Journal of Privacy and Confidentiality
Vol. 11 (2) 2021

Perspectives

Submitted January 2021
Published August 2021

DEVELOPING ACCESS TO CONFIDENTIAL DATA IN FRANCE: RESULTS AND NEW CHALLENGES

ROXANE SILBERMAN

CNRS & CASD, France
e-mail address: roxane.silberman@casd.eu

Abstract. Over the past twenty years, in various countries, secure access to data for the members of the research community was eased in a significant manner. Such data involve microdata and granular data from administrative records and detailed individual surveys. While some difficulties remain, the scene has been extensively redesigned, and new players emerged on both sides of the fence: data holders and users, both challenging what seemed to be well-established boundaries. In the French case, access to confidential data for research purposes has been carefully facilitated. Analyzing this process provides us with insights into how obstacles can be overcome and reveals newly emerging issues.

1. A wealth of confidential data available for research in France

Twenty years ago, in France, the practice of granting researchers access to anonymized microdata from public sources was just starting. In 1986, a first agreement was signed between the National Statistical Office, INSEE (*Institut National de la Statistique et des Études Économiques*) and CNRS (*Centre National de la Recherche Scientifique*), entrusting one of the CNRS research units specialized in quantitative sociology[1] with the responsibility of making a set of anonymized datasets available to all CNRS researchers, prior to any legislation in this regard. This process kept going for a decade, allowing access to an increasing number of survey data, including data provided by statistical departments in various ministries. However, significant difficulties remained to extend access beyond CNRS to all university-affiliated researchers. At the end of the 1990s, CNIL (*Commission Nationale de l'Informatique et des Libertés*), the French authority in charge of the implementation of the 1978 French Data Protection Act, *Loi Informatique et Libertés*, made the decision to drastically restrict the level of detail which would be available for the 1999 census in

Key words and phrases: Research, Privacy, Access, Legislation, France.
* Presented CRDCN..

Editorial note: This article is an edited version of the author's talk at the October 2020 Canadian Research Data Centre Network (CRDCN) conference. Information on the conference can be found at https://www.crdcn20.ca/crdcn20/program. *Articles in the Perspectives series reflect the author's opinions, and do not necessarily reflect the opinions of the journal's editorial board.*

[1] *LASMAS (Laboratoire d'analyse secondaire et de méthodes appliquées à la sociologie)*

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.788

the anonymized files provided by INSEE to the research community. Within this context, the present author was asked by the French Ministry in charge of research to organize a broad consultation, gathering various stakeholders, within the framework of a *Lettre de mission*[2]. The resulting report *"Les sciences sociales et leurs données"*, submitted in 1999, underlined the need to systematize access to government data, and ensuring access to highly detailed microdata. This involved changing the legal framework and setting up a secure access system.

It took a further eight years to open up access to confidential microdata in France. This process took a while to take off, but the initiative has now gained momentum, and France is probably one of the leading countries providing researchers with secure access to confidential data. The data available through the CASD (*Centre d'accès sécurisé aux données*) Secure Data Hub[3] set up in 2008 now covers a vast range of data from diverse fields and from public and private producers alike. Censuses, panels, and a large number of surveys conducted by *INSEE* are available at the finest level of detail. This is also the case for most surveys produced by the various statistical departments in the ministries (SSM, *Services statistiques ministériels*), with which CASD has progressively signed agreements. The available data thus covers a broad array of fields: work, employment, education and health, agriculture and the environment, with data concerning both households and companies. An agreement with the DGFIP (*Direction Générale des Finances Publiques*) within the Ministry of Finance has made tax data available, while another agreement with the Ministry of Justice provides access to ten years of criminal records. Data provided by various public agencies, such as ACOSS (*Agence centrale des organismes de sécurité sociale*), the Central Agency of Social Security Organizations (CNAF, *Caisse nationale des allocations familiales*), the National Family Allowance Fund, Pôle Emploi, the Unemployment Agency, and the Public Investment Bank (BPI France, *Banque publique d'investissement France*), were gradually added. The expanding set of agreements triggered new possibilities for matching these datasets via the national individual identifier, for research purposes. In addition, thanks to its certification as a highly secure remote access system, CASD was able to sign agreements to host data from several epidemiological cohorts (panels) supervised by academics, including one of the largest in Europe, CONSTANCES. Moreover, due to the legal obligation to comply with the rigid security baseline meant to protect privacy for all analyses based on the medico-administrative data of the the National Health Data System (SNDS, *Système national de données de santé*), many organizations active in the health sector now rely on CASD in order to pursue health data analysis, and to carry out data matching operations with SNDS data for clinical studies, or for enriching registers and cohorts.

