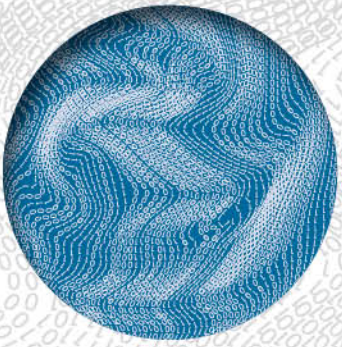
# Anonymat et confidentialité des données : l'expérience de beQuali

Selma Bendjaballah, Sarah Cadorel, Emilie Fromont, Guillaume Garcia, Emilie Groshens, Emeline Juillard

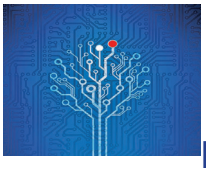
# La diffusion numérique des données en SHS

## Guide des bonnes pratiques éthiques et juridiques

sous la direction de
Véronique Ginouvès & Isabelle Gras



DIGITALES



DIGITALES

# La diffusion numérique des données en SHS

## Guide des bonnes pratiques éthiques et juridiques

sous la direction de

Véronique Ginouvès & Isabelle Gras

2018

Presses Universitaires de Provence

© Presses Universitaires de Provence
Aix-Marseille Université

29, avenue Robert-Schuman – F – 13621 Aix-en-Provence CEDEX 1
Tél. 33 (0)4 13 55 31 91

pup@univ-amu.fr – Catalogue complet sur presses-universitaires.univ-amu.fr

DIFFUSION LIBRAIRIES : AFPU DIFFUSION – DISTRIBUTION SODIS

# Anonymat et confidentialité des données

## L'expérience de beQuali

Selma Bendjaballah, Sarah Cadorel, Émilie Fromont, Guillaume Garcia, Émilie Groshens, Emeline Juillard
Centre de données socio-politiques, Sciences Po-CNRS

**Abstract**: *This experience, in the context of beQuali, CDSP's surveys bank, proposes to give a feedback of the concrete questions we have faced, for the anonymisation of qualitative data in sociology and political science. We highlight the resulting tensions, between scientific interest (the need to preserve the sociological wealth of data for reuse) and ethical and legal issues (the protection of respondents and researchers). It is a collective reflection based on the cross-examination of several cases of surveys (a dozen) already treated or being processed, and on the study of how other researchers, in the literature in particular, and other equivalent platforms, in France and abroad, have addressed this issue.*

## Introduction

Ce texte propose un retour d'expérience sur les problématiques d'anonymisation que nous gérons dans le développement de la banque d'enquêtes qualitatives en sciences sociales « beQuali ». Nous y restituons les questions que nous affrontons, ainsi que les solutions mises en œuvre, lorsque nous devons traiter des données qualitatives portant sur des individus « enquêtés » lors de recherches en sociologie et en science politique. Nous y explicitons les difficultés à articuler deux principes d'égale importance : préserver la précision des données et donc leur potentiel de réutilisation – ce qui renvoie à une nécessité d'ordre scientifique – et protéger les enquêtés et les chercheurs producteurs – ce qui renvoie à un impératif d'ordre juridique avant tout, mais aussi éthique ou déontologique [1]. L'anonymisation de telles « données », loin de se résumer à une opération technique, conséquence de l'application mécanique du droit, constitue un enjeu complexe, au carrefour de logiques hétérogènes. À défaut de pouvoir être standardisée, elle implique un traitement au

1 S'il est également question, dans la suite du texte, de contraintes d'ordre éthique ou déontologique, nous centrerons notre propos plus spécifiquement sur les contraintes juridiques, à la fois par manque d'espace mais aussi parce que les premières sont moins claires – du fait de leur non-codification – dans les disciplines visées, en particulier la sociologie et la science politique – les projets de charte de déontologie de l'Association française de sociologie (2009-2010) ou de l'Association française de science politique (2009) ayant par exemple achoppé.

cas par cas, via des réglages ad hoc, à la recherche d'un équilibre à chaque fois sur mesure. Pour ce faire, nous restituons les fruits d'une réflexion collective basée sur l'examen croisé de plusieurs cas de figure illustrés par des enquêtes déjà traitées ou en cours de traitement à beQuali.

## Mettre à disposition des données d'enquêtes qualitatives en sciences sociales

Depuis une vingtaine d'années, les dispositifs de mise à disposition des données d'enquêtes qualitatives en sciences sociales se sont multipliés dans le monde (Bishop et Kuula 2017). En France, beQuali a été créé en 2010 au sein du Centre de données socio-politiques (CDSP, UMS 828 Sciences Po-CNRS). Outil national, beQuali propose à la communauté scientifique, sous réserve d'autorisation, des archives d'enquêtes de sociologie et de science politique menées à partir de méthodes qualitatives, ainsi qu'une documentation restituant le contexte de leur production. L'objectif est de favoriser la réutilisation de ces archives à des fins scientifiques ou pédagogiques, pour produire de nouvelles recherches ou servir de ressources pour (se) former aux sciences sociales. À cette fin a été mis en place un site web (http://www.bequali.fr/) permettant d'explorer en ligne les corpus d'enquêtes, qu'il est également possible de télécharger depuis le portail Quetelet (https://quetelet.casd.eu/fr). Le traitement de ces archives, afin qu'elles puissent être proposées à la réutilisation, pose des problèmes particuliers que nous allons décrire.

Les données d'enquêtes qualitatives (transcriptions d'entretiens, notes d'observation, etc.) résultent de récits approfondis ou de descriptions fines des personnes, groupes ou situations étudiés. De par leur ancrage disciplinaire, les enquêtes de sociologie ou de science politique comportent fréquemment des données personnelles et « sensibles » – car touchant aux opinions politiques, à la religion, à la sexualité, à la santé, à l'ethnicité, aux difficultés sociales des personnes – au sens des diverses législations encadrant la communication de ces catégories d'informations. On se trouve face à un paradoxe, puisque si l'injonction d'une ouverture de ce type de données de recherche publiques est de plus en plus présente (loi pour une République numérique[2], directives européennes sur l'*open research data* et l'*open access*, loi du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public dite loi Valter[3]), la législation encadrant leur mise à disposition est en revanche très restrictive puisqu'elle en interdit la diffusion au bénéfice du droit et de la protection des enquêtés (Règlement européen sur la protection des données – transposée en France par la loi informatique et libertés[4] – qui sera remplacé par l'application en 2018 du Règlement européen relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, nouvelle loi pour une

2 Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, en particulier l'article 6 relatif aux données de la recherche et l'article 30 qui concerne l'*open access*.

