PROVIDING ACCESS TO SURVEYS AT THE FRENCH INSTITUTE FOR DEMOGRAPHIC STUDIES (INED): ORGANISATION AND OFFER

Arianna CAPORALI[1] (*), Amandine MORISSET (*), Stéphane LEGLEYE (*)

(*) Institut national d'études démographiques (Ined), Service des enquêtes et sondages

This is a summary of a forthcoming article in *Population*.

Please quote the following reference:

Caporali A., Morisset A., Legleye S. (à paraitre 2015), « L'Ined et la mise à disposition des enquêtes quantitatives pour la recherche en sciences sociales », *Population*, 70(3).

---

[1] arianna.caporali@ined.fr

Résumé

Visant à permettre le partage de données entre producteurs initiaux et utilisateurs secondaires, la mise à disposition des enquêtes représente un atout pour les sciences sociales. En effet, le partage permet de justifier les coûts élevés des enquêtes, d'éviter la répétition d'études similaires, d'approfondir et de valider les travaux de recherche scientifique. Néanmoins, les chercheurs sont parfois mal informés sur les activités de mise à disposition des enquêtes.

La première partie de ce papier retrace le développement des institutions gérant l'accès aux données quantitatives pour la communauté scientifique internationale. Ces institutions naissent dans les années cinquante, afin d'encourager les comparaisons en sciences politiques. Les premières archives d'enquêtes organisent des réseaux internationaux et fixent des normes pour la documentation des données. En France, la mise à disposition commence plus tardivement que dans d'autres pays. À l'Ined elle s'est développée au début des années 2000. Le Service des enquêtes et sondages (SES) fut alors co-fondateur du Réseau Quetelet, qui gère aujourd'hui l'accès à la plupart des enquêtes en France.

La deuxième partie de ce papier décrit les activités de mise à disposition des enquêtes quantitatives du SES de l'Ined. Elle détaille les différentes étapes et les enjeux qui concernent la préparation des enquêtes à mettre à disposition. Elle présente par ailleurs les normes réglant la documentation et l'accès aux données, ainsi que les fonctionnalités de l'outil utilisé pour publier, rechercher, visualiser et analyser les données et les métadonnées en ligne. La mise à disposition des enquêtes devient de plus en plus importante avec la demande croissante d'accès aux données. L'expérience du SES suggère que la préparation des données en vue de leur mise à disposition nécessite beaucoup de temps. Pour optimiser ce processus, la collaboration des producteurs des données est importante.

Abstract

Allowing data sharing, providing access is of key importance for social sciences. However, researchers are sometimes poorly informed about the activities carried out to provide access to surveys. First, this paper describes the development of social science survey data archives at the international level, in France and at the French Institute for Demographic Studies (Ined). Second, it gives an account of the activities carried out at Ined Survey Department to provide access to quantitative surveys. It details the steps and the issues involved in the data preparation, as well as the standards used to document surveys metadata. It also presents the functionalities of the software used to publish, browse, visualize and analyze data and metadata online. Providing access to surveys is becoming more and more important. To optimize the process required for data and metadata preparation, it is important to collaborate with researchers.

Mots-clés: données d'enquêtes quantitatives; archives de données en sciences sociales; partage des données ; métadonnées; Data Documentation Initiative.

Introduction: the importance of sharing survey data

Permitting data sharing, the activities which allow providing access to surveys are important for social research. For example, they allow testing existing studies. Furthermore, data sharing implies acknowledgement of survey producers' work and justifies expensive costs of surveys (King 1995; ICPSR 2012; Silberman 1999).

The activities to provide access to surveys are carried out by data producers and survey data archives. The latter assist data producers in assuring the wider access possible to their data (ICPSR 2012). To this end, they review the coherence of datasets, create exhaustive metadata records, publish survey data files in online catalogues, review requests for data access, and provide assistance to users. However, data producers are sometimes unaware about the specific activities carried out by archivists to provide access to data.

