodes et sources et, ce faisant, se confrontent aux promesses et aux écueils du travail scientifique des données.

Les méthodes de recherche et les pratiques en matière de données

Les pratiques en matière de données d'un domaine font partie intégrante de ses méthodes de recherche ; or, les sciences sociales explicitent bien plus leurs méthodes que les autres champs. Les manuels méthodologiques abondent pour guider les programmes universitaires sur le protocole de recherche, les statistiques, la recherche quantitative et qualitative et la visualisation. L'équilibre entre une description la plus riche possible du comportement humain et la nécessité de respecter les droits des individus, des groupes et des institutions étudiés est au cœur de ces méthodes.

Quelques dichotomies de base – quoique réductionnistes – suggèrent la palette des méthodes possibles. Ces dimensions ne sont pas incompatibles et peuvent se combiner de différentes manières. La première dichotomie met en regard l'explication idiographique et l'explication nomothétique. Les études idiographiques sont particulières à un lieu, à une condition ou à un événement. Cette démarche cherche à décrire et expliquer un cas donné de la manière la plus complète possible. À l'inverse, les études nomothétiques identifient des facteurs causaux qui influencent une classe d'événements ou de conditions (Babbie, 2013). La deuxième dichotomie distingue les méthodes reposant avant tout sur des techniques de décompte et de calcul pour étudier les problèmes sociaux, dites « quantitatives », de celles qui ont recours à l'interprétation, dites « qualitatives ». La troisième différencie les méthodes interventionnistes (*obtrusive*) et non interventionnistes (*unobtrusive*).

Les méthodes interventionnistes sont celles de travaux où il y a un certain degré d'interférence : le sujet de recherche est conscient d'être étudié, de préférence avec son consentement. Les méthodes non interventionnistes, quant à elles, ne supposent pas d'intervention : le chercheur ou la chercheuse travaille sur des traces d'activité humaine ou observe un comportement sans interférer.

Les notions de fiabilité et de validité transcendent ces dichotomies. La fiabilité correspond à la constance, à la probabilité que des observations répétées d'un même phénomène produisent le même résultat. La validité est la valeur de vérité, le degré auquel une mesure saisit le concept qu'elle est censée mesurer (Babbie, 2013 ; Shadish *et al.*, 2002).

Les scientifiques arbitreront différemment entre ces dimensions pour chaque étude. Les sondages, par exemple, sont généralement nomothétiques, quantitatifs et exigent un vaste échantillon pour être suffisamment fiables. À l'inverse, les ethnographies sont le plus souvent idiographiques, qualitatives, interventionnistes et s'inquiètent davantage de leur validité que de leur fiabilité. Les études ayant recours au *big data* nécessitent des méthodes statistiques et, parfois, des modélisations informatiques. Ces méthodes sont à même de produire des données anonymisables et réutilisables par d'autres. En revanche, les données qui reposent sur une analyse fouillée sont propres à engendrer des descriptions fécondes de phénomènes, mais les données peuvent s'avérer impossibles à anonymiser ou à partager. Le plus souvent, il faut faire des choix difficiles quant aux données pouvant être obtenues ou non, comment et pourquoi, ainsi que sur les façons de les rapporter et de les diffuser.

Études de cas en sciences sociales

Le travail scientifique des données est tout aussi divers dans les sciences sociales que dans les sciences exactes ; nous n'essayerons donc pas d'être exhaustifs. Les études de cas du présent chapitre abordent les raisons et les manières dont les gens utilisent les technologies de l'information, illustrant les dimensions des méthodes expliquées plus haut. La première série de cas, sur les enquêtes sur Internet et les médias sociaux comme c'est d'usage, compare des sondages, où les enquêtés évoquent leur usage d'Internet, avec des traces de leur comportement réel sur des technologies telles que Twitter. OxIS, petit nom de l'Oxford Internet Survey of Britain, est menée deux fois par an par l'Oxford Internet Institute depuis 2003 au moyen d'entretiens personnels. Plusieurs études qui utilisent les contenus de Twitter et d'autres services de microblogage comme ressources en données éclairent la manière dont les décisions relatives aux preuves et aux méthodes influent sur les pratiques en matière de données et les résultats.

La seconde étude de cas de ce chapitre s'intéresse à la manière dont les technologies de l'information sont conçues, mises en place et utilisées dans la recherche en science et en technologie. Les méthodes sociotechniques sont employées pour étudier les pratiques en matière de données du Center for Embedded Networked Sensing (Centre pour la télédétection intégrée en réseau). Les résultats de ces travaux sont rapportés dans l'étude de cas sur les pratiques en matière de données du CENS au chapitre 5. Dans le présent chapitre, nous résumons les méthodes employées dans une recherche sociotechnique sur dix ans et nous examinons leurs implications dans le travail scientifique des données. Ces méthodes sont largement idiographiques, mais comportent également une analyse de réseaux sociaux, ainsi que l'élaboration et l'évaluation de technologies. Ensemble, ces deux études de cas examinent la manière dont les spécialistes des sciences sociales abordent des méthodes et des problèmes nouveaux et les implications qui en découlent pour les infrastructures de la connaissance.

Les enquêtes sur Internet et l'étude des médias sociaux

Les études d'Internet tirent leurs méthodes d'autres domaines des sciences sociales. Les sondages sont une façon habituelle de poser les mêmes questions à un grand nombre de gens. Ils peuvent être menés en personne, par courrier, par e-mail ou en utilisant d'autres technologies comme des questionnaires sur le Web et des applications mobiles. L'analyse de réseaux sociaux, une méthode couramment employée dans la recherche sur Internet, est bien antérieure au Web et aux médias sociaux d'aujourd'hui comme Twitter, Facebook, LinkedIn, Flickr et Pinterest. Les sociologues modélisent des relations entre individus et groupes depuis les années 1920 environ, en ayant recours à tous les indicateurs disponibles : cartes postales, appels téléphoniques, listes de membres et autres liens sociaux (Freeman, 2004 ; Wellman et Haythornthwaite, 2002).

Pour toute méthode de recherche, il est nécessaire de disposer d'un savoir-faire scientifique pour concevoir l'étude, sélectionner la population à étudier et les méthodes d'échantillonnage et nettoyer, analyser et interpréter les données obtenues. Le danger de l'utilisation du *big data* dans les sciences sociales et ailleurs consiste à croire qu'on analysera les données aussi facilement qu'on les a obtenues. Or, mener des enquêtes et des études sur les médias sociaux est bien plus difficile qu'il n'y paraît. Certains travaux sur Internet sont très sophistiqués et prennent soigneusement en compte les limites de ces données en matière de fiabilité et de validité. D'autres sont naïfs et exploitent des flux de données comme d'intéressantes sources de preuves sans bien en comprendre les limites (Boyd et Crawford, 2012).

La taille compte

La mégascience ou science lourde (*big science*) au sens de Weinberg (1961) et Price (1963) se caractérise par la maturité du champ de recherche et par la sophistication des méthodes. La recherche par sondage, qui documente les tendances de la société depuis fort longtemps, est sans doute l'un des domaines les plus mûrs des sciences sociales. Des archives telles que l'Inter-University Consortium for Political and Social Research (2013) recueillent des données de sondage depuis plus de cinquante ans, originellement sur papier, puis sous forme numérique. Les archives de données des sciences sociales ont différentes priorités selon leur source de financement, le type d'étude, la région et d'autres critères. Les centres de recherche universitaires comme l'Institute for Quantitative Social Science à Harvard et la Science Data Archive à l'université de Californie à Los Angeles (UCLA) financent des référentiels de données, des didacticiels et instructions, le développement d'outils et d'autres services (Institute for Quantitative Social Science, 2013 ; Social Science Data Archive, 2014).

On peut effectuer des sondages de manière ponctuelle, mais ils s'avèrent particulièrement utiles quand ils sont menés sur de longues périodes. Des sondages d'opinion sur la politique, l'enseignement supérieur et les mentalités en général sont réalisés à intervalles réguliers depuis des décennies. En posant les mêmes questions clés à chaque fois, la comparabilité est maintenue. Les enquêtes sont adaptées pour répondre aux problématiques sociales du moment en ajoutant ou modifiant chaque fois quelques questions. Après qu'elles ont été déposées dans des archives, d'autres scientifiques peuvent en réanalyser les données, comparer les différentes études ou répliquer – entièrement ou partiellement – le protocole de recherche sur des populations différentes. Cependant, réutiliser ces données suppose un investissement considérable pour comprendre comment et pourquoi chaque enquête a été menée, en interpréter les résultats et déterminer quelles composantes peuvent être réutilisées ou comparées.

Les données des médias sociaux sont massives par leur volume absolu et peuvent produire dix fois, cent fois, mille fois, voire des millions de fois plus d'observations que les sondages, les entretiens ou les études en laboratoire. Suivre les communications quotidiennes de personnes au moyen de journaux intimes et d'autres méthodes manuelles, par exemple, ne fournit qu'un mince filet de traces. En revanche, s'il est difficile d'obtenir des chiffres fiables sur les transactions dans les médias sociaux, les volumes peuvent être colossaux. Ainsi, un des principaux fournisseurs de données de médias sociaux pour les entreprises affirme disposer d'un flux de trois milliards d'activités par jour (Gnip, 2013b).

Quand est-ce une donnée ?

Le déluge de données tiré de l'enregistrement numérique de l'activité humaine peut s'avérer une mine d'or pour les chercheurs et chercheuses en sciences sociales. La surabondance de sources et de ressources, à la fois contemporaines et historiques, amène toutefois son lot de difficultés. Découvrir ces richesses revient parfois à suivre une carte au trésor de pirate sur un chemin semé d'embûches, de chausse-trapes, de faux indices et d'indications trompeuses. La quête scientifique consiste à déterminer quelles entités peuvent validement mettre en évidence quels phénomènes. Souvent, les chercheurs et chercheuses sont déchirés entre les données qu'ils veulent et les données qu'ils peuvent obtenir. Les meilleurs travaux de recherche sont ceux qui adaptent au mieux leurs méthodes aux questions posées. Des protocoles novateurs peuvent mener à des découvertes sensationnelles, mais s'avérer difficiles à répliquer ou à expliquer à celles et ceux qui n'appartiennent pas à la discipline. Des méthodes nouvelles peuvent aussi produire des données plus difficiles à documenter, partager, réutiliser et conserver comparées à celles issues de protocoles traditionnels.

Sources et ressources

Lorsque des scientifiques recueillent leurs propres données, elles et ils ont davantage de contrôle sur leurs protocoles de recherche. Pour mener des travaux à plus grande échelle, ils devront souvent rassembler des données tirées de ressources externes. Beaucoup combineront différentes méthodes, en réalisant des entretiens ou des ethnographies qu'ils compléteront de traces issues de médias sociaux, de transactions économiques, de recensement ou autres.

Les sondages sont généralement nomothétiques par nature : ils mesurent un petit nombre de variables de manière à pouvoir en tirer des comparaisons sur de vastes populations. Les scientifiques contrôlent le plus de sources de variances possible en indiquant précisément les questions à poser, la population à interroger et le plan de sondage. Elles et ils savent ce qu'ils demandent, à qui et ce qu'ils comptent apprendre de chaque question. Ils contrôlent également la manière dont les questions sont posées en formant les sondeurs et sondeuses à s'adresser aux participants de manière cohérente, à formuler précisément les questions et à noter les réponses. Le protocole de recherche doit équilibrer la validité interne et externe, c'est-à-dire respectivement le degré auquel l'étude contrôle les variables pour isoler les phénomènes et le degré auquel l'étude est généralisable à des populations plus vastes et diverses (Shadish *et al.*, 2002). Ces choix méthodologiques influent sur la capacité à se fier aux sources de données, à déterminer la provenance, à employer des outils d'analyse et à interpréter les résultats.

Les sondages en ligne sur le Web peuvent toucher des populations bien plus importantes que les entretiens personnels, mais obtenir un échantillonnage cohérent est difficile et les taux de réponse sont souvent bas. La fiabilité et la validité tendent à être plus élevées dans les enquêtes qui ont recours à des enquêteurs et enquêtrices humains pour s'adresser en personne aux participants, mais elles sont aussi bien plus coûteuses. Lorsque les chercheurs et chercheuses recueillent leurs propres données, ils peuvent parfois contrôler la variance en comparant leurs sources aux ressources d'autrui. Les instituts de sondage interrogent généralement 1 500 personnes environ, en privilégiant le téléphone plutôt que le face-à-face. Les enquêtes peuvent être étalonnées grâce à de telles sources, comme le sont les instruments de télescopes et de réseaux de capteurs.

