

ACCOMPAGNER ÉTUDIANTS ET CHERCHEURS DANS L'EXPLOITATION DES SOURCES STATISTIQUES

L'EXPÉRIENCE DE LA PLATEFORME UNIVERSITAIRE DE DONNÉES DE LILLE (PUDL)

Jean-Marie DUPREZ¹ et Marie CROS²

TITLE

Supporting students and researchers in using statistical sources: the experience of the University Platform of Data in Lille (PUDL)

RÉSUMÉ

Utiliser les ressources statistiques particulières que sont les grands fichiers d'enquête suppose de mobiliser des compétences dans trois domaines à la fois connexes et distincts : statistique, informatique, production et mise en forme des données. A partir de l'expérience de la PUDL, on analyse les difficultés rencontrées par les utilisateurs dans la mise en œuvre de cette triple compétence. On insiste en particulier sur la nécessité de développer la compétence « données », qui se révèle un préalable indispensable pour tous ceux qui souhaitent s'engager dans cette démarche d'exploration des données statistiques.

Mots-clés : fichiers, enquête, compétences, données, Quetelet.

ABSTRACT

Using such specific statistical resources as large survey files demands skills in three fields which are both closely-related and distinct: statistics, computer science, data production and formatting. Drawing on the experience of PUDL, we analyze the users' difficulties in implementing this triple competence. We particularly insist on the necessity to develop "data competency", which is an essential prerequisite for those who wish to engage in this exploration of statistical data.

Keywords : files, survey, skills, data, Quetelet.

La plateforme universitaire de données de Lille³ (PUDL) a pour objectif d'accompagner les enseignants et les étudiants des universités lilloises dans l'exploitation des sources statistiques nationales et internationales, en particulier celles constituées par les grands fichiers d'enquête.

Elle fait partie du dispositif Quetelet et se situe au contact direct avec les utilisateurs, au moment même où ceux-ci recherchent et utilisent des sources statistiques diverses à des fins de recherche ou d'enseignement. La PUDL intervient très concrètement dans la recherche des sources et de leur documentation, l'exploitation des fichiers d'enquête, l'analyse des résultats. Dans ce sens, elle est un point d'observation privilégiée pour voir comment se réalise l'application de connaissances générales et statistiques en particulier dans un processus de recherche universitaire.

¹ Université de Lille 1, Fac de SES, jean-marie.duprez@univ-lille1.fr

² Université de Lille 1, Fac de SES, marie.cros@univ-lille1.fr

³ Voir la note de présentation en annexe.

L'objet de cet article ne portera pas sur l'apprentissage proprement dit des notions statistiques mais sur les difficultés rencontrées au moment de la recherche de sources statistiques et de l'exploitation de grands fichiers de données. En effet, les démarches qu'exigent cette recherche et cette exploitation font aussi, à notre point de vue, partie de la formation de statisticien en sciences économiques et sociales.

Précisons d'emblée, pour éviter toute ambiguïté, que :

- le public concerné est exclusivement celui des chercheurs et étudiants en sciences humaines et sociales, économiques et politiques. D'une façon générale, et si on excepte les économistes, l'essentiel des connaissances en statistique de cette population reste élémentaire ;
- les données concernées ne sont pas les résultats statistiques déjà constitués (il existe pour cela l'INSEE, le portail de la statistique publique française, l'OCDE, etc.), mais des données sources nécessitant d'être retravaillées, ou plus précisément encore des fichiers de données sur lesquels une analyse secondaire peut être réalisée.

1 Introduction

Comment analyser les difficultés rencontrées à l'occasion de la mise en œuvre des grands fichiers d'enquête ? On se propose de les examiner à partir des principales dimensions qui structurent la démarche d'utilisation de fichiers d'enquête.

Exploiter les données contenues dans les fichiers d'enquête requiert la mise en œuvre de compétences particulières (dimension compétence), au cours de différentes étapes d'un travail de recherche (dimension processus), sur la base d'un niveau d'expertise ou d'expérience plus ou moins élevé (dimension expertise).