This paper gives a brief historical account of data archives dealing with quantitative surveys for social research (thereafter refer to as survey data archives)[2] at the international level, in France, and at the French Institute for Demographic Studies (*Institut national d'études démographiques*, Ined). It then illustrates the current organization of the activities that provide access to quantitative surveys at Ined Survey Department[3].

1. Historical background and current context

1.1. Survey data archives at the international level

First social science survey data archives were created after the Second World War thanks to the initiatives of researchers in political science. The post-war geopolitical context encouraged international comparative studies which motivated the need for sharing survey data (Bisco 1966; Silberman 1999). The majority of first survey data archives originated in academic contexts where researchers could carry out their own surveys (Silberman 1999). This development was encouraged by conferences organized by UNESCO (United Nations Educational, Scientific and Cultural Organization) in the 1960s and in the 1970s. These conferences also played a role towards the development of international networks (Bisco 1966; Rokkan and Scheuch 1963; Rokkan 1966; Scheuch 2003). For example, the Consortium of European Social Science Data Archives (CESSDA) was created in 1976. These networks encouraged the definition of a common international standard for documenting metadata (Scheuch 2003; Silbermann 1999).

Today, CESSDA centralises the offer of European survey data archives in a unique online data catalogue (http://www.cessda.net/catalogue/). The inclusion in this catalogue requires the implementation of the standard for documenting social science survey data called Data Documentation Initiative (DDI) and of the software Nesstar (Network Social Science Tools and

---

[2] These archives deal with individual (micro) level data and are different from other types of research data archives, which include for instance electronic texts archives and historical and archeological data archives (Doorn and Tjalsma 2007).
[3] These surveys are produced for no profit by researchers with public funds with or without the collaboration of other public bodies. There are other types of data, such as qualitative surveys, administrative data, data produced for profit that are not dealt with in this article.

Resources). Launched in 2000, DDI allows organising surveys metadata down to the level of each variable[4] (Vardigan, Heus, and Thomas 2008). Nesstar helps prepare survey metadata according to DDI standards and provides an easy web interface to explore and analyze surveys online[5].

1.2. Survey data archives in France

First social science survey data archives in France were created in the 1980s, later than other countries. At the basis of this delay is a weak institutional support to the realization of university surveys, which favored the development of survey data archives in other countries. Furthermore, French legal framework put much emphasis on data protection (Silberman 1999).

Two main regulations protected individual data. First, the 1951 law no. 51-711 on the obligation, coordination and secrecy in statistics implied that individual data collected by official statistics (National Institute of Statistics and Economic Studies - INSEE, ministries, governmental agencies and administrations) could be shared only in anonymized or aggregated forms (Gaeremynck 2009; Silberman 2011). Second, the 1978 Data Protection Act (no. 78-17) established that personal data could be collected and processed only for specific purposes, with the notice of the National Commission for Informatics and Freedom (CNIL)[6] and within a limited time (Silberman 2011).

First survey data archives were created in the 1980s: the Bank of Socio-political Data (BDSP)[7], and the Laboratory of Secondary Analysis and Methods Applied to Sociology (LASMAS)[8]. These archives were established thanks to the initiative of single engineers rather than as a common institutional effort (Silberman 1999). Access to data was regulated through special agreements between the archives and research institutions. In this context, Silberman (1999), in a report commissioned by the minister of National Education, Research and Technology, called for a "real archival structure" (p.47) and for a reform of the French legal framework on access to survey data.

The Silberman report laid the foundations for the creation in 2001 of the *Centre Quetelet* (Decree no. 2001-139), then transformed into a *Réseau*, which today centralizes the access to social science surveys in France. It had three founding members: 1) BDSP, 2) LASMAS, and 3) Ined Survey Department. Following the recommendations by CESSDA, its members have adopted the standard

---

[4] DDI (http://www.ddialliance.org/) is based on XML schemas and consists of a group of items that enables the general description of empirical studies down to the level of each variable. It replaces paper codebooks with metadata on electronic human-readable format. There are two lines of developments (Hansen *et al.* 2011; Kramer *et al.* 2011): first, the DDI-Codebook (DDI-C or DDI 2), introduced in 2002, which is focused on the elements of a traditional social science codebook and suited for cross sectional studies; second, the DDI-Lifecycle (DDI-L or DDI 3), introduced in 2008, which is designed to document surveys across their entire life cycle and is particularly suited for longitudinal studies.