3 Loi n° 2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public.

4 Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

République numérique, législation sur le droit à l'image [5], législation sur le respect de la vie privée [6], loi CADA et Code du patrimoine [7] en particulier).

Devant l'impossibilité de mettre à disposition des données personnelles et sensibles, la solution consiste à les anonymiser. Les prescriptions juridiques ne disent cependant rien, ou presque, de la manière dont il convient de procéder pour anonymiser ces données à un niveau suffisant, *i.e.* sans risque pour les enquêtés, pour pouvoir les publiciser dans le respect de la loi. En particulier, les préconisations de la CNIL consistant à supprimer les informations indirectement identifiantes – celles qui, mises ensemble, peuvent potentiellement permettre d'identifier des individus par recoupements, selon une logique probabiliste – restent floues. Il existe une importante marge d'interprétation dans l'application des principes et dans l'évaluation du risque d'identification. Ce flou est d'autant plus problématique dans le cas des données de recherche, dont l'objectif de la diffusion est avant tout de garantir les possibilités de leur exploitation future. Il s'agit donc d'équilibrer le risque d'identification des personnes avec celui de perte d'informations, dans une tension entre contraintes juridiques et scientifiques. Nous y reviendrons.

Face à l'obligation pratique de répondre à ce flou, nous sommes partis des pratiques des professionnels ayant déjà affronté ces questions en réalisant un état des lieux à partir des publications produites sur ce sujet. Il apparaît clairement une plus grande maturité de la réflexion dans le monde anglo-saxon ou l'Europe du Nord. Sans chercher l'exhaustivité, nous avons identifié au moins une trentaine de publications traitant directement du sujet sur les vingt dernières années [8], formant l'ossature d'un débat adossé à l'existence, beaucoup plus ancienne (aux États-Unis, au Royaume-Uni, en Finlande, etc.), de dispositifs similaires à beQuali. On trouve même des guides de bonnes pratiques et des recommandations sur les sites de certaines plateformes, à destination des déposants potentiels d'enquêtes [9]. Mis à part des considérations générales sur les enjeux scientifiques et éthiques de l'anonymisation, ou des conseils génériques sur sa mise en œuvre pratique, ces ressources ne proposent toutefois que peu de retours réflexifs sur des cas concrets et problématiques – le plus souvent liés à des jeux de données quantitatives [10]. De son côté, la littérature francophone est très limitée et aborde le sujet essentiellement sous l'angle des publications, c'est-à-dire des problèmes qui se posent au moment de rendre compte, dans une publication, d'informations – partielles – sur les enquêtés, prélevées et sélectionnées dans les données brutes (Baude 2006 ; Béliard et Eideliman 2008 ; Weber 2008, 2010 ; Roux 2010 ; Zolesio 2011 ; Monge 2016 ; Coulmont 2017). En France, on n'a encore aucune idée ou définition précises

5 Notamment extraite de l'article 9 du Code civil, du Code pénal et de l'article 38 de la loi informatique et libertés.

6 Notamment l'article 9 du Code civil et l'article 8 de la Convention européenne des droits de l'homme et des libertés fondamentales.

7 Chapitre 3 du livre 2 du Code du patrimoine (régime de communicabilité : les archives publiques sont communicables de plein droit sauf cas listé dans l'article L213-2).

8 Pour un panorama voir Bishop (2009), Kuula (2011), Moore (2012), Saunders *et al.* (2014).

9 Voir par exemple le Finnish Social Science Data Archive (FSD), « Anonymisation and Identifiers Policy » : http://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html

10 Voir par exemple Clark (2006) et ICO (2012).

de ce que recouvre concrètement l'anonymisation (Béliard et Eideliman 2008), une pratique pourtant généralisée dans le travail de recherche en vue de la restitution des résultats. Cette situation assure à chacun une certaine liberté en coulisses, dans sa pratique individuelle, mais limite en contrepartie l'intelligence collective sur ce sujet éminemment complexe, laissant presque complètement de côté la question des conditions d'archivage des données brutes. Cette littérature révèle tout de même l'existence de normes déontologiques générales, assez largement partagées semble-t-il dans les communautés académiques concernées, autour notamment du « contrat de confiance » liant enquêteur et enquêté (Weber 2010 : 260 et s.). Ce contrat, souvent implicite, implique de protéger l'enquêté, la confidentialité des propos qui pourraient lui porter atteinte, ou porter atteinte à ses proches. L'existence de cette norme professionnelle, qui est d'ailleurs régulièrement mobilisée par les chercheurs déposants, nous oblige à être plus protecteur des enquêtés ou des enquêteurs que ce à quoi nous enjoint la CNIL[11]. En effet, au-delà des données personnelles et sensibles définies juridiquement, nous sommes amenés à masquer certaines informations traitant de l'histoire intime des enquêtés et potentiellement sensibles pour eux, tout comme des informations explicitement livrées par les enquêtés sous le mode du « *off* »[12]. Le dépôt des enquêtes pour leur mise à disposition suppose que soit rendue possible une extension de ce contrat de confiance. Parmi les conditions nécessaires à cette extension, figure notamment la participation active des chercheurs producteurs à l'examen détaillé des informations susceptibles d'être anonymisées, voire purement et simplement retirées. Ainsi, à côté des prescriptions juridiques, doivent également être prises en compte des considérations d'ordre déontologiques, tout en tenant compte de l'évaluation qu'on peut faire, au moment de la diffusion des données – et parfois de nombreuses années après la réalisation de l'enquête –, des risques professionnels ou personnels que comporte, pour les enquêtés, la divulgation de certaines informations qui ne sont pas explicitement couvertes par la législation[13].