2. Key drivers

As the process evolved, major improvements were made, including a change of the legal and regulatory framework pertaining to privacy. These changes will also be important for addressing new challenges.

Redesigning the legal framework turned out to be necessary in France, as in most countries. Indeed, the legal context was historically and widely viewed as the main obstacle to data access, sometimes presented as overwhelming. This proved inaccurate. The 1951 Statistics Law was first modified in 1984 to allow researchers access to corporate data, making

[2] *Les sciences sociales et leurs données, La Documentation française 1999*

[3] https://www.casd.eu/

France a pioneer in this domain. Access to individual and household microdata took longer, requiring changes to both the 1978 Data Protection Act (*Loi Informatique et Libertés*), and to the 1951 Statistics Law. The Data Protection Act, long perceived as untouchable, and the source of tensions between the research community and the data protection authority, was first amended in 1984 to include provisions for research pertaining to health data. In 2004, in the process of transposing the 1995 European directive to cover the protection of personal data, provisions for research, statistical and historical purposes were introduced into the French Act, now extended to cover data archiving, in line with the 2016 European General Data Protection Regulation (GDPR).[4] In line with these changes, the Statistics Law was once again amended in 2008, extending privileged access to the research community for individual and household microdata.

Over the last ten years, the legal and regulatory frameworks has undergone a new set of important changes in order to ease access to confidential data for research purposes.

Various administrations or public entities have been reluctant to make their data available to the research community, arguing that their professional deontology is at stake. Such was the case for tax data, to which researchers had been requesting access for many years. Access to these data became all the more necessary as they were increasingly used by INSEE instead of survey data. The exemption from tax secrecy for research undertakings required an amendment to the Book of Tax Procedures (*Livre des procédures fiscales*), implemented in 2013. Regarding health data, the 2017 Health Law (*Loi sur la santé*), granted wider access to a range of health-related data from medical and administrative sources, covering the entire French population. Other administrative data sources may contain information protected under professional confidentiality rules. In 2018, several provisions were included in the new Digital Republic Law (*Loi pour une République Numérique*), facilitating the use of all administrative data. A first provision, in connection with the Archives Act (*Loi Archives*) which covers all government data, authorizes data holders to provide access to their administrative data for research purposes, even when such data are protected by professional confidentiality rules. This presents an interesting attempt to address the issue of potentially having to change regulatory texts one by one for each body or service. Two other provisions, one for the statistical system coordinated by INSEE, the second for the researchers, authorizes the possibility to match these data using the national identification number (NIR, *Numéro d'inscription au répertoire de l'INSEE*). The matching procedure requires the involvement of trusted third parties, the use of a irreversible hashing process, and strict requirements regarding the destruction of the matching keys.

In order to convince the relevant institutional bodies to perform these changes, a variety of actors had to be involved.

Impact analyses required concrete examples of advances in research made possible by changing legal frameworks. Previous research projects, sometimes conducted within the various ministries through an *ad hoc* status allowing access to data, also played a role in gaining the support of data producers for changing the law. The involvement of INSEE was crucial, contributing a tradition of analysis and the existence of a research department, CREST (*Centre de recherche en économie et statistique*), initially set up within the institute.[5] The CNIS umbrella (*Commission Nationale de l'Information Statistique*), a meeting place for users and producers of official microdata, turned out to be a very useful sounding box for changing laws. Data requirements for the assessment of public policies (for

[4]The GDPR came into effect in 2018.

[5]CREST is now an independent entity affiliated with the French university system.

which researchers are strongly mobilized, some of these being in think tanks attached to Prime Ministerial services, such as the *Centre d'Analyse Economique* or *France Stratégie*) contributed substantially to changing the legal framework. In its recent annual report[6], the *Conseil d'État* thus underlined the decisive role of access to confidential data played by CASD for improving the assessment of public policies in France. Another factor has been the pressure exerted by the *Open Data* initiative. Despite the fact that it primarily concerns anonymized data available to the general public, it nevertheless sheds light on the necessity to design a framework enabling access to confidential data for authorized users. Significantly, provisions which facilitate access to and matches between administrative datasets are implemented in the 2018 Digital Republic Law.