Twitter a pris une place importante dans la recherche en technologie en raison de son usage international et de ses années d'existence, qui rendent possibles des analyses longitudinales. La plateforme permet d'envoyer de brefs messages – jusqu'à 140 caractères[1] – appelés « tweets ». Ces 140 signes peuvent contenir des informations supplémentaires grâce à des abréviations et des liens, lesquels peuvent être condensés grâce à des services de réduction d'URL. Les tweets peuvent également comporter des informations temporelles et des coordonnées géospatiales s'ils ont été envoyés depuis des appareils mobiles où ces fonctions sont activées. On peut y joindre des photos et d'autres images. Les comptes Twitter peuvent être nommés ou anonymes. Beaucoup appartiennent à des personnalités publiques, à des entreprises ou à des organisations grandes ou petites. Il n'y a pratiquement pas de limite à ce qu'un tweet peut dire : les particuliers évoquent leur journée, les scientifiques commentent les découvertes du moment, les entreprises présentent leurs nouveaux produits, les activistes motivent leurs troupes et les bibliothèques annoncent les modifications de leurs horaires et de leurs services. Les données de Twitter sont utilisées pour étudier les relations de communications, la santé publique, les événements politiques, les phénomènes linguistiques et le cours de la bourse, entre autres sujets (Bollen *et al.*, 2010 ; Bruns et Liang, 2012 ; Collins, 2011 ; Eysenbach, 2011 ; Murthy, 2011 ; Ozsoy, 2011 ; Shuai *et al.*, 2012 ; Simonite, 2013 ; Thelwall *et al.*, 2013 ; Zappavigna, 2011).

La variété des entités pouvant servir de données dans les flux Twitter et dans chaque tweet en fait une ressource attractive. Cependant, sa faiblesse réside dans la difficulté d'obtenir des données valides et fiables. À mesure que la valeur marchande des contenus de Twitter est apparue et que les préoccupations pour la vie privée ont émergé, les flux ont été rendus moins accessibles aux scientifiques. L'accès à l'ensemble du « flot » est limité, ce qui rend l'échantillonnage difficile. La distribution

1. NdT : 280 caractères depuis 2017.

des comptes Twitter et de leur activité par âge, sexe, race, pays, orientation politique, revenu et autres paramètres n'est pas homogène ; un échantillon d'utilisateurs ou utilisatrices ou de tweets ne constitue donc pas nécessairement une représentation fidèle de la population que le chercheur ou la chercheuse souhaite étudier. Cependant, la plus grande menace à la validité des tweets comme indicateurs d'activité sociale est l'évolution de l'usage des services en ligne. Une part croissante des comptes Twitter est constituée de robots servant à influencer la communication publique. Un compte peut ainsi attirer un grand nombre d'abonnés moyennant finance. Seulement 35 % des abonnés Twitter seraient de vraies personnes et jusqu'à 10 % de l'activité des médias sociaux proviendrait de comptes robotiques (Furnas et Gaffney, 2012 ; Urbina, 2013).

Les infrastructures de la connaissance

La diversité des ressources en données et de l'instrumentation dans les sciences sociales rend peu probable l'émergence d'une infrastructure de la connaissance comparable à celle de l'astronomie. Le comportement humain ne se prête pas à des descriptions normalisées comme le font les astres ou le spectre électromagnétique ; par ailleurs, l'échelle de coordination des télescopes spatiaux n'est pas transposable aux projets de ces disciplines. Les infrastructures de la connaissance de ces champs n'en comportent pas moins des outils et une expertise partagés. Une formation méthodologique, dispensée dans les cours de sciences sociales du second cycle, est essentielle au partage des méthodes et des données. Les cours de méthodologie abordent généralement l'utilisation d'outils analytiques, comme les progiciels statistiques, les logiciels de cartographie et les systèmes de codage qualitatifs.

Les archives des sciences sociales, que nous avons évoquées plus haut, comptent parmi les ressources les plus riches en données dont dispose la recherche par sondage. La collecte et la conservation de traces sociales sont une pratique bien antérieure à la recherche sur Internet (Boruch, 1985). Les registres de recensement, par exemple, prennent de la valeur avec les années, car ils constituent des observations sur un lieu et une époque. Le *Domesday Book*, le registre du recensement mené en Angleterre en 1085 et 1086 par Guillaume le Conquérant, est aujourd'hui considéré comme « le trésor le plus précieux de Grande-Bretagne » par les National Archives (2013). En 2011, le livre a été mis en ligne et doté d'une fonction de recherche des noms et des lieux mentionnés. Les registres de recensement contemporains, ainsi qu'une vaste gamme de traces sociales et institutionnelles, peuvent servir à la recherche.

Archiver les médias sociaux est un défi majeur, comme la bibliothèque du Congrès des États-Unis l'a appris à ses dépens. En 2010, elle a accepté de se charger de

la conservation de l'ensemble des archives de Twitter. En octobre 2012, environ cinq cents millions de tweets étaient envoyés quotidiennement. À l'heure où nous écrivons, la bibliothèque possède 133 téraoctets de données, qu'il faut deux jours pour interroger (Alabaster, 2013 ; Allen, 2013). Elle n'a pas encore réussi à rendre les données interrogeables ; les rendre interprétables est plus difficile encore. Les formes, les contenus et les usages des médias sociaux changent bien plus vite que les méthodes scientifiques pour les utiliser. Trouver un moyen d'incorporer ces nouveaux médias dans les infrastructures de la connaissance est une vraie gageure. Ils représentent des ressources précieuses, mais la question des moyens, du comment et du pour qui reste ouverte.

Métadonnées

La diversité des données et des questions de la recherche sur Internet rend difficile la mise en place des schémas de métadonnées communs, propres à permettre l'interopérabilité et l'échange de données. Pour les sondages, le standard de métadonnées le plus consensuel est la Data Documentation Initiative (DDI), qui est exprimée en XML. Comme mentionné au chapitre 2, la DDI a été largement adoptée par les archives de données des sciences sociales et s'applique à tout objet numérique que l'utilisateur ou l'utilisatrice considère comme une donnée (Data Documentation Initiative, 2012). L'un de ses objectifs est de faciliter la réutilisation de métadonnées (Vardigan *et al.*, 2008). La DDI est plus souple et moins prescriptive que les fichiers FITS utilisés en astronomie. Néanmoins, la DDI s'accompagne, comme la plupart des standards, d'une importante documentation. Les utilisateurs doivent investir du temps dans son apprentissage et son implémentation pour structurer et documenter leurs données. Une fois cette opération réalisée, les jeux de données sont prêts à être proposés à des archives ou échangés avec des collaborateurs et collaboratrices qui emploient aussi la DDI.

Parce qu'un sondage peut être très différent d'un autre, la DDI ne normalise pas les noms de variables, les guides de codification et les autres formes de documentation essentielles à l'analyse ou à l'interprétation des données. Dans les sciences sociales, les conventions de nommage et les pratiques documentaires sont souvent élaborées en interne. Même des conventions de base, comme la manière de codifier le sexe et l'âge dans une enquête, varient entre les projets de recherche et les jeux de données, y compris ceux déposés dans des archives. Un sondage peut codifier les hommes et les femmes comme 0 et 1 respectivement, un autre fera l'inverse et d'autres encore utiliseront 1 et 2 ou 2 et 1. De même, l'âge peut être codifié en années ou par l'année de naissance, qui sera notée avec deux ou quatre chiffres. Ainsi, une entrée « 45 » dans le champ « âge » peut signifier « le sujet a 45 ans » dans une étude et « le sujet est né en 1945 » dans une autre. Un jeu de

données n'a guère de valeur s'il lui manque des noms de variables, des guides de codification et d'autres formes de documentation.

Les données issues de médias sociaux sont particulièrement complexes, compte tenu de la variété des sources et la forme libre et dynamique des communications qui y sont effectuées. Les données tirées de services commerciaux tendent à être plus standardisées, surtout s'ils fournissent des API (Application Programming Interface, interface de programmation d'application). Les sources commerciales de données Twitter, par exemple, publient le format dans lequel les flux Twitter sont fournis (Gnip, 2013a). Ces formats sont eux-mêmes fondés sur des standards publiés d'échange de données, comme le JavaScript Object Notation (JavaScript Object Notation, 2013). Comme la DDI, ces formats ne sont qu'un point de départ pour les scientifiques qui sélectionnent des unités d'activité sur les médias sociaux pour s'en servir comme données.

Provenance

Établir la provenance dans les études sur Internet est difficile en raison de l'adaptabilité des méthodes, de la variété des ressources, des origines multiples des données et du manque de documentation sur les décisions prises à chaque étape du traitement. Ce sont généralement les enquêtes qui prévoyaient d'emblée une réutilisation qui présentent les meilleures informations de provenance. Le General Social Survey (GSS), qui fait partie de l'International Social Survey Programme (ISSP), est en cela exemplaire. Les données, les guides de codification et le reste de la documentation de ce sondage, qui est mené depuis 1972, sont à la disposition du public (General Social Survey, 2013 ; International Social Survey Programme, 2013). Pour les personnes formées à la recherche par sondage et à l'utilisation de ces données, repérer la formulation d'une question à une année considérée et la distribution des résultats est simple. Des usages plus complexes nécessiteront un examen plus poussé pour déterminer quelles questions – sur des sujets comme l'orientation politique – ont varié d'année en année, comment les variables ont été codifiées, comment elles ont été combinées dans des index, etc. Reportez-vous à la figure 6.1 tirée du General Social Survey pour en voir un exemple.

Malgré la documentation explicite de la provenance dans le General Social Survey, il faut beaucoup de savoir-faire et de travail pour utiliser ces données de manière fiable, comme le montrent les références croisées dans la légende du tableau. Les questions ont légèrement évolué de sondage en sondage, s'adaptant à leur époque. Les regroupements ont également changé, modifiant l'empreinte de chaque jeu de données, ce qui n'est pas sans rappeler les problèmes cartographiques du relevé COMPLETE, que nous avons vus au chapitre 5. Les choix

préliminaires de réduction et de nettoyage des données – que ce soit dans les sondages, les indicateurs économiques ou les observations astronomiques – peuvent influer profondément sur la capacité à retracer la provenance ou à réinterpréter des jeux de données (Blocker et Meng, 2013).

56. Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?

[VAR: PARTYID]

RESPONSE	PUNCH	YEAR										COL. 240
		1972-82	1982B	1983-87	1987B	1988-91	1993-96	1998	2000	2002	2004	2006	ALL
Strong Democrat	0	2197	143	1271	151	864	1050	370	414	408	455	700	8023
Not very strong Democrat	1	3482	109	1655	89	1282	1542	597	507	515	504	736	11018
Independent, close to Democrat	2	1768	44	904	51	578	887	349	325	267	281	527	5981
Independent (Neither, No response)	3	1736	30	855	32	721	1031	477	566	528	471	997	7444
Independent, close to Republican	4	1106	8	743	9	571	698	244	261	199	239	327	4405
Not very strong Republican	5	2011	8	1259	15	1170	1318	484	399	449	425	637	8175
Strong Republican	6	1009	8	751	2	662	808	239	285	315	396	495	4970
Other party, refused to say	7	243	0	75	1	44	104	63	48	48	29	65	720
Don't know	8	10	0	0	0	0	0	0	0	0	0	0	10
No answer	9	64	4	29	3	15	64	9	12	36	12	26	274

REMARKS: See Appendix D: Recodes, for original question format and method of recoding. See Appendix N for changes across surveys. If planning to perform trend analysis with this variable, please consult GSS Methodological Report No. 56.

Figure 6.1. Tableau du General Social Survey comparant l'affiliation à un parti politique au fil du temps

Crédit : National Opinion Research Center.

La documentation de provenance est essentielle aux données de médias sociaux et, pourtant, très difficile à établir. L'origine d'un tweet, d'un billet de blog ou d'une autre communication peut constituer un indicateur de sa valeur, de sa fiabilité ou de sa validité. La provenance peut inclure du contexte sur les émetteurs, les destinataires, le contenu de la communication et la chaîne de relations associées à chacun de ces éléments. Certaines études se préoccupent des relations tandis que d'autres s'intéressent à la teneur des communications. Les tweets mentionnent fréquemment des ressources en ligne et fournissent des liens pour y accéder. Ces derniers se brisent vite et les ressources disparaissent, rendant la provenance des tweets difficile à retracer (Salaheldeen et Nelson, 2013).