Lorsqu'il s'est agi d'affronter l'épreuve de l'anonymisation, nous disposions certes d'un certain nombre de repères, mais la difficulté a été de les convertir en arrangements pratiques, autour de quelques principes que nous allons maintenant décrire, à partir de cas exemplaires qui seront décrits plus loin.

---

11 On rejoint ici cependant, en partie du moins, les attendus de la législation sur le respect de la vie privée.

12 La notion de « sensibilité » des données est donc à entendre d'au moins deux points de vue : celui, *a priori* et général, des différentes législations encadrant la communication des informations personnelles ; celui des chercheurs producteurs qui vont procéder à une évaluation au cas par cas, en considérant la spécificité des problématiques et des terrains d'enquêtes.

13 En nous inspirant des propos de Florence Weber (2010 : 269), nous considérons que ce qui, pour les enquêtés, peut revêtir une importance très grande à un moment donné perd probablement une grande partie de sa pertinence de nombreuses années après, à la suite de changements importants de contexte.

## Les principes d'anonymisation suivis à beQuali : quelques exemples

Un constat s'impose : les enquêtes qui ne posent pas de problème de confidentialité sont rares. Nous avons reçu en dépôt quelques enquêtes dont les producteurs avaient déjà obtenu les autorisations des enquêtés – c'est le cas de Quand des Français, des Anglais et des Belges (francophones) parlent d'Europe[14] – ou procédé eux-mêmes à une pré-anonymisation des données – c'est aussi le cas de Représentation du champ social, attitudes politiques et changements socio-économiques[15] et de Deux Générations d'immigrés Africains[16]. Même dans ce dernier cas de figure, nous devons vérifier le degré d'anonymisation afin de la mettre en conformité avec les attendus juridiques et déontologiques de la diffusion des données.

Dans d'autres cas, lorsqu'il s'agit d'enquêtés ayant une importante notoriété[17], les données ne sont tout simplement pas anonymisables. S'efforcer de réduire la probabilité d'identification supposerait de supprimer une telle quantité d'informations qu'on appauvrirait démesurément les données, avec le risque que l'anonymisation demeure fictive. Pour pouvoir mettre ces enquêtes à disposition, nous devons obligatoirement disposer des autorisations explicites des enquêtés, même si elles sont obtenues *a posteriori*. Deux expériences positives – celle de L'Europe saisie par les rôles parlementaires[18] et Des femmes en politique[19] – montrent que cela est faisable, la quasi-totalité des personnes contactées, des années, voire des décennies après l'enquête, ayant donné leur consentement par écrit pour la diffusion de leur témoignage.

Pour les enquêtes mobilisant des enquêtés « ordinaires » nous disposons rarement de tels consentements écrits, les chercheurs ayant – encore – rarement recours à ce type de pratique. Face à la difficulté de pouvoir retrouver les enquêtés afin d'obtenir leur autorisation, l'anonymisation est souvent la solution privilégiée. Afin de réduire les risques de reconnaissance et de se conformer à la loi, nous avons adopté une position scrupuleuse sur les marqueurs d'identification directe au sens de la CNIL (nom, prénom, date précise de naissance, etc., sont systématiquement supprimés), mais plus souple sur les marqueurs d'identification indirecte (lieu de résidence, profession, affiliation partisane, syndicale ou associative, etc.). Nous

14 Enquête dirigée Sophie Duchesne, par entretiens collectifs filmés, dont les vidéos sont disponibles et pour laquelle les enquêtés avaient fourni un consentement écrit.

15 Enquête de Jean-Marie Donégani, Guy Michelat et Michel Simon, par entretiens individuels avec des citoyens ordinaires sur leurs attitudes politiques.

16 Enquête dirigée par Jacques Barou, par entretiens individuels et collectifs croisés mêlant les mêmes membres de plusieurs familles immigrées (2018).

17 Nous considérons ici une importante notoriété comme la caractéristique d'une personne occupant une fonction, une position, un poste, de premier plan ou unique, aisément reconnaissable au sein d'un milieu social donné, pour peu que ce milieu bénéficie d'une publicité suffisante, à un degré tel que n'importe qui, *a priori* et avec un minimum d'effort, pourrait identifier cette personne. Entrent par exemple dans ce cas de figure des personnes occupant des fonctions uniques ou de premier plan dans le monde politique, des médias, du sport, de la haute fonction publique.

18 Enquête d'Olivier Rozenberg, par entretiens individuels avec des parlementaires et ministres et quelques députés français au Parlement européen.

19 Enquête de Mariette Sineau, par entretiens individuels avec des femmes politiques.

supprimons donc le strict minimum de ces marqueurs d'identification indirecte de manière à nous conformer à la loi tout en préservant la valeur heuristique des données. Dans une logique juridique et éthique, nous sommes soucieux de protéger les enquêtés, les enquêteurs, les personnes citées et les auteurs de l'enquêtes. À ce titre, nous veillons particulièrement à respecter le contrat de confiance établi entre enquêteurs et enquêtés. À cette fin, la collaboration des chercheurs producteurs à ces opérations est nécessaire, c'est la raison pour laquelle ces derniers sont systématiquement consultés pour élaborer les protocoles d'anonymisation ou de retrait des informations.

Ces opérations ont pour objectif de permettre la réutilisation des enquêtes à des fins de recherche et d'enseignement. Il est donc crucial de maintenir un équilibre difficile entre protection des personnes et valeur informative des données. Pour en rendre compte, nous allons restituer les plans d'anonymisation mis en œuvre pour quatre enquêtes [20] comportant des données sensibles liées à la nature des informations personnelles, au statut ou aux fonctions sociales des enquêtés (qui les rendent plus facilement reconnaissables), aux implications que peuvent avoir les méthodes d'enquête sur le risque de reconnaissance, ou encore au caractère d'actualité des enjeux en cause pour les intéressés [21].