Security concerns were a key driver while opening access to confidential data. Creating highly secure access has been at the heart of this process. It has fostered the confidence of producers and citizens alike and greatly facilitated legislative development. The creation of the CASD Secure Data Hub within INSEE in 2007 thus preceded an amendment to the 1951 Statistics Law by a few months. As a trusted third party between producers and users, CASD became an independent public entity in 2019. Set up relatively late, compared to several other countries, where onsite access has been implemented since the early 2000s, CASD benefited from the technological advances made since then, opting without further delay for remote access as the sole method of access. Many remote access systems rely on software which needs to be installed on the user's workstation. The solution designed and fully controlled by CASD is an integrated system, involving a dedicated access device (the "SD-Box") and biometric authentication. This integration enables end-to-end control of security, offering both greater security for the data provider, and ease of deployment and usage for the users. Likewise, the CASD architecture facilitates the recommended security certifications, in particular ISO 27 0001 and RGPD certifications, because each component of the chain is fully controlled by CASD, without any additional constraint for the user. This high security level is combined with great ease of work for the user until the final drafting of the articles inside the "secure bubble" of the project.

Strong and transparent governance has been a third element in France to facilitate the opening of access to confidential data. As in most countries, an important component of security is the user authorization procedure for assessing "safe researchers and safe projects" in line with the *Five Safes framework*, an internationally recognized approach for managing decisions on access to confidential data.[7] Though developed without reference to the Five Safes, the overall process implemented in France for data access is fully in line with its guidelines. Data are classified, according to the level of details and risks, as *Public Use Files (PUF*), *Scientific Use Files* (*Fichier de Production et de recherche, FPR*) and *Secure Use Files*, in line with the Eurostat classification. While the *PUF* can be downloaded on producers' website (for instance the Labour Force Surveys on INSEE website), access to the *SUF* is provided by secure download from the *Réseau Quetelet* data archive, for researchers attached to universities or research institutions recognized by the *Comité du Secret Statistique (CSS)*, a committee chaired by a judge and which represents data producers and researchers. For *Secure Use Files*, authorizations are submitted for each project and research team by the CSS, ensuring projects and users are "safe" in coordination with the

[6] *Conduire et partager l'évaluation des politiques publiques,* La documentation française, 2020

[7] The Five Safes framework includes the five dimensions: safe researcher, safe project, safe data, safe settings, safe outputs. See Desai, Tanvi, Felix Ritchie, and Richard Welpton. "Five Safes: Designing Data Access for Research." Working Paper. University of the West of England, 2016. http://eprints.uwe.ac.uk/28124.

bodies in charge of data privacy protection (CNIL and National Archives). The principles are broadly comparable to those upheld by other countries: applicants have to provide a detailed description of the project, including the need for confidential data, the research or study purpose of the work, the methodology, and the list of data sources required. Secure settings are mandatory and secure remote access is provided for most data sources via CASD. User awareness of confidentiality issues is an important dimension, contributing to "safe" behavior among data users. Authorized users are required to attend CASD training sessions[8] on security covering legal aspects (user responsibility and sanctions in case of breach of confidentiality) and the anonymization rules applied to research outputs (safe outputs).[9] Transparency, which is important for producers, users and citizens alike, is also an element of the governance. All research projects are posted on the CASD website with links between data, projects, and publications. CASD has also set up a cooperation with the *cascad* (Certification Agency for Scientific Code and Data Agency, a collaboration between UMS 2007, École des Hautes Études Commerciales de Paris-HEC Paris, Université d'Orléans, and CNRS) to assess, prior to submitting a paper to a journal, the reproducibility of research results based on confidential data hosted by CASD. Such verifications would normally take a long time to obtain if reviewers have to go through the procedure for accessing confidential data.[10]

3. New challenges

This major initiative has undoubtedly proven successful: a number of sources covering various areas are now open for research in France, with more than 2,000 users across 600 institutions. This development generates new challenges that will become increasingly dominant in a data landscape, widely different from what it was 20 years ago: New data sources, new data holders, new tools. Two major challenges can be outlined and illustrated in the French context.