Le souci de la provenance et de l'interopérabilité a conduit à la diffusion d'outils ouverts pour la recherche sur les médias sociaux (Social Media Research Foundation, 2013). Ces outils dépendent parfois eux-mêmes de standards techniques pour documenter les relations de provenance dans le Web sémantique (Groth et Moreau, 2013). Cependant, la plupart des médias sociaux ne sont pas encore conçus d'après les technologies du Web sémantique, aussi les chercheurs et chercheuses utilisant ces données ont-ils toujours besoin d'élaborer leurs propres relations de provenance

(Barbier *et al.*, 2013). Étant donné l'évolution rapide des médias sociaux, des outils, des méthodes et des savoir-faire, il est peu probable que nous pourrons à court terme déterminer la provenance à un niveau suffisant pour pouvoir comparer ou réutiliser des jeux de données.

Les influences extérieures

Parce que les sciences sociales étudient le comportement humain, elles subissent davantage de contraintes extérieures qu'aucun autre domaine. Bien des problèmes de vie privée, de confidentialité et de propriété sont insolubles. D'autres peuvent être réglés ingénieusement dans le cadre de ce qui peut être demandé ou observé, des coûts d'acquisition des données et de l'usage de biens détenus par autrui (Brady, 2004).

Économie et valeur

La valeur des données des sciences sociales en général et de la recherche sur Internet en particulier réside souvent dans leur « emballage ». Les enquêtes sociales peuvent constituer des réservoirs communs de ressources si elles se trouvent dans des archives ouvertes à la plupart des scientifiques, comme la UK Data Archive. D'autres peuvent être des biens de club lorsque les archives où elles sont déposées ne sont accessibles qu'aux membres d'un consortium, comme c'est le cas pour l'ICPSR (Inter-University Consortium for Political and Social Research, 2013 ; UK Data Archive, 2014). Elles peuvent constituer des biens privés si elles ne sont échangées que de manière interpersonnelle ou des biens publics si elles sont publiées ouvertement pour être utilisables par quiconque. Acquérir des données de sondage est souvent coûteux. Par exemple, envoyer des enquêteurs et enquêtrices formés sur le terrain suppose un investissement considérable en argent, en temps et en expertise. Réutiliser des sondages permet d'économiser et d'étendre la gamme des ressources disponibles pour la recherche comparative et longitudinale.

De même, l'économie des données de médias sociaux dépend des sources et de l'emballage. Les sociétés commerciales comme Facebook, Twitter et Google possèdent les ressources les plus importantes. Ces dernières ont une immense valeur marchande et sont vendues à des entreprises pour leurs analyses d'affaires. Pour acquérir des jeux de ces données, les scientifiques doivent souvent payer : ce sont donc des biens de club. Certaines données peuvent être obtenues en copiant des sites web et par d'autres moyens. L'activité sur des *chat rooms* et d'autres sources publiques peut être accessible gratuitement, mais imposer d'autres contraintes. En effet, indépendamment d'éventuels frais, les contenus de médias sociaux sont généralement soumis à des restrictions de licences sur ce qu'on peut en faire et sur le degré auxquels ils peuvent être diffusés à d'autres.

Les sondages, les médias sociaux et d'autres observations de l'activité sur Internet peuvent être analysés à l'aide de logiciels. Beaucoup de ces outils sont commerciaux, quelques-uns sont *open source*. Une fois que les données sont entrées dans l'outil, on peut leur appliquer des routines prédéfinies et des scripts sur mesure. Analyser des données de médias sociaux peut nécessiter beaucoup de programmation *ad hoc*. Ces outils produisent généralement des jeux de données. L'utilisation ultérieure de ces derniers dépend de l'accès à un logiciel analytique, voire à une version particulière de celui-ci, auquel s'ajoute l'accès aux programmes ou aux scripts sur mesure. Quelles que soient les conditions de licence qui s'appliquent à un jeu de données, l'accès aux outils peut constituer une contrainte et déterminer s'il peut être utilisé et comment.

Droits de propriété

Les droits se rapportant aux données et aux outils – logiciels, par exemple – sont particulièrement complexes dans le cadre de la recherche sur Internet. Lorsque les données sont disponibles, elles peuvent avoir été diffusées sans documentation spécifiant les logiciels associés et leurs versions. Les outils sur mesure ne sont pas forcément divulgués. Les logiciels nécessaires à la réutilisation et à l'interprétation de jeux de données peuvent être coûteux, incompatibles avec le matériel informatique et les systèmes d'exploitation utilisés, n'être plus disponibles ou n'être pas intégralement diffusés. Lorsque des scientifiques recueillent des informations sur des personnes, les droits sur les données peuvent appartenir aux chercheurs et chercheuses, aux sujets de recherche ou aux fournisseurs d'informations. La possibilité, pour des sociétés, de vendre des données sur leur clientèle et la capacité de cette dernière à contrôler les données qui la concernent dépendent des pays et des juridictions. Le pouvoir de publier des données et la responsabilité afférente varient d'autant. Ces considérations de propriété influent sur les choix des études sur Internet et sur bien d'autres sortes de recherches en sciences sociales. Les scientifiques préfèrent souvent recueillir leurs propres données afin d'éviter ces problèmes.

Éthique

Les considérations éthiques dans la recherche en sciences sociales, y compris dans la recherche sur Internet, concernent le respect des personnes étudiées et leurs droits, que ce soit individuellement ou en tant que groupe. Les données personnelles sont soumises à de nombreuses réglementations quant aux données pouvant être recueillies et aux conditions de collecte. Aux États-Unis, les principes éthiques fondamentaux sur le traitement des sujets humains sont codifiés dans le rapport Belmont pour ce qui est du respect des personnes, de la bienfaisance et de la justice (US Department of Health and Human Services, 1979). L'application de ces règles varie selon les disciplines, les organismes de

financement, les juridictions et d'autres facteurs. La plupart des pays disposent de principes similaires pour la recherche sur des sujets humains.

Les données sur les sujets humains donnent lieu à un imbroglio de conflits. D'un côté, nous avons une prolifération d'informations personnelles divulguées par les internautes sur les médias sociaux et à travers leurs autres transactions sociales et économiques. Ces informations suscitent des débats sur la vie privée et la propriété des données. De l'autre côté, nous avons des principes établis pour protéger l'identité de sujets de recherche en raison du risque physique, psychologique ou économique que la divulgation leur ferait encourir. Entre les deux se trouvent des interrogations sur les règles autour du consentement éclairé qui seraient trop rigides et empêcheraient le partage de données susceptibles de bénéficier aux individus concernés. Les données biomédicales sont en première ligne de ces conflits. Certaines associations de patientes et patients souhaitent des contrôles plus étroits pour empêcher la réidentification, alors que d'autres veulent octroyer aux malades le droit de consentir au partage de données les concernant (Field *et al.*, 2009).

Les règles éthiques sur les sujets humains sont plus claires en ce qui concerne les méthodes traditionnelles, telles que les sondages. Les chercheurs et chercheuses déterminent quelles questions poser, de quelle manière et comment rapporter les données de façon à maintenir la confidentialité. La diffusion des données tirées de sondages dépend souvent de la capacité à les codifier de manière à préserver l'anonymat des sujets. Les archives de données disposent de procédures et de politiques pour garantir que les données sont fournies et publiées sous des formes qui respectent la confidentialité. Cependant, à mesure que le volume de données en ligne sur les sujets humains augmente, la capacité à identifier des personnes en combinant des jeux de données progresse. Il est parfois nécessaire d'ajouter des niveaux de contrôle, par exemple en ne donnant accès qu'aux métadonnées, et ce uniquement aux chercheurs certifiés. Il y a de plus en plus de données personnelles identifiables recueillies et conservées sous forme numérique et les risques changent en conséquence, de même que les moyens de les écarter (National Research Council, 2013).

Selon la plupart des directives en matière de sujets humains, les données publiques et les comportements adoptés en public, comme le fait de tweeter, de bloguer et de publier, ne sont pas librement utilisables par la recherche. L'Association of Internet Researchers occupe un rôle majeur dans ce domaine en établissant des principes pour la collecte de données issues d'activités en ligne (Association of Internet Researchers, 2012). En utilisant les services des médias sociaux, les personnes ne consentent pas nécessairement à devenir des sujets de recherche. Bien qu'il soit

difficile d'obtenir un consentement éclairé pour de telles études, d'autres moyens existent pour protéger les droits et la vie privée des sujets. Les politiques en matière de vie privée des sociétés de médias sociaux sont soumises à une surveillance étroite, ce qui a réduit d'autant plus l'accès aux données de Twitter, Facebook et autres sources (Bruns et Liang, 2012 ; Schroeder, 2014).

Une fois les résultats publiés, des questions peuvent se poser autour de l'usage éthique des données. Les scientifiques s'inquiètent de la façon dont leurs découvertes seront utilisées, potentiellement à mauvais escient. Ces questions se compliquent lorsque l'exactitude des articles est remise en cause, qu'ils soient rétractés plus tard ou non. L'erreur de tableur dans l'article économique de Reinhart-Rogoff n'a été décelée que lorsqu'un tiers a tenté d'en répliquer les résultats ; le coauteur et la coautrice ont alors dû déployer des efforts considérables pour retrouver l'erreur et l'étape du traitement où elle était apparue. D'autres questions ont été soulevées autour de la sélection des données, de leur poids et de l'utilisation de tableurs plutôt que d'un progiciel d'analyse statistique (Marcus, 2013 ; Wiesenthal, 2013). Un auteur influent a d'ailleurs réprimandé les décideurs et décideuses politiques pour s'être fondés sur un unique indicateur statistique et les encourage à examiner une vaste gamme de données, de méthodes et de théories avant de prendre des décisions importantes (Summers, 2013).

Mener des enquêtes sur Internet et des recherches sur les médias sociaux

Des individus et des équipes restreintes peuvent réaliser des entretiens, une expérience, un sondage en ligne ou une analyse de médias sociaux à petite échelle. En revanche, des enquêtes de l'ampleur d'OxIS supposent des ressources humaines et un capital financier substantiel. L'étude des médias sociaux nécessite un cadre théorique et méthodologique, une expertise computationnelle et statistique, la disponibilité de ressources informatiques et l'accès à des données.

L'enquête Oxford Internet Survey (OxIS) a été conduite pour la première fois en 2003 dans le cadre d'une étude internationale sur l'usage d'Internet (World Internet Project, 2013). Sur la base de cette réussite, OxIS a été menée de nouveau en 2005, 2007, 2009, 2011 et 2013 (Dutton *et al.*, 2013 ; Dutton *et al.*, 2005 ; Dutton *et al.*, 2009 ; Dutton et Blank, 2011). Le responsable de la recherche d'OxIS a assuré la continuité du projet pendant les six éditions de l'enquête, tandis que le personnel changeait après un ou deux cycles. OxIS a ainsi accumulé plus d'une décennie d'observations comparatives sur qui utilise Internet en Grande-Bretagne, comment, avec quels appareils et en quoi l'usage d'Internet est comparable à celui d'autres

médias. L'enquête pose également des questions sur les normes sociales, comme l'intérêt pour la politique ou l'opinion sur le gouvernement, qui permettent des comparaisons avec d'autres études.

En dépit de leur ressource commune, les études qui exploitent les contenus de Twitter diffèrent tellement en matière de théories et de méthodologies qu'il est difficile de les comparer. Les choix de phénomènes à étudier conduisent à sélectionner différentes entités. Par conséquent, un même contenu peut engendrer de nombreuses sortes de données. Meyer, Schroeder et Taylor (Meyer *et al.*, 2013), par exemple, ont comparé la théorie, la méthode et les résultats de trois études de Twitter où des informaticiennes et informaticiens évoquaient des questions sociales. L'une de ces études demandait si Twitter était utilisé plutôt comme un moyen de communication ou comme un réseau de relations, avec 1,47 milliard de relations Twitter (Kwak *et al.*, 2010) ; deux d'entre elles s'interrogeaient sur qui influence qui (Bakshy *et al.*, 2011 ; Cha *et al.*, 2010), s'appuyant respectivement sur 1,6 million d'utilisateurs et utilisatrices et sur 1,7 milliard de tweets de 54 millions d'utilisateurs. Ralph Schroeder (2014), dans un examen ultérieur des aspects sociaux de l'article de Kwak (*et al.*, 2010), remarque que l'envergure des données Twitter dont ils disposaient n'est plus accessible aux scientifiques aujourd'hui.

Les questions de recherche

Dans les enquêtes sur Internet comme dans les médias sociaux, il est difficile de séparer les questions de recherche de la théorie et de la méthodologie. Les études sont conçues pour saisir des informations sur des phénomènes particuliers, dans des populations particulières, à des moments – et parfois dans des endroits – spécifiques, qui seront mesurées d'une manière conforme aux théories et aux hypothèses.