L'enquête Choisir son école d'Agnès van Zanten [22] a été menée entre 1999 et 2005 dans l'Est et l'Ouest parisien, dans le but de comprendre pourquoi et comment les parents respectent ou contournent la carte scolaire. En étudiant des communes limitrophes deux à deux (Vincennes et Montreuil, Rueil Malmaison et Nanterre), il s'agissait de saisir les phénomènes de fuite d'une ville à l'autre. Deux groupes ont été interrogés : des parents d'élèves et des représentants d'associations de parents d'élèves ; des acteurs de l'institution scolaire (chefs d'établissement, professeurs, inspecteurs de l'éducation nationale et élus locaux). Quatre-cents documents composent le corpus, dont cent-cinquante transcriptions d'entretiens principalement avec des parents, et une dizaine de notes de terrain issues de rencontres avec des agences immobilières et des acteurs de l'éducation (mairie, commission de dérogation, réseau éducation prioritaire).

L'anonymisation de cette enquête s'est avérée complexe du fait de la diversité des publics étudiés et de sa dimension ethnographique (archiver les matériaux n'avait de sens que si l'on conservait les informations relatives aux propriétés situées – et relatives – des établissements et des zones d'habitation, qui sont centrales pour l'enquête). Pour ce qui concerne les parents, les informations géographiques ont été le plus souvent conservées, ainsi que certaines informations sociologiques. Nous avons considéré que les informations potentiellement sensibles (notamment les pratiques peu avouables liées au contournement de la carte scolaire ou les opinions

---

20 Nous n'en indiquons ici que les grandes lignes. Ponctuellement, des exceptions ou ajustements ont dû être faits, au cas par cas, en fonction de la combinaison particulière d'éléments potentiellement identifiants, qu'il n'est pas possible de restituer ici.

21 Les informations en cause dans les enquêtes que nous avons pour l'instant traitées ne sauraient être considérées comme particulièrement sensibles (par exemple, actes fortement illégaux). Plus largement, les enquêtes diffusées ne portent pas sur des petits milieux d'interconnaissance auxquels les enquêtés appartiendraient encore au moment de la diffusion des données.

22 Enquête « Choisir son école » d'Agnès Van Zanten [En ligne] http://bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bequali_s1/

politiques) demeuraient peu préjudiciables pour les intéressés et pouvaient être conservées, à condition de réduire le niveau de précision d'autres informations sociographiques (noms des employeurs, des associations de parents d'élèves ou des proches cités).

Pour les acteurs de l'institution scolaire, le nom de la ville et le poste occupé ont été autant que possible conservés, leurs témoignages livrant des informations inséparables des spécificités des secteurs scolaires au sein desquels ils œuvraient. Dans la mesure où ils livraient parfois des informations relativement sensibles sur le fonctionnement des établissements et certains arrangements avec le respect de la carte scolaire (par exemple, les coulisses des commissions de dérogation) et des principes de mixité sociale et ethnique, ce qui pouvait leur porter préjudice, le nom de l'établissement a pourtant été supprimé[23].

Tableau récapitulatif des éléments d'anonymisation de l'enquête Choisir son école

| **Types d'enquêtés** | **Données conservées** | **Données supprimées** | **Justification de la conservation de l'information** | **Justification du masquage de l'information** |
|---|---|---|---|---|
| **Parents d'élèves** | Données géographiques : ville, quartier, école<br>Données sociographiques : profession, délégué de classe, membre du CA d'établissement | ID, *nationalité*, *employeur (nom, lieu)*, nom des relations, *nom de l'association* | Les enquêtés occupaient des fonctions éphémères.<br>Difficulté d'accéder aux archives des écoles. | Enjeu de réputation (contournement de normes, opinions politiques, religieuses, etc.). |
| **Représentants d'associations de parents** | Ville, école, association/ responsabilité | ID, *nom de l'association (si président)* | | Risque d'identification si président (fonction unique et non éphémère). |
| **Acteurs scolaires (enseignants, directeurs, inspecteurs), élus locaux, agences immobilières** | Ville, *poste occupé* | *ID (sauf inspecteurs et maires)*, école | Impossibilité d'anonymiser les responsables hauts placés.<br>Risque mineur : parole contrôlée | Choix de privilégier le critère géographique par rapport au critère sociographique. |

**Légende - ID : identifiants directs (noms, coordonnées) ; *italique* :** évaluation au cas par cas du risque d'identification

L'enquête Dilapidation et prodigalité[24] d'Anne Gotman, réalisée au début des années 1990, porte sur la dilapidation d'héritages et plus largement sur les

23 Quelques cas étaient impossibles à anonymiser (inspecteurs, directeurs d'établissement), et les consentements difficiles à obtenir *a posteriori*. Dans la mesure où ces entretiens étaient secondaires pour la chercheuse, et peu sensibles (car ne livrant que des informations très générales sur les politiques officielles d'éducation et datant de plus de 15 ans), nous ne les avons pas anonymisés.

24 L'enquête est en cours de traitement et sera prochainement disponible.

phénomènes de prodigalité et de surendettement (pouvant éventuellement aboutir à des procédures de sauvegarde de justice). Plusieurs terrains ont été menés en région parisienne et dans une ville moyenne de l'ouest de la France, mêlant des témoignages de dilapidateurs, d'informateurs sur la dilapidation, de familles surendettées bénéficiant de l'aide sociale, de membres d'un groupe de « débiteurs anonymes », ainsi que d'« experts » (assistantes sociales, psychiatres, juristes et juges de tutelles)[25]. Les témoignages ont été obtenus principalement par entretiens (une soixantaine) parfois complétés par des questionnaires (une trentaine). Les archives comprennent également des notes prises sur des dossiers de demande de mise sous tutelle (une trentaine) ainsi que quelques notes d'observations. Une grande partie des données comporte des informations personnelles plus ou moins sensibles sur la vie privée des enquêtés et leur histoire familiale[26]. Le protocole d'anonymisation a été modulé selon les composantes de l'enquête. Globalement, le choix a été fait de sacrifier les informations sur les lieux – l'enquête n'ayant pas de dimension géographique – au profit des informations sociologiques sur les enquêtés, de manière à préserver la richesse des indications sur leurs positions sociales (profession, statut marital, etc.). Cela a été suffisant la plupart du temps, notamment parce que nous avons affaire à des individus dotés d'une faible notoriété et que l'enquête date d'il y a plus de 25 ans. Le cas des dilapidateurs s'est avéré plus complexe car il s'agit de récits de vie très longs, donnant des détails précis sur l'histoire de la famille élargie, le risque de reconnaissance étant accentué par l'appartenance de certains enquêtés à la grande bourgeoisie, donc caractérisés par un potentiel élevé de notoriété[27]. Il a fallu, au cas par cas, identifier les éléments susceptibles de favoriser des ruptures d'anonymat, et réduire la précision de l'information sur de nombreux éléments biographiques, aussi bien des enquêtés que des membres de leur famille, voire de leurs réseaux de sociabilité et des personnes citées.