Les compétences mobilisées relèvent de trois domaines à la fois connexes et distincts : celui des données proprement dites (la connaissance des sources et de l'organisation des données), le domaine informatique (connaissance d'un système d'exploitation et des logiciels statistiques), le domaine statistique (connaissance des traitements statistiques réalisables sur des données d'enquête). Il est illusoire en effet de penser que l'on puisse entreprendre un travail tant soit peu approfondi sur les fichiers d'enquête sans conjuguer ensemble ces trois types de compétences.

Les étapes du travail (la dimension processus) peuvent être déclinées en : 1) recherche et sélection des données pertinentes ; 2) exploitation des données ; 3) interprétation des résultats. Ici encore, on verra que les difficultés peuvent différer de façon importante selon la phase du travail où l'on se trouve.

Enfin, et de façon assez évidente, les difficultés ne sont pas les mêmes selon le niveau d'expérience (c'est la dimension expertise qui peut être très variable dans les trois domaines distingués). De ce point de vue, la PUDL accueille trois publics différents que, dans une typologie sommaire, on pourrait classer en trois catégories :

1. les experts : ils connaissent, parce qu'ils se sont déjà confrontés à ce type de travail, les données disponibles et les moyens informatiques et statistiques pour les exploiter. C'est un public autonome qui, sauf pour l'étape de recherche de données, n'a pas besoin de la PUDL ;

J.-M. Duprez, et M. Cros

2. les initiés partiels : ils disposent d'une ou plusieurs compétences dans les domaines distingués ci-dessus mais pas la totalité, ou ils maîtrisent une étape mais pas l'ensemble du processus de travail. C'est un public partiellement autonome qui recherche dans la PUDL les moyens de combler des lacunes qui se révèlent au fil du travail ;
3. les novices : ils n'ont jamais réalisé une exploitation de fichiers d'enquête de façon autonome. Il s'agit majoritairement d'étudiants de master ou de doctorat, mais aussi de chercheurs confirmés qui n'ont pas l'habitude ou eu l'occasion de traiter des données quantitatives. C'est un public de débutants, par définition peu autonome, qui demande un accompagnement aux différentes étapes de la recherche et maîtrise encore mal les trois compétences distinguées.

La PUDL accueille ces trois types de public dans des proportions différentes. Les experts, peu nombreux, viennent d'abord chercher de l'information sur les données disponibles et les moyens de les acquérir. Leurs demandes sont rarement « statistiques » ou « informatiques ». Il s'agit davantage d'utiliser la plateforme pour réaliser une documentation particulière (souvent sur des enquêtes anciennes pour lesquelles elle n'est pas constituée), chercher des ressources internationales nécessitant un travail plus ou moins long d'investigation, assurer le suivi administratif de la demande de fichiers d'enquête et la récupération matérielle des données.

A l'inverse, les novices et les initiés partiels sont les plus nombreux et leurs demandes diversifiées. C'est à partir des contacts noués avec ces deux catégories d'utilisateurs que l'on se propose d'exposer les difficultés rencontrées et de faire quelques propositions d'action. Les trois compétences que nous avons distinguées serviront ici de guide pour l'exposé.

2 Connaître et accéder aux données : une compétence à développer

La recherche de données d'enquête, sous la forme de fichiers de données, constitue l'étape initiale et souvent cruciale pour le reste de la recherche. Sans données, pas de travail statistique envisageable. Sans données pertinentes, pas de réponse aux questions posées.

Quelles sont alors ici les difficultés rencontrées ? Pour comprendre la situation, on peut la comparer, pour mieux la distinguer, à la constitution d'une bibliographie. Dans ce cas, la liste des publications et des ouvrages est (relativement) définie et les moteurs de recherche (par ex. INIST, Jstore, Web) permettent de retrouver les publications sur un thème donné. Il faudra ensuite trier, sélectionner, exploiter.

A l'inverse de la recherche bibliographique, la recherche en matière de fichiers d'enquête se heurte à deux difficultés importantes : l'inventaire des ressources disponibles est très partiel alors que l'on assiste à un accroissement notable des ressources.