Les sondages OxIS portent sur un large périmètre, à savoir l'usage d'Internet au sein d'un pays sur une longue période, mais sont en partie conçus pour être comparés à des questions semblables posées par des institutions partenaires dans le cadre du World Internet Project. Les chercheurs et chercheuses de l'Oxford Internet Institute (OII) élaborent les enquêtes avec une grande minutie, travaillant soigneusement chaque item du questionnaire, leur enchaînement, le plan de sondage et les consignes aux sondeurs et sondeuses. Afin d'assurer la continuité, une série de questions clés est répétée à chaque étude et parfois affinée pour s'adapter à l'évolution technologique. Les choix d'analyse de données font partie intégrante du protocole de recherche, déterminant à l'avance la codification de chaque item pour atteindre la plus grande puissance statistique. Certains items du sondage fournissent des données démographiques et descriptives sur la population des

internautes ; d'autres sont conçus pour explorer des questions théoriques sur qui utilise Internet, quand, comment et pourquoi.

Les manières dont les questions de recherche façonnent le choix des preuves sont particulièrement notables dans l'étude des médias sociaux. Par exemple, Kwak et ses collaborateurs (2010) s'intéressent à la topologie du réseau. Pour savoir si Twitter représente un réseau social ou un média d'actualité, elles et ils ont comparé la distribution abonnements/abonnés de l'ensemble du site à la mi-2009 à des schémas d'activités dans d'autres réseaux humains. En revanche, une étude très différente est celle de la censure dans les réseaux de microblogage chinois (King *et al.*, 2013). Ces chercheurs et chercheuses, bien connus dans les sciences sociales quantitatives, ont combiné analyse de réseau et entretiens.

La collecte de données

La part des ressources et du travail de recherche consacré à la collecte de données varie aussi considérablement. Dans les enquêtes, le travail de terrain pour mener des entretiens peut prendre des semaines, voire des mois, suivi du même temps d'analyse des données. Chacun des six sondages OxIS cite des réponses issues d'entretiens avec plus de deux mille individus (Dutton *et al.*, 2013). Bien qu'ils soient préparés et analysés par des scientifiques de l'Oxford Internet Institute, les entretiens sont délégués à une société spécialisée dans la recherche de terrain. L'OII reçoit alors des fichiers de données anonymisées codifiés selon ses spécifications.

Dans les études sur les médias sociaux, déterminer quelles données sont souhaitées, lesquelles peuvent être collectées et comment les recueillir peut s'avérer plus laborieux qu'obtenir le jeu de données. En effet, la collecte peut se faire en quelques heures ou jours à l'aide d'un algorithme. Elle est précédée par le développement et le test des algorithmes et peut être suivie de longues périodes d'analyse. Les données Twitter présentent, entre autres attraits, une structure relativement simple, un format standard et un vaste champ d'applications, malgré le peu d'informations transmises dans chaque tweet. Si l'on utilise des API pour collecter les données, les traces sont enregistrées avec l'heure, la date, le lieu et d'autres variables en syntaxe normée, qui peuvent être traités en métadonnées. Notons cependant qu'apprendre à utiliser l'API de Twitter représente un investissement non négligeable, à en juger par l'existence de manuels d'utilisation de plus de quatre cents pages (Makice, 2009).

L'étude de Kwak (*et al*., 2010) a utilisé l'API de Twitter pour collecter les profils de tous les utilisateurs et utilisatrices pendant trois semaines. Au cours des deux mois suivants, elles et ils ont relevé les profils d'utilisateurs évoquant des sujets en vogue. Une série d'informations annexes sur ces thèmes et les tweets associés

ont également été recueillies. De son côté, l'étude de King (*et al*., 2013) utilise une méthodologie très différente pour l'acquisition de ses données, obtenant 11,3 millions de publications de plus de 1 400 plateformes de microblogage et autres médias sociaux en Chine. Elles et ils ont conçu leurs propres algorithmes pour recueillir et codifier les données parce que leur but était d'enregistrer les publications rapidement avant que la censure ne les repère et ne les supprime.

L'analyse des données

Les méthodes quantitatives produisent des données complexes et difficiles à nettoyer, étalonner, interpréter et conserver. Un grand nombre de petites décisions sont effectuées lors de leur traitement et de leur analyse. La capacité à réutiliser des données peut absolument dépendre de la correction des erreurs et des anomalies, de la cohérence des décisions prises, de leur adéquation avec les buts de l'étude et de leur documentation claire et complète.

L'OII reçoit les jeux de données des entretiens sous forme de fichiers dans STATA, un progiciel d'analyse de données et de statistique largement utilisé dans les sciences sociales (STATA, 2013). Selon les consignes transmises par l'OII à l'entreprise chargée de conduire les entretiens en son nom, chaque réponse de chaque question est codifiée avec des noms de variables spécifiques dans le jeu de données. Les analystes de l'OII traitent ces noms comme des métadonnées pour analyser les données. La première passe sert à nettoyer le jeu de données, à repérer d'éventuelles anomalies et erreurs et, si nécessaire, à poser des questions à l'institut de sondage. Dans l'enquête de 2011, le personnel de l'OII a échangé à trois reprises avec l'institut avant de publier son rapport préliminaire. Afin d'analyser les données de manière plus poussée pour des articles de revue ultérieurs, le personnel a encore contacté l'entreprise plusieurs fois avec des questions plus précises.

La plupart des anomalies ne pouvaient être repérées que par des personnes dotées d'une connaissance intime des relations et des tendances attendues dans les données. Des relations surprenantes pouvaient indiquer des découvertes importantes ou bien être le fruit d'une codification incohérente. Ainsi, en voyant des jeunes déclarer des comportements plus typiques des personnes âgées et vice versa, les chercheurs et chercheuses de l'OII se sont interrogés sur la codification du champ de l'âge. Il se trouve que, bien que l'OII ait spécifié aux sondeurs et sondeuses de demander la date de naissance plutôt que l'âge des personnes interrogées, un examen attentif a révélé que certains avaient demandé aux enquêtés leur âge et l'avaient inscrit tel quel. Certaines personnes de 22 ans étaient donc codifiées par un 22, ce qui, dans le fichier STATA, voulait dire qu'elles étaient nées en 1922.

Kwak et ses collaborateurs (2010) ont formulé leur problématique de recherche en termes de topologie de réseau et utilisé les données de médias sociaux comme un banc d'essai. En toute logique, leur collecte et leur analyse de données se sont concentrées sur les liens relationnels. Elles et ils ont recueilli le plus d'informations contextuelles possible sur chaque tweet. Ces scientifiques ont expérimenté différentes techniques pour repérer les spams et plusieurs seuils pour les considérer comme de mauvaises données. Au bout du compte, ils ont supprimé plus de vingt millions de tweets et près de deux millions de comptes utilisateurs de leur jeu de données initial. Les flux d'informations étaient mesurés par le nombre d'abonnés de chaque utilisateur ou utilisatrice et par le degré de réciprocité de ces relations. L'équipe de recherche en a conclu que Twitter se rapprochait davantage d'un mécanisme de diffusion d'informations que d'un outil pour bâtir et entretenir un réseau de relations sociales.

L'étude de King (*et al.*, 2013) est, quant à elle, fondée sur des théories de la censure et de l'action gouvernementale. Les explications sur la théorie, la collecte de données, le codage et l'analyse y sont étroitement mêlées et occupent près de la moitié de l'article. Chaque décision de codage est minutieusement expliquée au regard des théories sociales. Certaines publications ont été codifiées à la main pour interpréter la teneur et l'intention du message. L'article se termine par une explication des implications sociales et politiques des schémas de censure identifiés.

La publication des résultats

Les résultats de ces études sur Internet sont publiés dans divers médias en fonction de leurs différences en matière de théorie, de méthode et de lectorat. Les rapports de chaque enquête OxIS sont diffusés à un très large public et notamment aux décideurs et décideuses politiques lors d'événements. Le rapport d'ensemble, disponible en ligne et sur papier, comprend des données récapitulatives sur chaque variable importante, des tableaux croisés et un résumé des méthodes de collecte. Des détails méthodologiques supplémentaires ainsi que l'instrument d'entrevue sont publiés en ligne. Une fois le rapport divulgué, les analystes d'OxIS explorent plus profondément les données de l'enquête aux fins de la recherche théorique et politique. Les articles de revue sont un endroit plus adéquat pour aborder les détails de la méthodologie et effectuer des comparaisons avec d'autres études (Di Gennaro et Dutton, 2007 ; Dutton et Shepherd, 2006). On peut les trouver dans les revues et les actes de recherche en communication, de science politique, de sociologie et de domaines proches.

Les études sur les médias sociaux sont rapportées dans tous les domaines, des sciences sociales à l'informatique en passant par les sciences exactes, la médecine,

les sciences humaines et la presse grand public. Les deux études évoquées ici (King *et al.*, 2013 ; Kwak *et al.*, 2010) ont été publiées dans des médias reconnus dans leurs disciplines respectives et toutes deux sont largement citées. Cependant, les différences théoriques et méthodologiques sont telles qu'effectuer des comparaisons entre travaux pourrait être vain – ou bien mener à des découvertes sensationnelles. Des études comme celles-là puisent leurs données dans des ressources communes, mais les entités choisies deviennent des preuves très différentes de phénomènes distincts. Ces domaines savants se recoupent de façon nouvelle. Néanmoins, il est rare qu'un politologue ou une informaticienne recherchent des éléments théoriques ou méthodologiques dans la littérature de l'autre.

La conservation, le partage et la réutilisation des données

Le cas des études sur Internet illustre les compromis entre méthodes conventionnelles et novatrices. L'innovation peut engendrer des découvertes importantes, mais entraîner des difficultés dans la documentation, le partage et la réutilisation des données. Les pratiques de conservation, de partage et de réutilisation semblent varier d'un projet à l'autre. La taille et la longueur des études, le renouvellement du personnel, les exigences externes de normalisation ou de dépôt et les attentes liées à la réutilisation sont autant de facteurs qui influent sur les pratiques en matière de données.

Les enquêtes OxIS sont conçues pour consigner l'usage d'Internet en Grande-Bretagne de manière à se comparer à des travaux analogues dans d'autres pays. Bien que beaucoup des questions clés soient les mêmes, le plan de sondage, les questions et l'interprétation des réponses différeront significativement. Alors qu'OxIS couvre des zones urbaines et rurales de la Grande-Bretagne, les études de pays en développement sont largement urbaines ; les pays riches disposent de plus de connexion haut débit alors que les pays pauvres utilisent essentiellement les réseaux mobiles, et ainsi de suite.

Les attentes en matière de réutilisation influent sur la capacité à interpréter les données ultérieurement. La première édition d'OxIS, en 2003, a été lancée rapidement en raison du financement et des collaborateurs et collaboratrices disponibles. Bien que le projet ait dès l'origine été pensé comme une étude continue, le personnel de recherche n'a d'abord constitué qu'une documentation minime sur la collecte, la manipulation, la codification et l'analyse des données. Le questionnaire et le jeu de données de l'enquête de 2003 ont servi de base à l'étude OxIS de 2005. Au fil des enquêtes, le personnel a élaboré des procédures de documentation de plus en plus détaillées et complexes pour permettre les comparaisons longitudinales. Les jeux de données cumulés ont gagné en valeur avec chaque nouvelle série d'observations.

L'OII diffuse les données d'OxIS et documente ses jeux de manière très complète pour leur réutilisation. Il s'agit cependant de deux procédures distinctes. Comme les organismes de financement n'exigent pas de l'OII qu'il dépose ses données dans un référentiel, celui-ci diffuse ses jeux directement auprès de scientifiques qualifiés à des fins non commerciales après une période d'embargo de deux ans. Par exemple, c'est au moment de la publication du rapport de l'édition 2011 que le jeu de données de 2009 a été divulgué. Les conditions d'accès et de licence sont expliquées sur son site web. Le jeu de données mis à disposition correspond à celui reçu de l'entreprise qui conduit les entretiens, sous forme de fichier STATA. La documentation en ligne comprend les noms des variables et les questions dont celles-ci sont dérivées, ainsi que la méthodologie de base et les statistiques descriptives publiées dans le rapport d'enquête. Bien que le personnel d'OxIS divulgue volontiers, sur demande, des informations supplémentaires sur les jeux de données, il a reçu peu de requêtes des deux cents parties environ qui ont fait l'acquisition du jeu. Les réutilisateurs et réutilisatrices peuvent opérer leurs propres transformations et codifications sur le jeu de données STATA « brut ». L'équipe d'OxIS connaît la DDI, mais ne l'a pas mise en œuvre : son utilisation du jeu de données ne justifierait pas les frais induits et aucun facteur externe ne lui impose d'implémenter cette norme.