L'enquête Formation des couples réalisée par Michel Bozon et François Héran entre 1982 et 1985 contient des données liées à l'intimité des personnes, notamment autour du thème de la sexualité. Cette enquête a été conduite en trois temps avec des méthodes différentes : une phase préliminaire de 28 entretiens semi-directifs, une enquête par questionnaires en face à face auprès de 2 924 personnes et une enquête post-quantitative par entretiens auprès de certains répondants au questionnaire. Les données à caractère sensible peuvent être liées, outre au thème, à l'interconnaissance entre les auteurs de l'enquête et les répondants aux premiers questionnaires ainsi qu'au recoupement entre la seconde phase d'entretiens et les questionnaires. Pour cette raison, nous avons décidé d'anonymiser les identifiants directs des enquêtés (nom, adresses et numéros de téléphone) et le nom des enquêteurs des questionnaires. En revanche, l'ancienneté de l'enquête

25 Anne Gotman a également réutilisé des entretiens de précédentes enquêtes sur l'héritage et sur l'accession à la propriété dans le parc du logement social.

26 Par exemple : conflits familiaux liés à des histoires d'argent, honte sociale causée par le dérèglement de la dépense et la mise sous tutelle, etc.

27 Une autre spécificité de cette enquête est un lien de connaissance – et donc de confiance – particulièrement perceptible entre la chercheuse et certains des enquêtés. On le relève grâce aux tutoiements dans les entretiens, au fait que les enquêtés nomment par leur prénom des personnes de leur entourage que connaît la chercheuse, et à des demandes explicites que lui font les enquêtés de garder une confidence pour elle.

(plus de trente ans) et le caractère restreint de la diffusion nous ont conduits à laisser les informations liées à la ville d'habitation et aux employeurs lorsqu'il s'agit de personnes morales (noms de grandes entreprises locales, etc.). En effet, un des principaux résultats de l'enquête étant la démonstration de l'endogamie et de l'homogamie dans la formation des couples, cela nous a semblé important de conserver ces informations idoines cruciales (profession, lieux de sociabilité et de rencontre, lieu d'habitation, etc.) pour comprendre les caractéristiques sociales des enquêtés car il s'agit d'éléments centraux dans l'analyse qui a été faite des données.

L'enquête Comparaison des ministères de l'Enseignement supérieur de France et d'Allemagne a été menée par Christine Musselin et Erhard Friedberg en 1993. Elle interroge la manière dont l'enseignement supérieur – établissements universitaires, programmes, ressources, orientations scientifiques, profession et statuts, etc. – est piloté, en France et dans trois Länder allemands (Basse-Saxe, Rhénanie-du-Nord-Westphalie, Bade-Wurtemberg) à la fin des années 1980. D'un point de vue méthodologique, l'équipe de recherche a adopté une approche organisationnelle qui se concentre sur le fonctionnement interne des administrations de tutelle chargées de l'enseignement supérieur. Au total, plus de 150 entretiens de hauts fonctionnaires des ministères français et allemands ont été conduits. Cette sociologie des organisations implique que les entretiens questionnent les rapports entre les directions et les acteurs chargés des universités et ils contiennent souvent des éléments sensibles sur les coulisses des prises de décision et les jeux de pouvoir entre collègues de services, comme l'atteste l'extrait d'entretien ci-dessous :

partir. Là, je suis plutôt content d'être venu. Mais je trouve le cabinet insupportable.

Question

Cela a-t-il une répercussion sur votre travail ?

Réponse :

Oui, par exemple, l'attaché parlementaire me demande si j'ai reçu les questions parlementaires. Il y a des retards et c'est intolérable. Il faut dire qu'elle ne s'entend pas avec l'autre membre du cabinet qui s'en occupe. Ca, ça rejaillit sur nous et c'est la partie la moins agréable. De plus, cela correspond à une séparation géographique des bureaux qui n'est pas bonne. Ca retarde. Ce cabinet ne fonctionne pas bien et ce n'est pas seulement lié à une personne. Mais j'aimerais que vous n'en parliez pas. Cet avis n'est pas lié à des opinions politiques. Moi, j'ai mon travail à faire et c'est tout. Et il y a des choses bizarres. Quand on pense que la direction générale n'a pas communication systématique des discours du ministre. Voyez, ce sont des petites choses come ça, ce n'est pas un bon cabinet. De plus, il n'y a pas de gens de l'administration minis-

Figure 1 : Enquête Comparaison des ministères de l'Enseignement supérieur en France et en Allemagne (C. Musselin et Erhard Friedberg). Extrait de l'entretien n° 4, Service administratif et financier du ministère français (cdsp_bq_s4_col_entr_indv_fr_entretien4_trans).

Dans ce cas, ce ne sont pas tant les informations personnelles des enquêtés qui sont sensibles, que les propos qu'ils tiennent sur leur environnement de travail qui

pourrait leur porter préjudice ; l'anonymisation consiste donc à protéger l'individu dans son milieu professionnel. Les noms des enquêtés ont été supprimés, et n'ont été conservées que les fonctions occupées (« un directeur », « un chef de bureau du budget », « un correspondant universitaire », etc.). Ainsi nous nous conformons à la manière dont les chercheurs les ont désignés dans leur publication (Friedberg et Musselin 1993), et préservons autant que possible les informations organisationnelles inhérentes à la sociologie des organisations et indispensables à la réutilisation des données de cette enquête. S'agissant des informations géographiques, les entretiens sont disponibles dans la langue où ils ont été conduits, et les noms de pays et Länder sont conservés.