On observe d'abord que, sans pouvoir répondre évidemment à tous les besoins, ce ne sont pas les ressources qui manquent. En France, en 2004, on dénombre 680 enquêtes inscrites au programme d'enquêtes statistiques des services publics français tel qu'il a été publié au Journal Officiel (voir Poulain *et al.*, 2006). Ce qui, en tenant compte de la diversité de la périodicité de la collecte (elle peut être mensuelle, trimestrielle, etc.), aboutit cette année-là à 2500 collectes effectivement lancées sur le terrain. Et l'on se limite ici à la statistique

publique. Au niveau international, l'ICPSR (Inter-University Consortium for Political and Social Research) a archivé plus de 5000 enquêtes dans le champ des sciences politiques, de la sociologie et de l'économie, accessibles pour la recherche et l'enseignement.

Encore faut-il en avoir connaissance et y retrouver l'information pertinente. Or, si on prend le cas de la France, il n'existe pas de répertoire qui permette de savoir quelles enquêtes statistiques sont en cours ou ont été menées au cours des années précédentes. Si les grandes enquêtes de l'INSEE et de l'INED sont connues, les fichiers divers constitués à l'occasion de travaux menés dans les ministères, les organismes de recherche et, a fortiori, les laboratoires universitaires ne sont pas répertoriés. Certes, on peut par le biais du CNIS en retrouver l'essentiel mais pas tous et encore moins ceux menés par les organismes privés (IFOP, SOFRES, etc.).

Diverses tentatives ont tenté de remédier à cette difficulté : dans les années 1980, l'INSEE a publié quelques volumes d'une publication appelée « la clé de sources » dont l'ambition était de répertorier toutes les sources nationales produites par un organisme public, mais la publication en est interrompue dès 1991. Claude Poulain de l'INSEE Auvergne a repris ce travail en l'étendant aux sources régionales pour la période 1995-2004, mais là encore ce travail n'a pas été prolongé ni même diffusé. Pour sa part, le Réseau Quetelet offre un moteur de recherche adapté aux fichiers d'enquête, mais il se limite aux enquêtes accessibles par le biais du réseau et ne peut prétendre couvrir l'ensemble de la production française. Quant au niveau européen, si le CESSDA (Council of European Social Science Data Archives) cherche à améliorer l'accès aux données européennes pour les chercheurs et les étudiants, il ne peut offrir mieux que ce qu'offre chaque pays de l'Union.

Le chercheur se trouve donc dans une situation paradoxale. D'une part, il existe de nombreuses enquêtes menées dans des secteurs très variés ; d'autre part, leur connaissance – ne parlons pas de leur accès – reste difficile et incomplète. Face à ce type de situation, la PUDL s'appuie sur les ressources du Réseau Quetelet et sur l'ensemble des moyens d'information existant pour répondre aux demandes de ses utilisateurs. Elle tente surtout de dépasser deux limitations récurrentes des moteurs de recherche : d'une part, la difficulté de fouiller les enquêtes pour aboutir aux questions elles-mêmes et aux réponses proposées ; d'autre part, l'absence de documentation sur les enquêtes les plus anciennes.

Mais le suivi des demandes faites à la plateforme en matière de recherche de sources statistiques permet de mettre l'accent sur une autre faiblesse propre cette fois aux utilisateurs et qui relève de ce qu'on pourrait appeler leur compétence « données ». On observe en particulier qu'une bonne partie de la population des novices méconnaît le plus souvent d'une part l'ensemble des producteurs et des diffuseurs de données, d'autre part les conditions de production des données (dont les conditions de réalisation du sondage, les instructions aux enquêteurs, etc.) et, enfin, l'organisation matérielle des fichiers et leurs principes de codification, autant d'éléments essentiels à la bonne compréhension et donc à la bonne utilisation des fichiers de données. C'est donc ici directement la compétence « données » des utilisateurs qui fait défaut et qui leur fait apparaître les fichiers d'enquête comme autant d'ovnis statistiques. Cette méconnaissance n'est pas en soi étonnante puisque la compétence « données » est encore peu développée dans l'enseignement universitaire et ne fait que peu ou pas partie de l'enseignement traditionnel des « statistiques ». Mais il faut bien convenir que si on ne sait pas ce que l'on cherche et sous quelle forme, on aura effectivement du mal à le trouver.