Les résultats rapportés par l'OII sont consignés selon les standards des différents médias de publication. La documentation rédigée en interne pour maintenir la fiabilité, la validité et la continuité au fil des enquêtes est considérée comme un contenu exclusif à son propre usage, ce qui n'est pas sans rappeler la façon dont on traite beaucoup de logiciels développés localement. Les informations internes comportent des détails sur les choix de transformation des variables, de combinaison des variables en index, de vérification des items et d'autres opérations de nettoyage et d'analyse effectués par les chercheurs et chercheuses. Cette documentation interne favorise la continuité et fournit à l'équipe un avantage concurrentiel.

La spécificité et la temporalité des études sur les médias sociaux constituent leur force, mais aussi leur faiblesse pour ce qui est de la réutilisation et de la réplicabilité. La nature temporelle des données de médias sociaux, à laquelle s'ajoute la dégradation rapide des sources auxquelles elles font référence, limite leur intérêt pour une réutilisation. En théorie, les études sur les médias sociaux pourraient être répliquées à intervalles réguliers. Néanmoins, les méthodes de collecte et d'analyse des données présentent une continuité moindre que la démarche des sondages, car elles reposent sur des outils qui évoluent rapidement. À mesure que ces médias deviennent des formes de communication courantes, leurs caractéristiques se transforment. L'apparition de robots sociaux, de messages qui s'autodétruisent et d'autres innovations rend les comparaisons difficiles. Faute d'indicateurs stables, valider et

étalonner ces méthodes pose problème. L'article d'informatique (Kwak *et al.*, 2010) et celui de science politique (King *et al.*, 2013) se concluent tous deux par des remarques comme quoi leurs résultats pourraient être largement appliqués à d'autres domaines. Ils n'explicitent cependant pas si leurs données sont disponibles, ni où ni comment. Que les jeux de données et les algorithmes puissent être réutilisés ou les conditions répliquées ou non, chacune de ces études suscite des questions de recherche qui peuvent être explorées dans différents domaines.

La sociotechnique

La sociotechnique, l'objet du second cas du présent chapitre, s'intéresse aux problèmes qui sont en partie sociaux et en partie techniques. L'étude de cas développée ici – qui correspond à une recherche en cours depuis 2002 sur les pratiques en matière de données et les infrastructures de la connaissance au CENS – combine des approches idiographique et nomothétique et applique un vaste éventail de méthodes de recherche.

Les méthodes idiographiques applicables à la sociotechnique comprennent les ethnographies, les entretiens, les récits oraux, les observations formelles et informelles et les analyses de traces d'activité humaine. Les technologies peuvent être étudiées du point de vue de l'interaction personne-machine, que ce soit pour explorer la cognition ou des phénomènes associés à des contextes spécifiques. Elles peuvent aussi être observées en tant qu'interventions dans les pratiques professionnelles, comme dans le domaine du travail coopératif assisté par ordinateur. Les projets dans l'un ou l'autre de ces champs peuvent être modestes, locaux et courts ou bien vastes, globaux et à long terme.

Les explications idiographiques sont unies par la place qu'elles accordent à l'examen minutieux et à l'interprétation. Les démarches interprétativistes font le lien entre les sciences sociales et les sciences humaines, qui explorent et comparent de multiples points de vue. Bien que ces méthodes aient une longue histoire, elles font toujours l'objet de polémiques sur des questions d'épistémologie, de critères de preuves et de philosophies du savoir. Nous prenons ici acte de ces débats, sans tenter de les traiter ou de les trancher (Garfinkel, 1967 ; Geertz, 1973 ; Glaser et Strauss, 1967 ; Latour et Woolgar, 1986 ; Latour, 1987 ; Lévi-Strauss, 1966 ; Lofland *et al.*, 2006 ; Roth et Mehta, 2002).

La taille compte

La plupart des travaux sociotechniques correspondent à de la science légère, puisqu'ils s'intéressent à des problématiques émergentes et confuses. Ils sont menés par de petites équipes de recherche ou même par un chercheur ou une chercheuse isolés, selon l'ampleur du projet. Il faut parfois des mois, voire des années de labeur pour transcrire, compiler, codifier, analyser et interpréter des données qualitatives. Les réseaux sociaux des communautés de recherche peuvent, grâce à l'analyse de publications ou d'autres indicateurs sociométriques, fournir des comparaisons quantitatives. De même, la fouille de texte (*text mining*) dans les publications peut révéler des récurrences dans les objets de recherche et leur évolution. Le but est généralement de combiner plusieurs méthodes. En appliquant des démarches présentant divers degrés de fiabilité, de validité interne et externe ainsi que différentes échelles, on peut entériner les résultats par triangulation.

Quand est-ce une donnée ?

Les travaux du CENS rapportés dans la présente étude de cas se fondent sur une littérature sociotechnique sur les pratiques de terrain, notamment dans les sciences de l'environnement, de plus en plus foisonnantes. Il s'agit essentiellement d'études qualitatives ayant recours à une combinaison d'ethnographie, d'entretiens, d'observation participante et d'analyse de documents (Aronova *et al.*, 2010 ; Cragin *et al.*, 2010 ; Cragin et Shankar, 2006 ; Jackson et Buyuktur, 2014 ; Jackson *et al.*, 2011 ; Karasti *et al.*, 2006 ; Olson *et al.*, 2008 ; Zimmerman, 2007). Comme pour la recherche sur Internet, les scientifiques peuvent puiser dans une pléthore de ressources en données. Les spécialistes de la sociotechnique doivent se montrer ingénieuses et ingénieux pour identifier les sources de données et évaluer leur crédibilité, leur indépendance et leurs relations avec d'autres formes de données.

Sources et ressources

En dépit du déluge de données dont elles disposent sur la pratique scientifique, les études sociotechniques sur les pratiques en matière de données tendent à se faire *in situ*. C'est pourquoi les chercheurs et chercheuses sont davantage susceptibles de recueillir leurs propres observations que de se fonder sur des sources externes. Néanmoins, les traces produites par les sujets de recherche, dont leurs publications, peuvent constituer des sources auxiliaires. Les catégories de sources et de ressources que nous évoquons ci-après sont courantes dans la recherche sur les pratiques de terrain citées plus haut, mais ne représentent pas une liste exhaustive de toutes les sources de preuves potentielles.

Observations de terrain et ethnographie

La recherche sociotechnique prend place sur un site de recherche, qui n'est pas nécessairement un lieu physique unique. On appelle ethnographies multisites des travaux qui effectuent des comparaisons entre différents sites (Marcus, 1995). Les observations peuvent également s'effectuer en ligne, grâce aux méthodes de « l'ethnographie virtuelle » (Hine, 2000). Les scientifiques peuvent observer les activités qui les intéressent de loin et mener des entretiens à l'aide d'une communication vidéo ou audio. Si elles augmentent l'échelle et la distance du projet de recherche, ces démarches renoncent néanmoins à un certain nombre d'informations contextuelles.

Toute recherche qui étudie des individus ou des communautés *in situ* nécessite leur permission. C'est ce qu'on appelle « l'entrée ». Une fois arrivé sur le site de recherche, y rester peut s'avérer tout aussi difficile. Dans les sciences sociales, le but de l'observation sur le terrain est d'étudier des phénomènes sans perturber l'environnement plus qu'absolument nécessaire. Une fois sur place, les chercheurs et chercheuses se présentent et expliquent la raison de leur présence plus en détail, puis s'efforcent d'être discrets et de perturber le moins possible le milieu. Ils cherchent ainsi à éviter « l'effet Hawthorne », qui tire son nom d'une célèbre étude sur une centrale de Western Electric où la productivité a augmenté à la suite d'expérimentations sur la luminosité, avant de décliner après la fin de l'étude. Le simple fait d'être observé modifie le comportement des personnes (Landsberger, 1958).

Une manière de minimiser les risques pour la validité des observations de terrain est de rester suffisamment longtemps pour que la présence du chercheur ou de la chercheuse fasse partie de l'environnement quotidien. Ce processus peut prendre des semaines, des mois ou des années. Les anthropologues peuvent ainsi passer une carrière entière à étudier une communauté particulière, qu'il s'agisse de physiciens ou de gangsters. Une manière de se fondre dans le décor consiste à participer activement et ainsi à en apprendre plus sur le site. Par exemple, s'ils étudient les pratiques scientifiques en matière de données, les chercheurs peuvent contribuer à la collecte ou à l'analyse des données de l'équipe, se charger des commissions pour les fournitures et le matériel ou aider à des tâches physiques comme installer des batteries sur le terrain. La participation aux activités étudiées comporte cependant ses propres risques. Les chercheurs peuvent craindre de se montrer trop subjectifs, de « se convertir » et de parler au nom de leurs sujets de recherche plutôt que de parler d'eux.

Les chercheurs et chercheuses appliquent de nombreuses techniques pour rassembler des observations et les transformer en données. Sur le terrain, il vaut souvent mieux participer d'abord et prendre des notes plus tard. Selon les circonstances, les

chercheurs prennent des notes, enregistrent de l'audio ou des vidéos ou prennent des photos, avec la permission des personnes observées. Les enregistrements sonores peuvent être retranscrits, mais cette tâche est longue et coûteuse. Les enregistrements eux-mêmes restent précieux en raison des nuances de la voix et des indices visuels qu'on y trouve. Les notes, les transcriptions, les entretiens et les autres documents seront codifiés selon les thématiques, les événements et d'autres indicateurs des phénomènes étudiés. La codification peut être simple et manuelle ou complexe et assistée par des technologies. Ainsi, des outils d'analyse commerciaux et *open source* existent pour encoder des événements, des personnes, des thématiques et d'autres catégories. Une fois les documents codifiés, ces outils simplifient l'agrégation et la visualisation.

Entretiens

Dans la recherche sociotechnique idiographique, les entretiens consistent généralement en des questions ouvertes. Une entrevue de cinq à dix questions peut susciter une conversation d'une heure ou plus. Bien que les questions soient générales, elles peuvent produire des explications des activités : par exemple, comment les participantes et participants sélectionnent les terrains ou les technologies, ou quelles difficultés ils rencontrent dans la conservation de leurs données. En posant les mêmes questions à plusieurs personnes, les scientifiques peuvent comparer les pratiques d'individus, d'équipes, de domaines, de terrains et autres.

Dans le cadre d'une recherche qualitative, un échantillonnage aléatoire est rarement réalisable. Si le site est petit, il peut être possible d'interviewer chaque participante et participant. Le plus souvent, la population du site est trop diverse pour qu'un même ensemble de questions soit pertinent pour tous les enquêtés. Ainsi, un sujet de préoccupation des cheffes et chefs d'équipe peut ne pas concerner celles et ceux qui fabriquent les instruments, et vice versa. Une démarche alternative consiste à stratifier l'échantillon par catégories théoriques du phénomène étudié, en équilibrant le nombre et la distribution des participants dans chacune.

Traces et documents

Les traces et les autres formes de documents associées à un site de recherche peuvent être des ressources en données très utiles, mais elles ne constituent pas des preuves en elles-mêmes, comme le savent bien les spécialistes de l'histoire et de l'archivistique. La compréhension des traces d'activité humaine dépend de la date, du lieu et des circonstances où elles sont utilisées (Furner, 2004a). Les personnels de recherche doivent se renseigner le plus possible sur le contexte de chaque trace et ainsi tisser une trame à partir de ces informations rassemblées. Par exemple, des différences entre les recueils de données de deux laborantines peuvent tenir aussi

bien de leurs habitudes de conservation que des données qu'elles collectent. De même, des différences dans les éléments biographiques sur des personnes reflètent la façon dont elles choisissent de se présenter et la fréquence à laquelle elles le font.

Les spécialistes de la sociotechnique ont recours à des sources publiques pour en apprendre autant que possible sur les sites potentiels. Il est souvent facile de trouver des informations sur le financement, le personnel, les équipements, les événements, les activités de recherche et les publications. S'il faut généralement vérifier les détails des documents publics avec les enquêtés, recueillir ces informations à l'avance permet aux chercheurs et chercheuses de mieux employer leur temps sur place. Lorsqu'ils observent les participants et participantes ou s'entretiennent avec eux, les chercheurs rassemblent souvent autant de documents internes pertinents que les enquêtés veulent bien leur en communiquer. Ainsi, des traces telles que les carnets de laboratoire peuvent fournir un éclairage sur la conduite de la recherche. Certaines représentent des contenus sensibles, que les chercheurs protégeront dans le cadre d'accords de confidentialité.