Au total, ces cas de figure illustrent la diversité des configurations qu'il est nécessaire de gérer d'une enquête à l'autre, et les arrangements spécifiques qu'il faut nécessairement opérer pour trouver un équilibre entre les contraintes juridiques et scientifiques. Selon les particularités de l'enquête et des objectifs du ou des chercheurs, c'est par exemple tantôt le degré de précision des informations géographiques, tantôt celui des informations sociographiques qui est diminué, afin de conserver un potentiel de réutilisation optimale tout en protégeant raisonnablement (notamment, mais pas seulement) les enquêtés du risque de reconnaissance et surtout des préjudices qui pourraient en découler pour eux. Dans cet ajustement, nous intégrons d'autres paramètres, comme le degré de sensibilité des informations pour les enquêtés pondéré par l'ancienneté de l'enquête pour arriver à un résultat *ad hoc* pour chaque enquête, voire pour chaque composante du corpus.

## Retour sur les opérations concrètes d'anonymisation et de respect de la confidentialité

Les problèmes et les opérations d'anonymisation ont aussi, inséparablement, une dimension pratique qui pose parfois des difficultés que nous allons illustrer à l'aide de cas concrets. Pour ce faire, revenons sur la spécificité des matériaux que nous devons traiter. Dans beQuali, l'unité « enquête » est composée d'un ensemble de documents d'archives allant au-delà des matériaux de terrain, c'est-à-dire des « données brutes [28] », et qui renseignent aussi bien sur la préparation de l'enquête [29] que la conduite du terrain [30], ou encore la phase d'analyse [31] préalable à la publication des résultats. Ces archives peuvent être nativement numériques ou conservées sur des supports physiques (papier, cassettes audio, photos, etc.) et concerner une diversité de formats (textuels, sonores, visuels, etc.). Réunies au sein d'un corpus, elles forment une unité intellectuelle et matérielle reconstruite *a posteriori*, mêlant documents spécialement produits pour mener à bien la recherche et documents extraits (à la pièce) des fonds de chercheurs – par exemple des documents plus personnels, comme de la correspondance, des journaux de bord, etc. – ou de

28 Entretiens, notes ethnographiques, photos, etc.

29 Projets de recherche, dossiers de financement, prise de contact avec des institutions ou des informateurs, etc.

30 Documentation collectée sur le terrain, correspondance, etc.

31 Tableaux récapitulatifs, fiches de synthèses, résumés, etc.

laboratoires – des documents plus administratifs, comme des rapports d'activité, des documents comptables, etc. Ces archives forment un maillage informationnel dont la complexité résulte de la multiplicité des documents (souvent plusieurs centaines), des recoupements potentiels de données (nombreuses références croisées), de la diversité des formats de notation de l'information, ou encore de la nécessité d'intégrer les métadonnées comme éléments susceptibles de dévoiler des informations sur les enquêtés.

Pour commencer, lorsque nous collectons les archives des enquêtes, celles-ci mêlent documents physiques et/ou numériques. Nous sommes tenus par le respect de l'intégrité des archives : toutes les opérations d'anonymisation sont faites soit sur des copies, soit sont réversibles quand il s'agit des originaux. Elles peuvent se faire avant la numérisation, ou après sur des fichiers numériques.

Avec des documents manuscrits ou « sales », qu'il n'est pas possible ou trop coûteux de soumettre à une reconnaissance optique des caractères, l'anonymisation est faite directement sur les documents, avec des post-its de couleur. L'information est complètement masquée, pour des raisons essentiellement techniques.

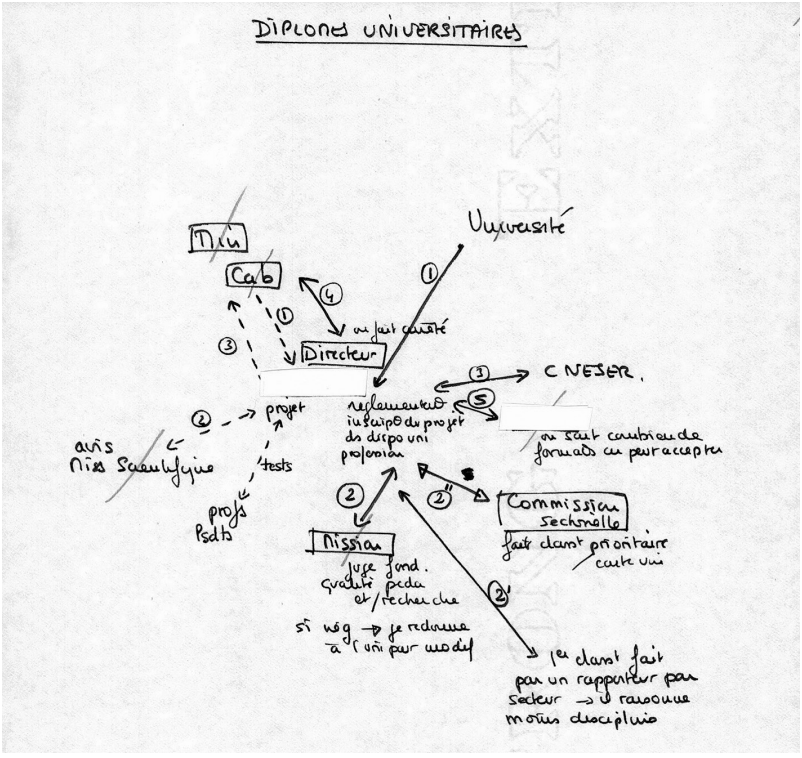



**Figure 2** : Enquête Comparaison des ministères de l'Enseignement supérieur en France et en Allemagne (Musselin et Friedberg). Sociogramme des diplômes universitaires (France) (cdsp_bq_s4_ana_inte_figu_fr_sociogr_fiche1).

Avec des documents manuscrits ou mixtes (tapuscrit-manuscrit) qu'il est possible de soumettre à une reconnaissance optique des caractères, ou des documents nativement numériques, l'anonymisation est réalisée directement sur une copie du fichier numérique, avec davantage de marge de manœuvre puisqu'il est possible de remplacer facilement l'information originelle par une autre. Nous utilisons ici des balises de remplacement de l'information selon la technique de l'hyperonyme, du type ((anonym : hyperonyme)).