[5] Nesstar (http://www.nesstar.com/) is today owned by the Norwegian Social Science Data Services (NSD). This software permits the preparation of survey documentation in DDI-C format, without the archivists having to know the XML language. It also allows publishing data and metadata on internet. The online window allows accessing survey metadata, visualizing variables frequency tables and the related documentation, creating personalized crossed tabulations and graphs, and performing simple analyses (i.e., correlations and regressions). Survey documentations as well as results of analysis can be exported in various formats.

[6] In particular, a declaration to the CNIL is sufficient in case of personal data allowing for direct or indirect identification. Whereas, in case of sensitive data (i.e., ethnic origin, political and religious opinions, health, sex behaviors), it is necessary to ask for its authorization.

[7] The BDSP was established at the Institute of Political Studies in Grenoble and then integrated to the Centre of Computerization of Socio-Political Data (CIDSP). It was then taken over by the Center for Socio-Political Data (CDSP, http://cdsp.sciences-po.fr/) located in Paris at SciencePo.

[8] LASMAS was created in Paris and then transformed in the *Centre Maurice-Halbwachs* (CMH, more information is available here: http://www.cmh.ens.fr/greco/adisp.php). One of its missions was to provide access to data produced by official statistical offices (INSEE, ministries, governmental agencies and administrations).

DDI to document metadata surveys and the software Nesstar, to publish data and metadata online. They manage data access requests through a common website (http://www.reseau-quetelet.cnrs.fr/spip/) which catalogues all the available surveys. Access is granted free of charges and only for research purposes (data access conditions are detailed online).

Furthermore, in 2004 a reform of the Data Protection Act (law no. 2004-801), introduced the possibility to reuse for statistical, historical and research purposes, personal data collected for other purposes (Silberman 2011). In 2008, the new archives law (no. 2008-696) introduced the possibility to use very detailed data[9] collected by official statistics for statistical, historical and research purposes (Gaeremynck 2009). The access to these data is submitted to the notice of the Committee on Statistical Confidentiality and managed through the Centre for Remote Secure Access to Data (CASD)[10] created in 2010 (Le Gléau and Royer 2011). The CASD is a member of the *Réseau Quetelet*.

1.3. The organization of survey access procedures at Ined

Since its creation in 1945, Ined has conducted surveys often in collaboration with other public institutions, dealing with socio-economic and demographic topics such as family, employment, fertility, sexuality, migrations, health, life conditions, etc. By the 1970s, Ined has focused on its obligation to archive data, in collaboration with the National Archives (*Comité d'archivage de l'Ined* January 2001). Access to Ined surveys was based on an implicit consensus and any potential problem was solved in an informal way. In the 1990s, the increasing requests for using Ined data in secondary analyses generated the need to organize formally the data access procedures. This need was expressed in the strategic counseling document for the years 2002-2005 (*Orientations stratégiques de l'Ined* 2002-2005). It was necessary to preserve priority access for Ined researchers and partners participating in the funding and design of the surveys. Ined could provide access only to surveys conducted by Ined in whole or in partnership with other institutions. The Survey Department, co-founder of the Réseau Quetelet, was given the task to enforce the Ined policy about providing access to its surveys.

The implementation of this policy started to be effective in 2004 (*Archives de l'équipe mise à disposition de l'Ined* 2004). Since 2006, on the impulse of François Héran and Cécile Lefèvre, head of the Survey Department, greater resources were put on the activities to provide access to surveys (*Comité d'archivage de l'Ined* 2006; *Orientations stratégiques de l'Ined* 2006-2009). As a result, some recent surveys began to be made available before being sent to the National Archives. In 2009 the Survey Department was given the task of providing access to surveys of the European research infrastructure "Generations and Gender Programme" (GGP, www.ggp-i.org/). As recommended by CESSDA and the *Réseau Quetelet*, the entire catalogue of Ined surveys was organized through the Nesstar software and documented according to the DDI standard.