J.-M. Duprez, et M. Cros

Dans une large mesure, la PUDL offre à ses utilisateurs cette compétence « données », d'abord par la connaissance des producteurs et de leur production, ensuite en leur offrant les moyens d'accéder aux ressources existantes. Mais, plus largement, elle plaide pour un double développement de cette forme de compétence. D'abord dans l'enseignement universitaire où un enseignement sur les sources statistiques et leurs conditions de production et d'utilisation devrait être largement proposé⁴ aux étudiants en sciences sociales, économiques et politiques. Ensuite, elle suggère qu'à l'exemple des documentalistes et bibliothécaires, le métier de spécialiste en ressources statistiques soit développé au sein des universités et des organismes de recherche.

3 Exploitation des données et compétence informatique

Une fois les données repérées, encore faut-il les récupérer et les exploiter. Dans cette étape de la recherche, c'est surtout la compétence informatique qui va être mobilisée. On examine sa mise en œuvre pour souligner une nouvelle série de difficultés.

La récupération matérielle des données est une première occasion de mobiliser la compétence informatique. Si l'envoi par voie postale d'un CD peut être utilisé, c'est généralement par le biais du WEB et de ses différents outils que les données seront récupérées sous forme de pièce jointe à un courriel, par téléchargement sur un site web, par logiciel de type FTP. Avec le développement de l'usage des navigateurs internet et du courrier électronique, les problèmes rencontrés ont nettement diminué et c'est surtout dans l'usage d'un logiciel de type FTP que des demandes d'aide sont formulées. D'une façon générale, la récupération des données auprès du producteur ou du diffuseur a été très notablement améliorée ces dernières années, du moins en ce qui concerne les données du réseau Quetelet. La mise en place du nouveau portail Quetelet au deuxième semestre 2009 devrait encore améliorer la situation.

La mise en forme des données dans un ou des fichiers de travail constitue une étape informatique plus délicate. En effet, récupérer des données ne veut pas dire qu'on en dispose sous une forme exploitable par son logiciel favori. Il y a souvent ici deux moments qui posent problème : a) la conversion des données dans le format du logiciel statistique qui sera utilisé (c'est, par exemple, passer d'un fichier transport de SAS à un jeu de données qui soit exploitable) ; b) la constitution d'un fichier de travail adapté au problème de recherche : ainsi, il arrive fréquemment que les données se présentent sous la forme de plusieurs fichiers (exemple : les enquêtes FQP, Emploi, Patrimoine de l'INSEE) et il faut alors réaliser un fichier de travail par fusion des variables et observations pertinentes. Et lorsque les données sont dans un fichier unique, il est souvent utile de sélectionner les seules variables et observations utiles à la recherche pour ne pas manipuler un fichier trop lourd et complexe.

L'exploitation informatique du fichier suppose que les données soient suffisamment documentées. On observe ici une double difficulté. D'une part, malgré les efforts réalisés, la documentation est loin d'être complète et disponible pour tous les fichiers ou, dans les enquêtes les plus anciennes, il faut littéralement la reconstituer... D'autre part, et dans tous les cas, il faudra intégrer les libellés des variables et des modalités dans le logiciel utilisé. Ce qui peut paraître élémentaire peut se révéler un travail lourd et fastidieux.

⁴ Une formation de ce type a été mise en place dans le cadre du Master IESS (Ingénierie de l'enquête en sciences sociales) de l'Université de Lille depuis 2005.

Enfin, l'exploitation des données proprement dite suppose une bonne connaissance des possibilités offertes par le logiciel pour les opérations classiques de création de nouvelles variables, regroupement de modalités, production des tabulations et analyses statistiques.