Concevoir et évaluer des technologies

Un élément de l'étude d'un site de recherche consiste à observer comment les enquêtés utilisent les technologies dans leur travail. Des travaux sur les bibliothèques numériques, les archives de données, les technologies éducatives, les outils collaboratifs, les modèles climatiques, les systèmes de traitement de texte, le courriel et bien d'autres technologies nous ont renseignés sur la manière dont nous travaillons (Blomberg et Karasti, 2013 ; Bowker et Star, 1999 ; Edwards *et al.*, 2007 ; Edwards, 2010 ; Karasti *et al.*, 2006 ; Olson *et al.*, 2008 ; Ribes et Finholt, 2009).

Certains spécialistes sociotechniques bâtissent des systèmes dans le cadre d'interventions dans les communautés de pratique. Dans les bibliothèques numériques, on peut concevoir de petits systèmes pour tester de nouvelles interfaces utilisateur, des fonctionnalités de recherche d'information, des théories pédagogiques ou des processus cognitifs. L'Alexandria Digital Earth Prototype Project, par exemple, était une initiative conjointe de géographes, d'informaticiennes et informaticiens, de psychologues et de spécialistes des sciences de l'information pour utiliser des données des sciences de la terre dans des cours de premier cycle (Borgman *et al.*, 2000 ; Janee et Frew, 2002 ; Smith et Zheng, 2002). Le Science Library Catalog consistait en une interface utilisateur graphique qui servait à étudier les compétences cognitives et développementales d'enfants de huit à douze ans lorsqu'elles et ils cherchaient des informations scientifiques. Le système a été bâti progressivement et amélioré en fonction des capacités des enfants. Ainsi, au cours de plusieurs années de recherche, des hypothèses sur

les différences d'âge dans la capacité à manier les hiérarchies, à catégoriser des informations, à chercher par ordre alphabétique ou sur un graphique et à persévérer dans l'accomplissement de tâches informationnelles ont pu être éprouvées (Borgman *et al.*, 1995 ; Hirsh, 1996).

Les infrastructures de la connaissance

La sociotechnique est une problématique plus qu'une discipline. Les chercheurs et chercheuses convergent vers un problème présentant des aspects sociaux et techniques, comme les pratiques en matière de données, chacun apportant ses propres théories, ses méthodes et ses perspectives. Ils amènent également avec eux, si elles existent, les infrastructures de la connaissance de leurs disciplines respectives. Comme pour la recherche sur Internet, l'expertise partagée, les méthodes et les outils sont plus essentiels à l'infrastructure que les archives. Les données résultant de ces méthodes mixtes sont particulièrement difficiles à organiser, partager ou conserver. Il n'existe que peu de référentiels ou de standards communs qui puissent être appliqués aux notes de terrain, aux sites web, aux enregistrements audio et vidéo, aux fichiers, aux logiciels, aux photographies, aux échantillons physiques et aux myriades d'autres types de données pouvant être recueillies dans ce genre de travaux. Diffuser les données peut revenir à les disperser par genre et par matière, ce qui est contre-productif pour la réplication.

Métadonnées

Étant donné la combinaison des sources de données, on tend à élaborer les conventions de nommage et les pratiques documentaires en interne. Certains personnels de recherche rédigent des guides de codification pour favoriser un encodage cohérent à long terme et par des codeurs et codeuses multiples. En principe, la DDI peut servir à structurer des données et métadonnées qualitatives (Data Documentation Initiative, 2012 ; Vardigan *et al.*, 2008). Dans la pratique, seuls les grands projets comptant de multiples collaborateurs et collaboratrices ont suffisamment intérêt à investir dans des pratiques formelles en matière de métadonnées.

Il est difficile de normaliser les pratiques d'encodage de données qualitatives parce que les sources de données varient beaucoup d'un projet à l'autre et d'un site à l'autre. La nature itérative des observations sur le terrain et de la codification décourage aussi la normalisation : ainsi, la théorie ancrée encourage les scientifiques à enrichir et réviser leur structure de codification à mesure qu'elles et ils collectent des données et découvrent davantage d'informations contextuelles sur celles qu'ils possèdent déjà (Anderson, 1994 ; Glaser et Strauss, 1967 ; Star, 1999). Les hypothèses sont élaborées à travers la codification, puis testées sur d'autres parties du corpus.

L'itération améliore la cohérence et donc la fiabilité des méthodes. Néanmoins, la validité externe peut diminuer à mesure que la validité interne augmente.

Provenance

Consigner la provenance pour des méthodes mixtes nécessite des registres pour chaque type de données et pour les relations qui les unissent. Imaginons, par exemple, qu'une photographie soit prise à un moment et dans un lieu donné dans le cadre d'un entretien donné. Un ensemble de documents est également acquis à cette occasion. La photographie présente un intérêt si elle est comparée à d'autres prises par le même laboratoire à d'autres moments ou à des photographies d'autres laboratoires prises à peu près au même moment. Chaque série d'entretiens, de photographies, d'enregistrements, de documents et d'autres formes de données peut avoir de multiples lignages. La provenance, dans le cadre des méthodes mixtes et des autres formes de recherche interprétative, peut comprendre de multiples réseaux de relations.

La réplication, en soi, représente rarement un problème, car l'explication idiographique est ancrée dans des personnes, des lieux, des moments et des situations. En revanche, la sociotechnique s'inquiète tout autant de la véracité que les autres champs. Les scientifiques acquièrent, quand c'est possible, des informations de sources plurielles et indépendantes. L'information de provenance peut répondre aux besoins du chercheur ou de la chercheuse plutôt qu'à ceux de la réutilisation par d'autres. Il peut conserver ses données sa vie durant, s'assurant ainsi que ses observations puissent être vérifiées. Dans d'autres cas, les comités de protection des personnes exigent que les données soient détruites à la fin du projet. En effet, la confidentialité des traces et l'anonymat des individus dans la présentation des données sont primordiaux. Plus la provenance d'une trace quelconque sur un enquêté est détaillée, plus il est difficile de diffuser les jeux de données.

Les influences extérieures

La recherche sociotechnique implique des sujets humains et est donc contrainte par une série de facteurs économiques, éthiques et de propriété comparables à ceux des autres domaines des sciences sociales. La démarche idiographique étant généralement fondée sur une relation de travail étroite entre les scientifiques et les sujets, ces types de données peuvent être plus sensibles que celles issues des enquêtes sur Internet ou des études des médias sociaux, qui saisissent des traces publiques d'activité.

Économie et valeur

L'« emballage » des données influe sur leur valeur. Les données issues de travaux à méthodes mixtes peuvent être combinées de tant de façons qu'il est difficile de les situer dans les quadrants des biens économiques. Les spécialistes de la sociotechnique peuvent puiser dans des réservoirs communs de ressources associés à leurs sites de recherche, comme des archives de données ou des dépôts de publications, dont les enquêtés sont les auteurs et les autrices. Les informations concernant les sujets de recherche peuvent être publiques ou privées, selon la manière dont on les a obtenues. Cependant, une fois possédés, les produits de données sociotechniques deviennent parfois des biens privés qui ne peuvent pas être partagés sous peine de révéler l'identité des sujets. Par exemple, une liste des pages web des sujets de recherche peut être constituée de liens vers des informations publiques, mais la liste elle-même représente le groupe des participantes et participants et doit être tenue confidentielle.

Droits de propriété

Les personnels de recherche en sociotechnique accèdent à des informations qu'ils peuvent fouiller pour découvrir des éléments probants ; cependant, y accéder ne signifie pas qu'ils peuvent les diffuser ou les reproduire. Par exemple, les carnets de laboratoire demeurent généralement la propriété de l'équipe de recherche et les publications, celle des ayants droit. Toute source de données des études sociotechniques peut être grevée de droits de propriété, que ce soit du matériel, des logiciels, des documents, des échantillons ou d'autres matériaux.

Éthique

Les travaux sociotechniques sont généralement assujettis aux politiques en matière de sujets humains. Néanmoins, les règles de consentement, d'anonymat, de confidentialité et de contrôle des registres peuvent varier légèrement selon les méthodes appliquées. Les risques et les responsabilités ne sont pas toujours les mêmes pour les ethnographies, les observations sur le terrain, les entretiens, les évaluations de technologie, etc. Des problèmes éthiques se posent dans la manière de recueillir les données, de protéger les registres et de rapporter les résultats. Par exemple, lorsqu'on publie des travaux, le niveau de détail fourni dépend parfois des risques encourus par les sujets. Les pratiques d'identification varient selon les revues et les communautés. Dans la plupart des cas, les sites de recherche sont dissimulés en supprimant des détails locaux et en attribuant des pseudonymes aux enquêtés ; mais dans d'autres, les sites et les personnes sont nommés. Les résultats sont agrégés dans des groupes suffisamment grands pour qu'il soit difficile d'identifier les individus.

Les observations peuvent comprendre de longs récits à travers les mots des participantes et participants, des éléments biographiques et des vidéos de visages. Tout travail de recherche qui repose sur un grand niveau de détail concernant des individus et des groupes suppose un investissement considérable dans la sauvegarde de la confidentialité et des droits des personnes. Parmi les propositions de nouvelles règles nationales pour la protection des sujets humains aux États-Unis, on trouve un cadre en quatre parties sur la confidentialité : « données sûres, endroits sûrs, personnes sûres et livrables sûrs » (National Research Council, 2013, p. 50). Dans le modèle proposé, les études sont conçues pour une meilleure sécurité des données avant, pendant et après la collecte. Une méthode consiste à recueillir des données sur plusieurs sites, pour que le lieu soit plus difficile à déceler. Les données confidentielles peuvent être partagées dans des « endroits sûrs », comme des « enclaves de données virtuelles » : elles sont conservées dans un *data center* où les scientifiques peuvent, sur demande, manipuler les données, mais pas les transférer sur leurs propres ordinateurs. Les personnels de recherche qui diffusent des données et ceux qui les réutilisent peuvent devenir des « personnes sûres » grâce à des formations et des certifications. Les livrables peuvent être rendus plus sûrs en catégorisant les données selon leur nocivité. Ainsi, pour les données qui risquent peu de nuire, de simples accords sur les conditions d'utilisation peuvent suffire. Pour les données « véritablement radioactives », des règles plus strictes prévaudraient (National Research Council, 2013, p. 52). Quand bien même ces règles s'appliqueraient à toutes les catégories de données sur les sujets humains, c'est la recherche qualitative qui connaîtrait les changements les plus marqués. Certains des amendements proposés visent à favoriser le partage et la réutilisation des données sur les sujets humains.

Mener des recherches sociotechniques au CENS

Le Center for Embedded Networked Sensing offrait une occasion rare d'étudier, sur le long terme, la formation de collaborations et la collecte, l'analyse, la gestion et la publication des données. Comme nous l'avons expliqué au chapitre 5, le CENS était un centre scientifique et technologique relevant de la National Science Foundation, fondé en 2002 avec un financement pour cinq ans. Une deuxième subvention pour cinq années supplémentaires a prolongé l'existence du CENS jusqu'en 2012. La responsable de la recherche sur les pratiques en matière de données était également l'une des cofondatrices du centre. L'équipe sociotechnique chargée des pratiques en matière de données, qui rejoindrait plus tard l'équipe des statistiques, était une unité de recherche du CENS. Elle comptait également, comme beaucoup des autres unités, des collaborateurs et collaboratrices en dehors du centre.

Une grande partie de la recherche sur les pratiques concernant les données au sein du CENS consistait en de l'observation participante, puisque l'équipe était intégrée au centre. Les professeures et professeurs, les étudiantes et étudiants et le personnel de l'équipe prenaient part à des activités formelles et informelles et passaient le plus de temps possible sur le site. Au cours de l'existence du CENS, l'équipe a réalisé des ethnographies et des entretiens ouverts, observé des essais pilotes en laboratoire et sur le terrain, participé à des déploiements sur le terrain de quelques heures à plusieurs semaines, assisté à d'innombrables réunions d'équipe, présenté ses résultats lors de colloques, de journées d'actualités et de retraites scientifiques et analysé des documents tels que des carnets de laboratoire et de terrain, des fichiers de données de formats divers et des publications. La collecte de données se poursuit aujourd'hui encore, car l'équipe étudie la postérité du CENS dans les années suivant sa fermeture officielle (Borgman *et al.*, 2014).

Les observatrices et observateurs participants sont rarement des membres officiels de l'organisation qu'ils étudient, mais dans le cas du CENS, l'équipe de recherche en sciences sociales avait des responsabilités administratives analogues à celles des autres équipes. Elle a également contribué à l'activité du centre en concevant des technologies pour la collecte et la gestion de données scientifiques, dont un site de dépôt en accès ouvert pour les publications.