---

[9] These data concern variables that if crossed may allow for identification of interviewees (e.g., disaggregated geographic locations or nationalities).
[10] The data are stored on a server, cannot be copied, and are accessible for a limited period of time through a personal code. The contract signed with CASD commits applicants to anonymize their analyses results.

2. Ined survey catalogue

2.1. Content and data access procedures

Ined survey catalogue was officially launched in June 2012. It replaces the dissemination tool that was previously used and it is available at: http://nesstar.ined.fr/webview/. Since its opening, it was visited by more than 5800 users. As of November 2014, Ined survey catalogue contains about 250 references. In the catalogue, surveys are divided into two categories (each organized by decades):

1) surveys that are accessible, i.e., for which it is possible to order the data file(s);

2) surveys that are not accessible, i.e., for which it is not possible to order the data file(s) (either because it does not exist, or because the access to it is not given yet). These surveys are described through a fact sheet.

For the accessible surveys, the catalogue offers a complete and downloadable documentation down to the variable level, as well as the possibility to perform and export some basic analyses. Access to the data can be asked through the portal of *Réseau Quetelet* and follows its procedures.

2.2. The work of data preparation and data documentation

In order to provide access to a survey, a meticulous and 'invisible' work of preparation and documentation is carried out. It is necessary to gather all the documentation about the survey itself and its data file(s). This documentation is often scattered: some documents can be stored in electronic files, some others in boxes of Ined archives, some other documents consist of articles published thanks to survey data. There can be a lack of information or some information can be duplicated. Some other times it is necessary to select the correct information. As to the data files (if available), two scenarios are possible. For the more recent surveys, files are in "current" formats, such as SAS, SPSS or Stata that don't require to be transformed. In case of old surveys, it is not rare to have only paper documents and data files in text format (or no data files at all). Data files are sometimes incomplete. For example, variable labels can be missing. In case of non-anonymous files, additional work is required to detect and remove potentially identifying variables.

Once all the data files and related documentation are gathered, they are uploaded in Nesstar. Metadata are documented according to specific DDI fields and classified in three groups:

1) Document description: information on the Nesstar file (survey concerned, author, etc.);

2) Study description: information on the survey (abstract, keywords, researchers involved, producers, funding institutions, dates of the data collection, data collection methodology, sampling procedure, weighting, description of questionnaire, response rates, etc.), links to the questionnaire(s), interviewers' instructions, associated surveys, and the related bibliography ;

3) Data files description: information about the dataset(s), such as how they are constructed, missing data, which variables are replaced with derived ones, etc.

Additionally, each variable is examined in great detail. Variables labels and modalities are informed, as well as the question texts (text before the question, literal question, text after the question, instructions to interviewers), the universe (i.e., the persons to whom the question was asked), the questionnaire from which the variable comes from (when the survey involves several

questionnaires) and, in the case of derivate variables, the variables and the scripts used in the calculation. To provide all this documentation, each dataset has to be entirely reviewed, and variables are reorganized so as to follow questionnaires' sections.

This work of documentation requires a great amount of time that can be reduced by the research teams who conducted the surveys. Exhaustive, clean and anonymous data files, good codebook and extensive documentation on the study methodology can accelerate the work required to provide access to surveys.

Conclusion

The need to organise the access to survey data started in the 1950s within the field of political science, to develop comparative research. International networks fostered the creation of standards for documenting metadata. In France, due to a weak academic tradition in large scale surveys and a legal framework protective towards individual data, this development is more recent. The *Réseau Quetelet* was created in 2001 as a structured organisation of until then scattered survey data archives. Today, it centralises the access to most French social science surveys. The development of the activities to provide access to surveys at Ined followed the same path. It was not before 2001 that the Institute began these activities.

Ined survey catalogue covers a large spectrum of socio-demographic topics and offers the possibility to browse and visualise data for accessible surveys. Survey metadata are provided according to the DDI standard and are downloadable for each survey. Because survey metadata are often not structured to the DDI standard, the work of data preparation requires a lot of time. A greater collaboration from data producers in preparing surveys and their metadata according to DDI standards could optimise the process. The trend toward open data and data deluge in social science (Silberman 2013) is increasing the need for providing rapid and high quality access to social science surveys. It is likely that greater and greater resources will be devoted to providing access to surveys.