L'expérience accumulée à la PUDL montre la grande diversité des difficultés concrètes rencontrées par les candidats à l'analyse des fichiers d'enquête selon leur formation et leur niveau d'expertise informatique. Si pour l'un la simple récupération des données semble poser des problèmes insurmontables, ce sera pour un autre l'intégration de la documentation au processus de travail ou encore la réalisation d'opérations informatiques, ou enfin la mise en œuvre de procédures statistiques sophistiquées. La réponse apportée est d'abord individuelle de façon à répondre à la fois à la demande précise de l'utilisateur et à son urgence, mais elle peut aussi être collective (atelier « logiciels » autour de SAS). Elle peut porter sur une aide directe à la récupération des données, l'installation des logiciels, la constitution et la fourniture d'une documentation de travail pour les enquêtes les plus demandées, la formation aux logiciels (essentiellement SAS, SPSS, Excel, plus rarement Tri2, Modalisa, Sphinx), l'aide à la résolution de problèmes dans la mise en œuvre des logiciels.

Les difficultés précédentes amènent à se demander s'il ne serait pas plus simple de passer à une exploitation des données directement sur le site du producteur (ou diffuseur) en utilisant des logiciels adaptés (Nestar, Beyond, etc.). Il existe déjà sur plusieurs sites (INSEE, OCDE, quelques sites internationaux) la possibilité de réaliser des tabulations à distance. L'intérêt de ce type d'offre est d'éviter une partie des manœuvres précédentes et de permettre au producteur de garder le contrôle sur ses données. Mais les limites sont vite rencontrées : en dehors de tabulations plus ou moins faciles à mettre en œuvre, il n'y a guère de possibilité de traitements statistiques approfondis. Or, c'est bien là l'intérêt de disposer des fichiers d'enquête : celui de fouiller les données et de leur appliquer la grande variété des traitements qu'offre aujourd'hui la statistique.

Comment alors développer la compétence « informatique » ? En sus de la maîtrise du système d'exploitation lié au poste de travail (c'est quand même un préalable indispensable), c'est dans un double sens que l'on suggère de développer les efforts :

- Généraliser dans l'enseignement l'approche des différents moments que nous avons distingués. Les contraintes du travail pédagogique amènent souvent à travailler sur un fichier nettoyé et limité. Ce qui tend à faire croire aux étudiants que les données sont propres et facilement accessibles. Lorsqu'on en a le temps, c'est l'ensemble de la chaîne qu'il faudrait faire expérimenter aux étudiants : récupération des données sur un serveur, constitution du fichier de travail et de la documentation utile au travail envisagé, création des variables manquantes, constitution des résultats. Ceci suppose évidemment qu'en parallèle soit effectuée une initiation plus ou moins approfondie à un logiciel spécialisé.
- Initiation minimale à l'interrogation à distance. Exemple : OCDE, INSEE.

4 Difficultés liées à la compétence « statistique »

Même si nous l'abordons dans cette communication en dernière position, la compétence « statistique » est présente aux trois étapes que nous avons distinguées dans l'exploitation des fichiers d'enquête. Elle décide du choix des outils statistiques, se manifeste dans l'analyse des résultats et leur rédaction. L'expérience montre que cette compétence est très variable d'un

J.-M. Duprez et M. Cros

utilisateur à l'autre, plutôt bien développée chez les économistes, souvent très limitée chez les utilisateurs en sciences humaines, sociales, politiques.

Le niveau d'expertise induit directement la plupart des opérations statistiques qui seront réalisées – tabulations et tests associés (chi² essentiellement), plus rarement AFC et classifications, chez les moins pourvus ; régressions linéaire ou logistique, modélisation chez les autres – et donc l'aide qui sera éventuellement demandée à la PUDL.