Comme pour toute méthodologie de recherche, il fallait en permanence trouver des compromis en matière d'échelle, d'envergure et d'objectivité. En tant que membre à part entière du CENS, l'équipe disposait d'un bien meilleur accès aux individus, aux sites et à la documentation que la plupart des spécialistes des sciences sociales. Elle était aussi consciente du risque de subjectivité. Être conscient de ces risques contribuait déjà à les atténuer, de même qu'avoir des partenaires extérieurs qui posaient des questions critiques sur les méthodes et l'interprétation. Généralement, on demandait aux spécialistes des sciences exactes et des technologies du CENS de vérifier les descriptifs de leurs activités avant qu'ils ne soient publiés par l'équipe sociotechnique. Elles et ils les ont corrigés et améliorés de bonne grâce et sans essayer d'altérer les interprétations effectuées.

Les questions de recherche

L'équipe chargée des pratiques en matière de données a débuté son étude du CENS en se demandant comment les chercheurs et chercheuses individuels et leurs équipes créaient des données et comment ces pratiques varient selon les équipes et les domaines. Étant d'entrée de jeu des membres de la communauté, les spécialistes sociotechniques avaient la possibilité de suivre les données de leur conception initiale au traitement final en passant par les étapes de nettoyage, d'analyse et

de présentation. À mesure que le contexte du travail scientifique des données était mieux appréhendé, les questions de recherche ont évolué pour s'intéresser au travail collaboratif des équipes, au contrôle et à la propriété des données, aux métadonnées et à la provenance, au partage et à la réutilisation et aux applications éducatives des données du CENS. Plusieurs thèses de doctorat, un mémoire de master et de nombreux projets étudiants sur les pratiques en matière de données et les collaborations au sein du centre ont encore élargi les questions de recherche traitées depuis 2002.

Les questions de recherche et le financement ont été réaffirmés tout au long des projets de l'équipe. Les fonds de départ fournis par le CENS ont financé le traitement de la problématique initiale s'agissant des données et des pratiques. Ces résultats ont mené à des demandes de subvention, qui ont financé de nouvelles découvertes qui ont à leur tour conduit à d'autres demandes et à des publications. La plupart des bourses étaient attribuées à plusieurs collaborateurs et collaboratrices, permettant des comparaisons des pratiques en matière de données avec d'autres sites de recherche (Borgman *et al.*, 2006, 2007a, 2007b, 2012 ; Borgman, 2006 ; Edwards *et al.*, 2011 ; Mandell, 2012 ; Mayernik *et al.*, 2007, 2012 ; Mayernik, 2011 ; Pepe *et al.*, 2007, 2010 ; Pepe, 2010 ; Shankar, 2003 ; Shilton, 2011 ; Wallis *et al.*, 2007, 2008, 2010a, 2010b, 2012, 2013 ; Wallis, 2012).

La collecte de données

Le terrain des spécialistes sociotechniques englobait tout endroit où se trouvaient les chercheurs et chercheuses du CENS. Les équipes étaient basées dans l'une des cinq universités participantes et avaient parfois des partenaires dans d'autres institutions. Les données étaient recueillies dans les laboratoires et les espaces publics du campus, ainsi que dans des sites de recherche aux États-Unis et ailleurs dans le monde. Différentes méthodes étaient appliquées simultanément par une équipe allant de deux à huit membres selon les périodes. Certains chercheurs et chercheuses se concentraient sur l'observation des terrains, d'autres sur la conduite d'entretiens, l'analyse de documents, la conception et l'évaluation de technologies et l'identification des réseaux sociaux au sein du CENS. Chaque personne de l'équipe a pris part à au moins deux méthodes de collecte pour bénéficier d'une formation polyvalente et faciliter la combinaison de données à partir de méthodologies multiples.

L'une des premières activités de l'équipe sociotechnique a consisté à identifier des standards et des formats de métadonnées adaptés aux activités de recherche de cette communauté interdisciplinaire naissante. La tâche s'est avérée bien plus difficile que prévu et a orienté les premières années du projet. Les membres du CENS n'ayant guère d'expérience des standards de métadonnées, leur demander directement leurs préférences n'a pas abouti à des renseignements utiles. L'étape

suivante dans ce but a donc consisté à identifier la gamme des données recueillies. Les observations sur le terrain, les entretiens et l'analyse des publications du CENS ont été les principales sources d'information sur ce sujet. L'équipe a repéré quelles données les scientifiques et technologues acquéraient et lesquelles elles et ils conservaient le plus souvent en vue d'une analyse et d'une réutilisation ultérieure.

Une fois les besoins généraux en métadonnées identifiés, on les a comparés aux standards existants ou émergents. Plusieurs standards pour les données environnementales et pour les données de capteurs constituaient des candidats prometteurs. En effet, ils pouvaient, en principe, être utilisés seuls ou combinés pour décrire la majeure partie des données du CENS. Les observations écologiques recueillies par les capteurs ou à la main pourraient être décrites avec une structure et une terminologie communes. Par ailleurs, les caractéristiques des capteurs seraient enregistrées automatiquement si ces standards XML étaient intégrés dans les algorithmes de collecte.

Les standards de métadonnées ont alors été présentés aux chercheurs et chercheuses du CENS en expliquant leurs avantages et leurs inconvénients. Bien que reconnaissants des efforts de l'équipe sociotechnique pour les aider à gérer leurs données, les membres du CENS ont estimé que les standards existants ne répondaient pas à leurs besoins. Comprendre les raisons de cette non-adoption a nécessité encore plusieurs années de recherche approfondie, comme nous le verrons plus en détail au chapitre 8 (Borgman *et al.*, 2006, 2007a ; Borgman, 2006 ; Pepe *et al.*, 2007 ; Shankar, 2003 ; Wallis *et al.*, 2006).

À mesure que le CENS voyait le nombre de ses participantes et participants passer de quelques dizaines à plusieurs centaines, le besoin de meilleurs registres de provenance s'est fait sentir. En effet, la culture orale ne fonctionnait plus. Les équipes de recherche partaient pour un site de déploiement, souvent à plusieurs heures de route, pour se rendre compte une fois sur place qu'il leur manquait une pièce matérielle essentielle ou un savoir-faire particulier détenu par un ou deux participants – typiquement, un étudiant ou une étudiante de second cycle. L'équipe sociotechnique s'est efforcée de combler cette lacune grâce à un logiciel simple, baptisé « CENS Deployment Center » (Mayernik *et al.*, 2007). CENS DC, comme on l'appelait, était rempli de descriptifs du matériel et du personnel présents lors des déploiements précédents – données que l'équipe sociotechnique avait recueillies en amont. Le système comportait des fonctions de modèles pouvant servir à générer un projet de déploiement et à créer des registres sur le terrain, par exemple pour indiquer ce qui avait fonctionné ou non, ce qui manquait ou s'avérait particulièrement utile et ainsi de suite. Ces fonctionnalités visaient à rendre les

déploiements plus efficaces et productifs grâce à l'observation des précédents. Le système enregistrait également des catégories d'information fréquentes dans les publications scientifiques : il pouvait ainsi aider à la rédaction d'articles sur la recherche de terrain. Fondé sur le travail de certaines équipes, CENS DC a été testé par elles avant d'être mis en service. Il a partiellement rempli ses fonctions chez les quelques équipes qui l'ont intégré dans leurs activités (Mayernik, 2011 ; Wallis, 2012).

L'analyse des données

Comme les enquêtes OxIS, la recherche sur les pratiques en matière de données du CENS a démarré de manière modeste. À mesure que l'équipe grandissait et qu'il devenait possible de réaliser des études plus longues et importantes, l'analyse de données s'est faite plus formelle. Les entretiens étaient enregistrés et transcrits de façon professionnelle. On a rédigé des guides de codification pour maintenir une cohérence entre les multiples enquêtes et les nombreuses observations sur le terrain. NVivo, un progiciel commercial d'analyse de données pour la recherche qualitative et à méthodes mixtes, a été utilisé pour encoder les personnes, les événements, les thématiques et d'autres catégories mentionnées dans les notes et les transcriptions (NVivo 10, 2013). À chaque collecte de données, deux étudiantes et étudiants de second cycle encodaient les mêmes entretiens chacun de leur côté. Ils les comparaient ensuite et discutaient leur codification, s'efforçant de concilier leurs différences d'interprétation. On répétait la procédure jusqu'à obtenir une concordance entre codeurs. Les guides de codification étaient annotés d'explications sur les interprétations, ce qui contribuait à leur cohérence. Le défi consistait à élaborer et éprouver des hypothèses à partir de ces données tout en restant ouvert à des résultats imprévus.

Les transcriptions d'entretiens se prêtaient davantage à la concordance entre codeurs que d'autres types de données. En effet, bien que les entretiens soient menés par plusieurs membres de l'équipe, tous travaillaient à partir d'un même ensemble de questions à poser aux chercheurs et chercheuses du CENS. En revanche, sur le terrain, chaque membre prenait copieusement des notes sur ses observations. Ces notes devaient être codifiées par l'observateur ou l'observatrice. S'ils s'en chargeaient rapidement et régulièrement, les observateurs pouvaient combler les lacunes de leurs carnets ou déterminer si un élément manquant pouvait être acquis par d'autres moyens. On encourageait les observateurs à prendre beaucoup de notes, ainsi que des photos quand c'était possible. La prise de notes s'est améliorée avec l'expérience. Néanmoins, il y avait d'importantes variations dans l'attention au détail et dans la connaissance des phénomènes observés chez les membres de l'équipe. L'étudiante ayant une formation en biologie a noté de nombreux détails

sur les espèces étudiées, tandis que le diplômé en génie s'est davantage concentré sur l'instrumentation mise en place. Disposer sur le terrain d'observateurs multiples ayant des parcours complémentaires a élargi le spectre de l'information acquise sur les pratiques en matière de données du CENS.

Les données encodées avec NVivo pouvaient être agrégées par thématiques, événements et diverses autres catégories. On pouvait extraire des sous-ensembles de données et ainsi effectuer des comparaisons entre études. Plus il y avait de données recueillies, plus les comparaisons et les interprétations seraient riches. Les fichiers NVivo, les mémos, les photographies et d'autres traces sont devenus matière à discussion pour l'équipe sociotechnique et ont ainsi contribué à orienter les publications, les présentations et les posters scientifiques.

La publication des résultats

La première étape de partage des résultats de l'équipe sociotechnique se faisait souvent au sein même du CENS. Celui-ci mettait sur pied des présentations par posters et des démonstrations dans le cadre des événements publics qu'il organisait régulièrement, comme les visites annuelles de la National Science Foundation, les journées annuelles d'actualités ouvertes au public et les retraites scientifiques. L'équipe sociotechnique a réalisé des posters à partir de son propre travail et a contribué à ceux d'autres équipes du CENS. Dans ces manifestations, les visiteurs et visiteuses se rencontraient autour de trente à quatre-vingts posters, offrant autant d'occasions de discuter et de bénéficier de critiques. Le CENS organisait également des séminaires hebdomadaires à l'heure du déjeuner pour rapporter les travaux du moment. L'équipe sociotechnique assistait régulièrement à ces événements et y présentait de temps à autre ses résultats.

Les résultats de la recherche sur les pratiques en matière de données du CENS étaient publiés dans différents médias pour atteindre plusieurs publics. L'essentiel des fonds du centre provenant de l'informatique, de nombreuses communications ont été proposées aux conférences de l'ACM (Association for Computing Machinery). Il s'agit de lieux de publication prestigieux et très sélectifs. Intervenir lors de conférences ACM était donc une occasion importante de présenter des découvertes sociotechniques à un auditoire technique. D'autres publications étaient destinées aux chercheurs et chercheuses en sciences de l'information et en sociologie des sciences et des technologies. Quelques publications ciblaient le corps enseignant, notamment sur le thème de la réutilisation de données à des fins pédagogiques. Outre ces publications, cette littérature a donné lieu à de nombreuses présentations qui ont atteint des publics plus variés encore dans les sciences exactes, les sciences sociales, la technologie et les sciences humaines.