References

[1] Arduin, P. (2004), Présentation du Centre Quetelet : description des missions, activités, partenaires et projets de mise à disposition des données, Presentation to the *Journée d'information sur l'archivage historique et la mise à disposition des données d'enquêtes en SHS*, Institut national d'études démographiques (Ined), January 29.

[2] Archives de l'équipe mise à disposition (2004), Internal documents of Ined Survey Department.

[3] Bisco, R. L. (1966), Social Science Data Archives: A Review of Developments, *The American Political Science Review*, 60 (1), 93−109.

[4] Comité d'archivage de l'Ined (2001-2009), Comptes-rendus et relevés de décisions, Ined internal documents.

[5] Doorn, P. and Tjalsma, H. (2007), Introduction: Archiving research data, *Archival science*, 7, 1−20.

[6] Gaeremynck J. (2009), Le comité du secret statistique, *Courrier des statistiques*, 128(3), 15-18.

[7] Hansen, S.H., Iverson, J., Jensen, U., Orten H. and Vompras J. (2011), Enabling Longitudinal Data Comparison Using DDI, *DDI Working Paper Series – Longitudinal Best Practice*, n. 2, http://www.ddialliance.org/resources/publications/working/BestPractices/LongitudinalData.

[8] Inter-university Consortium for Political and Social Research (ICPSR) (2012), *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (5th ed.), Ann Arbor, MI, http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf.

[9] King, G. (1995), Replication, Replication. *PS: Political Science & Politics*, 28(3), 444−452.

[10] Kramer, S., Banks, R., Chang, V., Sieber I., Vardigan M. and Zenk-Möltgen, W. (2011), Presenting Longitudinal Studies to end Users effectively Using DDI Metadata. *DDI Working Paper Series – Longitudinal Best Practice*, n. 4,

http://www.ddialliance.org/resources/publications/working/BestPractices/LongitudinalData.

[11] Le Gléau J. and Royer J. (2011), Le centre d'accès sécurisé aux données de la statistique publique française : un nouvel outil pour les chercheurs, *Courrier des statistiques*, 130(1), 1-5.

[12] Orientations stratégiques de l'Institut national d'études démographiques (2002-2005) and (2006-2009), Ined internal reports.

[13] Rokkan, S. and Scheuch, E. K. (1963), Conference on Data Archives in the Social Sciences, *Social Science Information*, 2(4), 109.

[14] Rokkan, S. (1966), *Data Archives for the Social Sciences*, Mouton Co. and Ecole Pratique des Hautes Etudes, Paris – The Haugue.

[15] Scheuch, E. K. (2003), Les services de données en sciences sociales : historique et perspectives. *Revue internationale des sciences sociales*, 177(3), 433−449*.

[16] Silberman, R. (1999), Les sciences sociales et leurs données, Report for the Ministère de l'éducation nationale, de la recherche et de la technologie, June 1999, http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/004000935/0000.pdf.

[17] Silberman, R. (2011), La protection des données individuelles en France et la recherche en Sciences Sociales, in Chenu A., Lesnard L. (eds.), *La France dans les comparaisons internationales. Guide d'access aux grandes enquêtes statistiques en sciences sociales,* Paris, Presses de Science Po, 183−204.

[18] Silberman, R. (2013), Transnational access to official micro-data: The *Data without Boundaries* European network, in Kleiner B., Renschler I., Wernli B., Farago P., Joye D. (eds), *Understanding Research Infrastructures in the Social Sciences*, Zurich: Seismo Press, 47−66.

[19] Vardigan, M., Heus, P. and Thomas, W. (2008), Data Documentation Initiative: Toward a Standard for the Social Sciences, *The International Journal of Digital Curation,* 3(1), 107−113, http://www.ijdc.net/index.php/ijdc/article/viewFile/66/45.