En soi, ces demandes dépassent le cadre d'action qui a été fixé à la PUDL et relèvent directement des directeurs de recherche ou de thèse. Mais il est bien difficile de ne pas prendre en considération dans l'analyse de la demande initiale comme dans l'accompagnement qui sera apporté en matière informatique, les opérations statistiques qui sont envisagées sur les données. L'expérience montre bien qu'il s'agit là de trois éléments enchevêtrés et qui doivent être considérés en même temps.

Très concrètement, il nous est donné d'observer trois grands types de difficultés. Les premières concernent les plus débutants et les moins bien pourvus pour qui la première expérience d'utilisation d'un fichier d'enquête représente une véritable expérimentation grandeur nature des connaissances engrangées pendant leur formation. On relève des erreurs élémentaires auxquelles les enseignants sont bien habitués : par exemple, une interprétation des moyennes sans tenir compte de l'écart-type, une lecture inappropriée des tabulations (en clair, on lit en colonne quand il faut lire en ligne, ou l'inverse), une interprétation abusive des différences de pourcentages observées (à partir de quand une différence a du sens), une extrapolation sur la base d'effectifs particulièrement limités (on travaille sur les pourcentages en oubliant les effectifs), une interprétation incorrecte ou magique des tests (le chi² est mis à toutes les sauces et devient le critère décisif de l'association entre variables), etc. Ceci montre clairement que, pour une série d'utilisateurs, les problèmes de la causalité (dans le cas de lecture incorrecte des résultats), de la significativité des données (dans la lecture des différences dans les pourcentages), de l'échantillonnage (dans le cas d'extrapolation abusive), de la compréhension des tests (dans le cas du chi²) sont encore mal résolus.

Le second type de difficultés fréquemment rencontrées concerne l'interprétation des résultats. Il s'agit là d'une phase délicate même pour les plus aguerris. Le chercheur a réalisé une AFC, une régression, une classification, une modélisation, se trouve face à des listings où les résultats sont présentés de façon diverse selon le logiciel utilisé et se demande comment les interpréter. C'est là un moment où le besoin d'échanger peut être très fort... mais pas toujours comblé.

Le troisième type de difficultés, largement partagées, concerne la rédaction des résultats. Comment exposer par écrit ou oral les résultats des analyses réalisées en leur donnant du sens ?

La PUDL n'a pas pour mission d'entrer dans les problématiques ni dans les conclusions des chercheurs. Mais elle est bien placée pour témoigner de quelques-unes de leurs difficultés. Elle ne peut y remédier que très partiellement soit en intervenant à la demande directe d'un utilisateur (tant que l'analyse reste à un niveau élémentaire), soit en sollicitant d'autres chercheurs dans un but de partage d'expérience. Plus largement, l'expérience acquise permet de formuler deux types de vœux.

Après des enseignants, on ne peut qu'insister sur la compréhension du codage et du tableau des données, la lecture des tableaux, la compréhension des tests statistiques et la rédaction des résultats. Aux chercheurs, une première suggestion serait de davantage

documenter les analyses réalisées pour qu'elles puissent servir à des fins pédagogiques. En incluant une description des données, de la programmation utilisée, des résultats et de leur interprétation. Travail lourd et peu rémunérateur mais qui serait ô combien utile ! Le but final serait de pouvoir constituer une sorte de catalogue d'applications décrites complètement, depuis le fichier initial jusqu'à l'interprétation.

Une seconde suggestion, déjà émise mais jamais réalisée, serait de réaliser un forum par grande enquête. Les chercheurs travaillant sur les mêmes données pourraient y mettre en commun une série d'informations et de conseils, de programmes et d'analyses. Utopie ?

5 Conclusion

Nous voudrions, en guise de conclusion, faire deux séries de suggestions plus générales tirées de l'observation des usages des fichiers d'enquête menée au sein de la PUDL.