> Enquête De l'Afrique à la France : deux générations d'immigrés africains de Jacques Barou *et al*. Exemple de balisage par hyperonyme.
>
> **((anonym : Prénom Nom)), née en 1969 est une ravissante jeune femme aux dents très blanches, caissière depuis 12 ans au Franprix de la rue ((anonym : nom de la rue)).**
>
> Dans ce petit supermarché, les relations entre clients et caissières stables sont très amicales, je vois que les gens les embrassent, les complimentent, les draguent.
>
> Donc elle accepte le rendez-vous dans le café d'en face, où nous nous retrouvons le vendredi 12 janvier à 15 h 30, nous serons chassées par les élèves du lycée ((nom du lycée)) à 16 h 30 qui viennent en bande, faire du bruit et fumer.
>
> Elle est née à Dakar.
>
> Ses grands-parents des deux côtés sont Casamançais, et catholiques.
>
> Son père est venu à Dakar, il s'appelle ((anonym : Prénom)), il est né en 1920, il a été engagé comme tirailleur sénégalais pour la guerre de 39-40 et pour l'Indochine ensuite.
>
> Une sœur est à Paris, une autre en Italie, une autre, malade, 40 ans, dans une région de Bretagne, mariée à un français en deuxièmes noces qui travaille à EDF. Elle a un enfant d'un premier mariage.
>
> Les autres sont à Dakar : ((anonym : prénom du frère 1)) a une entreprise de décoration, ((anonym : prénom du frère 2)) a un restaurant (où vont les gens du Paris-Dakar), et ((anonym : prénom du frère 3)) son jumeau a aussi un restaurant.
>
> Ils ont très bien réussi.

Nous suivons dans ce cas une règle stricte : nous ne transformons jamais l'information – par exemple nous ne modifions pas les noms ou les prénoms des enquêtés par des pseudonymes[32].

Par principe, nous conservons l'intégrité de l'information ; un dévoiement de l'information originale pourrait en effet induire en erreur les utilisateurs et conduire à de mauvaises interprétations. En ce qui concerne les données brutes, la technique de l'hyperonyme est plus satisfaisante que la pseudonymisation car bien qu'elle

32 Dans l'enquête De l'Afrique à la France ce sont les auteurs qui ont choisi les pseudonymes pour les besoins de la publication, sauf exception pour deux enquêtés annexes où il nous a fallu anonymiser avec des hyperonymes (cf. exemple précédent).

implique de soustraire l'information elle n'implique pas de la falsifier[33], puisque l'indication de la catégorie d'information dont il était question est conservée, ce qui permet de guider l'utilisateur sur la nature de l'information masquée.

Ensuite, ce travail doit être réalisé non seulement sur les données contenues dans les « matériaux » (typiquement les transcriptions d'entretiens ou les notes d'observation) qui concentrent la plus grande partie des informations sur les enquêtés, mais aussi sur d'autres documents constitutifs des corpus d'enquêtes. Ces derniers peuvent contenir des informations risquant de révéler l'identité des enquêtés. Cela est le cas par exemple de certains documents préparatoires (mentions de prises de contact avec des groupes, des institutions, etc.) ou de documents d'analyse (lorsque, par exemple, des cas sont décrits en reprenant quelques informations plus ou moins transformées des données, des extraits de verbatims, etc.)[34].

Les métadonnées, au sens large, peuvent également contenir des informations pouvant favoriser l'identification des enquêtés, et il faut être particulièrement vigilant à cet égard, en vérifiant non seulement les métadonnées originelles contenues dans les propriétés des fichiers mais aussi, bien sûr, les en-têtes des documents (notamment dans le cas des transcriptions), ainsi que les métadonnées descriptives que nous rajoutons après coup dans le processus de documentation des données.

Cela a déjà été évoqué, nous supprimons également les passages ou informations sensibles susceptibles de porter préjudice aux personnes citées dans les enquêtes[35], mais aussi aux chercheurs eux-mêmes – en effet des informations intimes sur ces derniers peuvent avoir été consignées et conservées dans les documents de l'enquête (notes, journal de bord, etc.). Dans ce cas, nous évaluons au cas par cas quelles informations doivent être supprimées, en collaboration avec les chercheurs producteurs.

Pour compenser le retrait d'informations nous produisons des outils pour préserver la compréhension du contexte, et donc le potentiel de réutilisation des données. Ils peuvent prendre la forme de tableaux de correspondance entre les listes d'enquêtés et les pseudonymes éventuellement utilisés par le chercheur dans les publications, comme c'est le cas dans la restitution de l'enquête Choisir son école[36].

Plus généralement, nous explicitons systématiquement dans un rapport réalisé par beQuali (« l'enquête sur l'enquête ») les stratégies d'anonymisation adoptées pour chaque enquête, à destination des utilisateurs afin qu'ils puissent les réutiliser en connaissance de cause.

Dans l'impossibilité de tout anonymiser, nous prenons des précautions supplémentaires qui renvoient au respect de la confidentialité. L'accès aux enquêtes est sécurisé

---

33 Ce faisant, c'est aussi une garantie du respect de l'intégrité des documents.

34 Une vigilance élevée est nécessaire pour détecter des risques d'identification qui se logent parfois dans les détails. Par exemple, dans l'enquête Dilapidation et prodigalité, une note manuscrite d'Anne Gotman mentionnait incidemment, en marge, un arrêt de bus et le numéro de la ligne, indication – dès lors supprimée – qui renseignait indirectement sur le nom de la ville du tribunal d'instance où la chercheuse interrogeait des juges des tutelles et consultait des dossiers de prodigues.

35 Voir *supra* l'enquête Comparaison des ministères de l'Enseignement supérieur en France et en Allemagne.

36 Voir le document cdsp_bequali_s1_add_transcr_liste_FR_pseudonymes, accessible sur demande à l'adresse : http://bequali.fr/fr/les-enquetes/demander-lacces/cdsp_bq_s5/.

et restreint à la communauté scientifique (enseignants, chercheurs, ingénieurs et étudiants) qui peut consulter les données sur justification d'un projet de recherche ou d'enseignement et après signature d'une convention de réutilisation. Les conditions de réutilisations y sont encadrées : les utilisateurs s'engagent à ne pas dupliquer les fichiers, à ne pas les transmettre à des tiers et à ne pas chercher à briser l'anonymat.