Joint use of data from different sources will be increasingly decisive for the advancement of knowledge and our future society. Issues such as sustainable development and major risks and their impact on society require an analysis of the interactions between various fields. This involves a joint use of highly detailed public data sources with academic and private sources, combining socio-economic data, health data and environmental data. Already, only 10% of research projects hosted by CASD are based on a single data source. The recent global COVID-19 outbreak provides an insightful illustration in this respect. In order to deal with issues such as climate change, energy transition, and sustainable development, there is an increasing need to use data from private companies demanding a secure environment in order to be matched at a very fine spatial level with other socio-economic or health data. This entails several challenges:

(1) Research projects requiring access to very large files will become more frequent, requiring requiring increased storage and computing power, which must be provided without violating legal requirements and security constraints;

[8]Training is repeated every four years.

[9]Research outputs are verified by CASD staff.

[10]For more details, see Pérignon, Christophe, Kamel Gadouche, Christophe Hurlin, Roxane Silberman, and Eric Debonnel. "Certify Reproducibility with Confidential Data." Science 365, no. 6449 (July 12, 2019): 127–28. https://doi.org/10.1126/science.aaw2825.

(2) Silos will become increasingly problematic if data are stored in multiple secure centres with different technologies and systems of governance, as is currently the case for much health data or for central banking data. This will require either the development of cooperation networks – currently under discussion in France between CASD (which provides access to INSEE data) and the *Open Data Room* of the Banque de France – or more centralization into a small number of data centers, along with the fears this may arouse for citizens;
(3) New matching possibilities raises the problem of long-term conservation of administrative databases in a context where the legal framework for personal data tends to limit their availability over time. In France, the applicability of the *Loi Archives* (Archives Act) for selecting and keeping data beyond the legal duration was discussed with the data protection authority in the context of a CNIS report. It will also be increasingly necessary to support matching requests, developing reference methodologies and tools which can be shared.
(4) Data held by private companies are valuable resources, typically data on energy consumption. Meanwhile national statistical institutes are beginning to collect corporate data: INSEE is already computing price indexes based on supermarket data; discussions are underway with telephone operators – provisions may not include researchers' access through this channel.

Overcoming barriers for transnational access to confidential microdata for research still remains a major challenge, despite many past efforts. Decentralized remote access is still not in place for European integrated microdata, although it was authorized by the 2013 European regulations. Access to confidential national data across borders is still unevenly and differently organized, making the use of non-integrated data, especially administrative data, from several countries a challenge[11]. Discussions with international partners such as OECD, and projects such as the 2011-2015 European 7th Framework Programme, *Data without Boundaries (DwB)* and the *Nordic Microdata Access Network* (*NordMAN)* project revealed the difficulties to build from scratch an "ideal world" which might allow researchers to access data from several countries with a single piece of equipment and a single access procedure. Drawing lessons from those earlier efforts, the *International Data Access Network (IDAN)* initiative aimed to create a concrete operational international framework based on a "quick wins" strategy via reciprocal provision of Safe Room Remote Desktop Access facilitating researchers' access to confidential microdata from various countries. *IDAN* https://idan.network/, founded in 2018, involves six Research Data Centres (RDCs) from France (CASD, coordinator), Germany (FDZ of IAB - Research Data Centre of the Institute for Employment Research, and GESIS - Leibnitz Institut für Socialwissenschaften), the Netherlands (CBS - Statistics Netherlands), and the United Kingdom (SRS-ONS, Secure Research Service - Office for National Statistics, and UK Data Service). As a first step, access points in partners' *safe rooms* have been set up via bilateral contracts, and practical agreements have been discussed for handling differences in requirements for surveillance of the *safe rooms* (see Figure 1). Although this requires users to physically visit one of the Network's Research Data Centres, it does enable them to work remotely on analyzing data provided by partner countries solely from within their local RDC. Building trust between partners/countries, involving the research fellows who have started working using the new

[11]Silberman Roxane, 2013, "Transnational access to official microdata: The Data Without Boundaries European Network" in Kleiner Brian, Renschler Isabelle, Wernli Boris, Farago Peter, Joey Dominique, 2013, *Understanding Research Infrastructures in the Social Sciences,* Seismo Press," Zurich, p.47-66.