Les premières concernent la formation des étudiants : l'exploitation des sources statistiques, en particulier des grands fichiers d'enquête, exige de notre point de vue de faire converger trois types de compétences qui interviennent chacune à leur manière aux différentes étapes du processus de recherche. Jusqu'à présent, sans doute parce que l'accès aux fichiers d'enquête était à la fois difficile et méconnu, la compétence « données » a été peu développée au cours de la formation à l'inverse des compétences informatique et statistique. Cependant, l'observation menée au sein de la PUDL montre qu'elle est aussi cruciale que les autres et qu'elle devrait être développée si l'on veut favoriser l'utilisation des ressources contenues dans les fichiers d'enquête.

Les secondes concernent le niveau organisationnel et institutionnel : la création de relais (type PUDL) entre producteurs, diffuseurs et utilisateurs apparaît indispensable pour accompagner les utilisateurs et les aider aux différentes étapes de leur analyse. Il est vain d'espérer que chaque utilisateur possède le niveau d'expertise suffisant dans les trois compétences à mobiliser aux différentes étapes de sa démarche de recherche. Autant que bibliothèques et centres de documentation, les PUD apparaissent comme un équipement nécessaire pour le développement de la recherche.

Références

- [1] Poulain, C., P. de Fouchier et M.-M. Grissonnanche-Tartière (2006), Dix ans d'enquêtes statistiques en France 1995-2004, *Courrier des statistiques*, n°117-119, 21-32.
- [2] Silberman, R. (1999), Les sciences sociales et leurs données, rapport pour le Ministère de l'éducation nationale, de la recherche et de la technologie, <http://www.education.gouv.fr/cid1925/les-sciences-sociales-et-leurs-donnees.html>

J.-M. Duprez, et M. Cros

Annexe : note sur la Plateforme Universitaire de Données de Lille (PUDL)

L'exploitation, sous forme d'analyse secondaire, des enquêtes menées par des organismes publics ou privés (INSEE, INED, Services statistiques ministériels ou SSM, etc.) a été une demande constante de la part des chercheurs en sciences économiques et sociales. L'accès aux fichiers d'enquête a été notablement facilité, à la suite du rapport Silberman sur « Les sciences sociales et leurs données », par la création par décret en 2001 d'un Comité de Concertation pour les Données en Sciences Humaines et Sociales (CCDSHS, <http://www.enseignementsup-recherche.gouv.fr/comite/shumaines.htm>) et la mise en place d'un dispositif spécifique de diffusion des données : le Centre Quetelet, devenu par la suite Réseau Quetelet, s'est ainsi vu attribué pour mission l'archivage des données, leur diffusion et leur valorisation.

La création de plateformes universitaires de données a été voulue pour servir d'intermédiaire entre les producteurs/diffuseurs de données et les utilisateurs, mais aussi pour les aider localement dans la compréhension et le traitement des données, et les former à leur utilisation.

La Plateforme Universitaire de Données de Lille (PUDL) a été créée en 2003 et a été pendant longtemps la seule plateforme de ce type en France⁵. Elle est ouverte aux chercheurs, enseignants-chercheurs, doctorants et étudiants de master des universités et laboratoires de la région Nord-Pas-de-Calais.

Son action s'organise autour de trois axes principaux :

- l'aide à la recherche et à l'acquisition de fichiers statistiques ;
- l'aide à la mise en œuvre de ces fichiers ;
- la formation à l'utilisation des données, à l'utilisation des logiciels statistiques.

La PUDL développe également des actions d'information : plaquettes de présentation, site web, lettres d'information indiquant les enquêtes disponibles, les séminaires méthodologiques, les colloques...

Depuis son ouverture au public en 2003, on peut chiffrer à 250 le nombre de personnes qui ont demandé, sous une forme ou une autre, un soutien de la PUDL, et on observe un nombre croissant de fichiers empruntés suite à un contact avec la PUDL sur des thèmes aussi divers que l'emploi, la formation, le logement, les revenus, la santé...

Site web de la PUDL : <http://www.meshs.fr/PUDL/>

⁵ En 2009, une seconde plateforme s'est mise en place à Lyon dans le cadre de l'Institut des Sciences de l'Homme (ISH). Elle est accessible à l'adresse suivante : http://www.ish-lyon.cnrs.fr/DATA_SHS/index_fr.php