> Extraits d'un contrat de réutilisation concernant l'enquête Les Français et la politique (Étienne Schweisguth)
>
> _____ s'engage à utiliser à des fins de recherche ou d'enseignement des méthodes les données auxquelles beQuali lui a donné accès, à ne pas les transmettre à des tiers, à les utiliser loyalement (sans chercher à nuire notamment à la réputation du chercheur primaire), et à respecter l'anonymat des enquêtés.
>
> _____ s'engage également à mentionner, pour toutes les publications ou autres formes de communication des résultats qui en résulteront, la source des données selon la norme suivante : [Les Français et la politique, 1982-1988, Étienne Schweisguth, CEVIPOF (producteur), beQuali (diffuseur)], ainsi que la publication suivante : Étienne SCHWEISGUTH, « Les avatars de la dimension gauche-droite », in Élisabeth Dupoirier et Gérard Grunberg (dir.), *Mars 1986 : la drôle de défaite de la gauche*, Paris, Presses universitaires de France, 1986, p. 51-70.

Les documents originaux, une fois traités, sont soit reversés au producteur de l'enquête, soit versés auprès du service d'archive compétent. De ce fait, il sera toujours possible à un utilisateur potentiel de faire une demande de consultation des originaux en dehors des documents anonymisés mis à disposition par beQuali. L'information sur les données n'est donc jamais perdue : son degré de précision est simplement suspendu pour les besoins du partage numérique, pour un temps donné.

## Conclusion

En conclusion, nous voudrions insister sur le fait que l'anonymisation, telle que nous y sommes confrontés dans le cadre de beQuali, est un mélange d'opérations et de compétences métier, un compromis entre des contraintes et des logiques scientifiques, juridiques, archivistiques et techniques basé sur une « bonne gestion du risque ». Dans la pratique, il est impossible de produire des règles systématiques, dont les préceptes seraient applicables pour toutes les situations. La réalité résiste à de telles règles supposément générales et l'anonymisation, en tant que prescription, se matérialise par des arrangements au cas par cas. L'ensemble des éléments que nous venons d'exposer résulte d'une réflexion fondée sur notre expérience sur des problèmes spécifiques posés par chaque enquête, mais aussi sur des échanges collectifs avec d'autres acteurs concernés par cette question[37], dont les chercheurs

37 Discussions avec des producteurs d'enquêtes lors des visites que nous effectuons dans les laboratoires, consultation du CIL du CNRS, réflexions menées avec des experts au sein d'un groupe de travail organisé par beQuali sur l'anonymisation, etc.

producteurs, que nous consultons systématiquement pour élaborer les protocoles d'anonymisation et dont la collaboration est nécessaire, même si son degré peut varier d'une enquête à une autre. Nous sommes encore loin d'avoir réglé tous les cas problématiques (celui par exemple des enquêtes ethnographiques menées sur des terrains ou des groupes très localisés, ou encore des enregistrements audio ou vidéo). Si l'anonymisation est une pratique généralisée dans le travail de recherche, étape obligée pour restituer publiquement les résultats des enquêtes, les quelques réflexions visibles dans la littérature des sciences sociales abordent cette question essentiellement sous l'angle des publications, laissant souvent de côté les aspects archivistiques. Dans un contexte où les enjeux de la capitalisation des données d'enquêtes ont atteint une actualité saillante, nous espérons, à travers ce témoignage, montrer en quoi les exigences de la réutilisation obligent à décaler quelque peu ce point de vue pour aborder sous un autre angle les enjeux de l'anonymisation. En effet, se poser la question de la manière dont on anonymise des données d'enquête à des fins de réutilisation, dans le cadre d'une mutualisation entre membres de la communauté scientifique qui n'entretiennent pas de relations de collaboration a priori, oblige à dépasser la situation où la question est réglée individuellement, par chaque chercheur selon sa propre éthique ou sa perception de la déontologie d'enquête. Cette situation aurait le mérite de rendre davantage publiques et partageables des pratiques qui intéressent aussi bien les enquêtés que, collectivement, les (futurs) chercheurs et enseignants.

# Table des matières

# La diffusion numérique des données en SHS

## Guide des bonnes pratiques éthiques et juridiques

DIGITALES

La collection « Digitales » s'intéresse aux rapports entre les sciences humaines et le monde numérique, qu'il fournisse des outils critiques ou qu'il soit un domaine de création.

Produire, exploiter, éditer, publier ou valoriser des données numériques fait partie du travail quotidien des chercheurs en sciences humaines et sociales (SHS). Ces données sont aujourd'hui disséminées sous de multiples formats dans le monde de la recherche et, au-delà, auprès de citoyens de plus en plus curieux et intéressés par les documents produits par les scientifiques. Dans un contexte de mutation fulgurante des méthodes de travail, ce guide aborde avec simplicité des questions et des enjeux complexes auxquels se confronte quotidiennement la communauté des SHS. De leur collecte à leur réutilisation, les données de la recherche sont manipulées, éditorialisées, interrogées, mises en ligne… par tous les acteurs du monde académique qui ne savent pas toujours répondre aux questions juridiques et éthiques ou même, ne parviennent pas à les poser clairement. C'est à eux que s'adresse cet ouvrage, fondé sur des réflexions et des retours d'expériences qui présentent les bonnes pratiques pour accompagner celles et ceux qui s'inscrivent dans la dynamique de la science ouverte.

conception graphique
et illustration de couverture
J.-B. Cholbi

**Véronique Ginouvès** est responsable des archives sonores et audiovisuelles à la Maison méditerranéenne des sciences de l'homme (AMU-CNRS) à Aix-en-Provence.

**Isabelle Gras** est conservatrice des bibliothèques au Service commun de la documentation de l'université d'Aix-Marseille (SCD AMU).

Presses Universitaires de Provence
Aix*Marseille université initiative d'excellence
Bibliothèques universitaires
cnrs
Maison méditerranéenne des sciences de l'homme USR 3125
Huma-Num

9 791032 001790

20 €