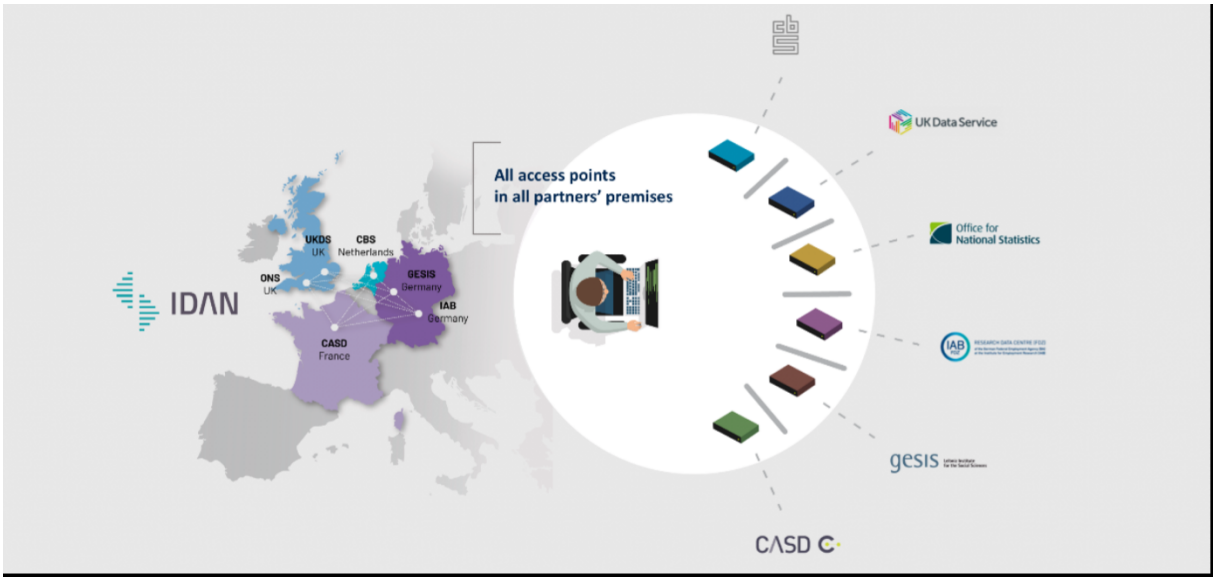

Figure 1: IDAN Network

facilities, presenting showcases, developing work on data comparability[12] are all paving the way for the next steps towards a multilateral network agreement (a basis for extension to other countries and partners), and laying the grounds for progress enabling research teams to combine data from different sources in the same secure environment.

4. Conclusion

Many obstacles were overcome in the last twenty years, increasingly opening up access to a wealth of resources which have proven useful for advances in research as well as for shaping public policies. Breaking down both national and international silos to facilitate interdisciplinary use of data from different domains, data holders and countries will be crucial in the years ahead. The COVID-19 pandemic, with its implications reaching far beyond health in all areas, has strongly demonstrated this urgency and should be a driver for meeting these challenges. Security and transparent governance will be essential for maintaining citizens' confidence.

[12]Laible, Marie-Christine; Seilles, Marine; Alkhoury, Maria; Fleureux, Raphaëlle (2020): *New opportunities for comparative cross-country research in France and Germany.* FDZ-Datenreport, 03/2020, DOI: 10.5164/IAB.FDZD.2003.en.v1.

5. Glossary

- ACOSS: Agence centrale des organismes de sécurité sociale
- BPI France: Banque publique d'investissement France
- CASD: Centre d'accès sécurisé aux données
- CNAF: Caisse nationale des allocations familiales
- CNIL: Commission nationale de l'informatique et des libertés
- CNIS: Conseil national de l'information statistique
- CNRS: Centre National de la Recherche Scientifique
- CSS: Comité du secret statistique
- DGFIP: Direction générale des finances publiques
- FPR: Fichier de Production et de Recherche
- GDPR: European General Data Protection Regulation
- IDAN: International Data Access Network project
- INSEE: Institut national de statistique et d'études économiques
- NIR: Numéro d'inscription au répertoire de l' INSEE
- SNDS: Système national des données de santé
- SSM: Services statistiques ministériels
- SUF: Scientific Use File

Disclosures

Mme Silberman is scientific advisor to the CASD, which is discussed in this article. The section editor of the "Perspectives" articles, in which this article appears, is the chair of the scientific advisory board of the CASD.