Ces publications dévoilent l'identité du CENS, mais les participantes et participants individuels ne sont pas nommés. Les personnes citées sont désignées par une catégorie (par exemple, scientifique, technologue) ou reçoivent un pseudonyme. Le nom du centre est révélé pour plusieurs raisons. La première est que les particularités de la recherche en réseaux de capteurs intégrés sont essentielles pour expliquer ses pratiques en matière de données ; le site est donc trop singulier pour être dissimulé aisément. Il aurait fallu, pour anonymiser le CENS, supprimer tant de contexte que les résultats en auraient été rendus vides de sens. Une seconde raison est de promouvoir le travail du CENS, puisque l'équipe sociotechnique appartenait à cette communauté. La National Science Foundation a reconnu la contribution du centre aux sciences sociales en plus de ses réussites scientifiques et techniques. En faisant du CENS un site exemplaire pour la recherche sur les pratiques en matière de données, elle a encouragé d'autres centres à accueillir des études en sciences sociales. Néanmoins, parce que le site de recherche était nommé, il a fallu prendre plus de précautions encore en agrégeant les résultats à un niveau qui dissimule l'identité des individus.

La conservation, le partage et la réutilisation des données

Les données de l'équipe sociotechnique, constituées d'enregistrements audio, de transcriptions, de notes de terrain, de fichiers NVivo et d'une variété de traces publiques et privées, sont stockées en sûreté sous le contrôle de l'équipe et de ses membres. Conformément aux recommandations de l'Institutional Review Board (IRB), les registres papier et les autres documents sont conservés dans des meubles de classements verrouillés dans des bureaux fermés. Les registres numériques sont stockés dans des serveurs sécurisés. Alors que l'IRB exige parfois que les données soient détruites en fin de projet, l'équipe renouvelle chaque année son autorisation pour continuer à analyser ses données et les comparer à celles recueillies ultérieurement. Selon les consignes actuelles, les données d'un projet ne peuvent plus être analysées une fois que l'autorisation de l'IRB a expiré.

En ce qui concerne les subventions impliquant plusieurs universités, chaque chercheur et chercheuse a dû obtenir une autorisation de l'IRB pour les données collectées par son équipe. Bien que les politiques universitaires soient globalement similaires, la possibilité de partager des données entre universités était limitée. De manière générale, les chercheurs pouvaient se transmettre des données encodées et anonymisées entre universités, mais pas les transcriptions et les autres traces comportant des informations permettant d'identifier les sujets humains. Plutôt que mettre en commun les données des collaborateurs et collaboratrices, chaque laboratoire universitaire participant gérait ses données isolément. Lorsque de nouveaux étudiants et étudiantes et chercheurs et chercheuses postdoctoraux rejoignaient une

équipe, ils devaient être certifiés et ajoutés au protocole IRB pour les jeux de données avec lesquels ils seraient amenés à travailler.

À l'heure où nous écrivons, les organismes de financement n'ont pas demandé que les données soient divulguées et aucun chercheur ou chercheuse hors de l'équipe n'a sollicité d'accès. Ces données représentent une mine de ressources pour l'équipe et sont utilisées en comparaison avec des études sur les pratiques en matière de données dans d'autres domaines. Lorsque le financement prendra fin et si les autorisations de l'IRB expirent sans être renouvelées, il sera difficile de les réutiliser, même par les spécialistes sociotechniques qui les ont produites. Les sujets de recherche ont signé des formulaires de consentement pour participer à ces études. Ces documents, qui font plusieurs pages et subissent un examen approfondi de la part de l'IRB, promettent, en échange de la participation à l'étude, la confidentialité des données permettant d'identifier les personnes dans l'analyse et la présentation des résultats.

En bref, ces données sont conservées pour l'usage de l'équipe de recherche. Elles ne peuvent être diffusées qu'aux individus figurant dans le protocole approuvé par l'IRB. Il est possible qu'une partie de ces données, comme les traces tirées de sites publics, soient diffusables un jour. Les documents obtenus dans le cadre des entretiens tombent sous le coup des accords de confidentialité des formulaires de consentement, sauf mention contraire expresse. Certaines notes de terrain pourraient demeurer utiles à d'autres si les noms et les autres informations identificatoires étaient retirés. En revanche, il est peu probable que les enregistrements audio, les transcriptions d'entretiens, les traces créées par les sujets de recherche et les autres informations étroitement liées à l'identité des personnes soient diffusés, pour des raisons éthiques et du fait des conditions de consentement qui ont permis d'obtenir ces données. Pris dans son ensemble, le jeu de données contient des informations pluridimensionnelles sur les activités du CENS lors de ses dix années de fonctionnement, ainsi que sur sa postérité. En revanche, s'il était étudié de manière parcellaire, le contexte et la provenance ne pourraient jamais être retracés.

Conclusion

La recherche sur Internet et la sociotechnique sont des domaines exemplaires des sciences sociales. Toutes deux appliquent des méthodes novatrices, interrogent le comportement humain et attirent des savantes et savants de diverses disciplines. Leurs infrastructures de la connaissance sont davantage caractérisées par un savoir partagé que par des infrastructures techniques ou des ressources informationnelles communes. Les accords sur les méthodes de recherche, telles qu'elles

s'incarnent dans les manuels et les cours universitaires, forment un substrat d'expertise partagée au sein des sciences sociales. Ces méthodes emploient des outillages techniques communs, comme les logiciels d'analyse de données et de modélisation informatique.

La recherche par sondage, qu'elle s'applique à l'usage d'Internet ou à n'importe quel autre comportement humain, s'appuie sur une longue tradition de savoir-faire méthodologique. Ces pratiques arbitrent la richesse des informations que l'on peut recueillir sur les personnes, le contrôle précis de la variance et la confidentialité des données. Les éléments probants sont rapportés avec une explication détaillée des plans de sondage, des populations, des questions posées, de la distribution des réponses et des conclusions au regard des résultats, le tout en garantissant l'anonymat des participantes et participants aux études. Les données issues de travaux présentant des niveaux adéquats d'assurance qualité peuvent être versées dans des archives de données ou diffusées directement par les chercheurs et chercheuses.

Les études sur les médias sociaux, comme les exemples sur le microblogage que nous avons présentés ici, s'appuient également sur de longues traditions méthodologiques, comme celles de l'analyse des réseaux sociaux. Les chercheurs et chercheuses venus d'autres champs apportent leurs propres traditions, comme les méthodes de topologie de réseau en informatique. Ces utilisations contrastées de ressources en données communes peuvent contribuer au transfert d'expertise entre disciplines ; mais ces démarches peuvent aussi avoir si peu en commun qu'elles deviennent opaques aux spécialistes d'autres champs. Le plus frappant dans cette comparaison est que chaque scientifique considérera comme des données des choses très différentes. Par exemple, on peut estimer qu'un lien entre des comptes Twitter met en évidence une influence, une structure de graphe, une communication illicite, une relation personnelle ou d'autres phénomènes. Le contenu d'un tweet ou d'un autre message sera traité, selon la question de recherche et la méthode, comme preuve d'une multitude de phénomènes. Déterminer quels messages représentent des conversations intentionnelles entre êtres humains, lesquels ont été échangés entre des êtres humains et des robots sociaux et lesquels constituent des échanges robotiques – et déterminer quand ces catégories ont une importance – requiert une expertise sociologique et technologique sur l'environnement étudié.

La sociotechnique, telle qu'elle s'est pratiquée dans les dix ans de travaux au sein du Center for Embedded Networked Sensing, applique à un problème un éventail de méthodes de recherche complémentaires. Dans notre cas, la problématique était de comprendre les pratiques en matière de données des nombreux participants et participantes du CENS et concevoir des outils et des services pour les

renforcer. Les questions et les méthodes de recherche ont évolué à mesure que la compréhension des pratiques s'améliorait, que les technologies étaient conçues et testées et que la recherche du CENS gagnait en maturité. L'infrastructure de la connaissance du centre incorporait une expertise partagée dans la science et la technologie des réseaux de capteurs intégrés, un espace collaboratif commun et d'autres formes d'assistance administrative et collégiale. L'investissement dans des ressources informationnelles a cependant été minime, puisque les technologies formaient, plus que les données, le terrain d'entente. L'équipe sociotechnique gérait son propre référentiel de données en interne. Ces données restent utiles pour des comparaisons avec d'autres environnements de recherche et pour poursuivre l'étude de la postérité du CENS. Leur richesse et la diversité de leurs genres les rendent cependant difficiles à partager. En effet, plus les informations sur des individus et des groupes sont complètes, moins elles sont anonymisables. Quand bien même on surmonterait ces difficultés, il est peu probable que l'on trouvera une demeure commune pour les enregistrements audio, les transcriptions, les photographies, les notes de terrain, les interviews codifiées, les publications, les posters scientifiques, les innombrables documents papier et numériques acquis auprès des sujets de recherche et les divers objets technologiques ; or, les disperser selon leur forme ou leur matière les rendra moins intéressants pour la recherche.

Au regard des provocations du chapitre 1, ces études de cas en sciences sociales montrent les diverses façons dont une même entité peut être traitée en donnée, la variété des méthodes de recherche applicables à une source de données commune et la diversité des infrastructures de la connaissance qui émergent pour faciliter ces types de travaux scientifiques. En ce qui concerne la première provocation, la propriété et le contrôle des données ne s'exercent pas de la même façon entre ces différents cas. L'Oxford Internet Institute recueillait les données des enquêtes OxIS et en conservait le contrôle. Le sondage ne saurait être reproduit, puisqu'il s'agit d'observations liées à des moments et des lieux spécifiques, mais il peut être répété plus tard avec de nouveaux échantillons. Les données d'OxIS sont diffusées après deux ans d'embargo, mais elles ne sont pas versées à des référentiels de sciences sociales. Les études sur le microblogage se prêtent moins à la reproductibilité ou au partage. Les travaux sociotechniques sur le CENS ne sont pas réplicables, mais les protocoles en ont été appliqués ultérieurement à d'autres sites de recherches pour effectuer des comparaisons. Ces données restent sous le contrôle des chercheurs et chercheuses et le partage en est limité par la réglementation sur les sujets humains.

La transférabilité, objet de la deuxième provocation, se joue aussi différemment selon les études de cas. Si les accords sur les méthodes de recherche fournissent un terrain d'entente à l'ensemble des sciences sociales, les techniques spécifiques

de manipulation des données – nettoyage, interpolation de valeurs manquantes, suppression des valeurs aberrantes, etc. – varient néanmoins suffisamment pour que l'analyse et l'interprétation des jeux de données divergent. L'évolution des fonctions de la communication savante, telle qu'évoquée dans la troisième et la quatrième provocation, semble plus graduelle que radicale dans les exemples présentés. Nos scientifiques s'inquiètent bien plus de publier leurs résultats que de diffuser leurs données. Elles et ils puisent à des sources de données plus volumineuses et variées que par le passé et recoupent des questions de recherche traitées dans les sciences exactes, la recherche en technologie et les sciences humaines. Malgré ces sources de données communes à plusieurs domaines, les différences de méthodes, de questions et de représentations peuvent limiter la découvrabilité et la réutilisation des jeux de données créés.

Ces études de cas montrent aussi la réorientation de l'expertise nécessaire à la recherche en sciences sociales, exposée dans la cinquième provocation. La combinaison de savoir-faire qualitatifs et quantitatifs dont il est fait preuve varie selon les études. Un certain degré de compétence informatique est nécessaire dans chacun de ces domaines, que ce soit pour développer de nouveaux outils, rédiger des scripts dans des outils existants ou exploiter des routines statistiques complexes. La capacité à réutiliser ces données est contrainte par l'accès à ces outils, ces scripts et ces routines.

Enfin, dans les domaines que nous abordons ici, les infrastructures de la connaissance et le partage de données en particulier attirent relativement peu d'investissements. La découvrabilité et la conservation des données de sondage se taillent la part du lion et leur infrastructure est encore loin d'être complète. Les médias sociaux représentent le front de recherche qui évolue le plus vite. Les objets d'étude et les outils pour l'examiner se transforment trop vite pour les partager fructueusement. Combiner des données issues d'études multiples ou effectuer des méta-analyses implique un investissement substantiel dans l'intégration des données. Dans tous les exemples cités, les jeux de données tendent à rester sous le contrôle des personnels de recherche, que ce soit du fait d'une préférence, d'un manque de solutions alternatives, de problèmes de confidentialité insolubles ou d'un mélange de ces raisons. De ce fait, ces spécialistes des sciences sociales ont peu de motivations pour investir dans des schémas de métadonnées ou de classification qui faciliteraient le transfert de leurs données vers d'autres domaines. Une vision à plus long terme des besoins des infrastructures de la connaissance dans les sciences sociales doit se préoccuper de l'acquisition et de la gestion des ressources en données au sein de chacune de ces disciplines, entre elles et avec d'autres domaines scientifiques, ainsi qu'entre ressources

publiques et privées. Compte tenu de la dilution des sources de données, le flou des frontières disciplinaires, le caractère politique et sensible des sujets et la diversité des parties prenantes, les investissements dans les infrastructures de la connaissance des sciences sociales s'annoncent litigieux dans les années à venir.