Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Paola Tubaro & Roxane Silberman

Access to Governmental Microdata for Research: Recent Developments and New Challenges in Europe.

IASSIST Conference 2009

Session B2, Wednesday, May 27

cessda ppp

SciencesPo. INED

cnrs LECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES Ens

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Introduction

Technical improvements and advances in statistical tools and methods have boosted researchers' interest in government microdata.

At the same time, production of these data has mushroomed, taking multiple forms.

Access to government microdata has now become a crucial issue for researchers, and consequently for Data Archives.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Introduction (cont.)

It is unclear what kind of infrastructure can address these needs, and whether it can be a simple extension of the 70s solution of national Data Archives, with CESSDA at European level.

Parallel –and potentially competing– structures may emerge. Some of them are already under way, involving a variety of actors, products, and practices.

A new environment appears: Data Archives must take it into account, positioning themselves and their know-how.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

This talk

- an assessment of the current data situation across Europe, based on a survey conducted in the framework of CESSDA PPP and discussed at a workshop co-organized with Eurostat and ONS (December 3-4, 2008);

- a reflection on possible directions for the future.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Outline of the talk

1. A complex, heterogeneous data landscape;
2. Evolving modes of access;
3. Who gives access to what;
4. Researchers' accreditation;
5. Costs of access;
6. Access to Eurostat data;
7. Conclusions.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

1. A complex, heterogeneous data landscape

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

A complex, heterogeneous data landscape

The legal framework (Statistical Law and Privacy Protection Law) sets limits to data dissemination.

Since the 1990s, growing concerns about privacy protection have intensified problems of access to confidential data.

The EU 1995 Directive on protection of personal data first recognized the need to process data for purposes of scientific research.

Some national laws have followed suit (France, Portugal, UK), but differences in legislation persist across countries.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

A complex, heterogeneous data landscape (cont.)

Besides heterogeneity in legislation, countries differ in their interpretation of the law, especially of confidentiality safeguards.

Differences in the structure of national statistical systems (e.g. degree of centralization) also affect access conditions.

Surging demand of administrative data and of combined survey / administrative datasets raises new problems.

Technical developments make available new modes of access and favor a model in which producers directly disseminate their data.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

2. Evolving modes of access

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Major existing modes of access

Public Use Files and Scientific Use Files:

- For researchers vs. for the general public and/or students (CAMPUS files);
- Differing degrees of detail (*de facto* vs. full anonymization);
- License vs. no license.

Tabulations:

- Public tabulations vs. special (bespoke) tabulations;
- Spectacular recent development of web-based tools to prepare tables.

New solutions for confidential data:

- Secure on-site data centers on the premises of statistical agencies;
- Secure remote connections through the Internet: remote *access* vs. remote *execution*.

	1. Public Use Files	2. Scientific Use Files	3. Extracts (subsets)	4. Public tabulations	5. Special (bespoke) tabulations	6. Secure remote access / execution	7. On-site safe centers
a. Population Census, Register, or Microcensus							
b. Main Household Surveys							
c. Some Data from Administrative Registries							
d. Some Business Data							
e. Some other economic and financial data							
f. Other							

Table 1. Current modes of access, regardless of distributor, in European countries.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Table 1 provides evidence that:

There has been significant progress in granting access to anonymized household survey data and to census data.

Yet even access to fully anonymized data remains difficult and opaque in some countries.

On the whole, access to administrative data, business data, and other economic data (e.g. tax data) remains problematic and uneven.

Secure modes of access to more detailed datasets (safe centers and remote connection facilities) are undergoing a rapid evolution.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

3. Who gives access to what

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Multiple actors

Government data production is undertaken by a variety of actors:

- National Statistical Institutes (NSIs); statistical services of ministries; registries; Central Banks; and other public administrations at national and sub-national levels.

Dissemination of government data may be undertaken by the producers themselves, or by third-party intermediaries:

- CESSDA Data Archives; other national data organizations; internal data services of particular higher education and research institutions.

	Access granted directly by producer through the Internet	Access granted directly by producer in other form (onsite, post)	Access from a CESSDA data archive	Access from a non-CESSDA distributor
Public Use Files				
Scientific Use Files				
Extracts (subsets)				
Public tabulations				
Bespoke tabulations				
Secure remote access / remote execution				
Onsite safe centre				

Table 2. Who distributes data, by mode of access, in countries in which there exists a CESSDA Data Archive.

no country | 1-3 countries | 4-6 countries | 7-9 countries | 10-12 countries | ≥ 13 countries

does not apply

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Table 2 provides evidence that:

About half of CESSDA Data Archives disseminate fully anonymized and/or factually anonymized data (PUFs / SUFs).

In addition, a small number of data archives (France, Norway, UK) actively cooperate with governmental statistical agencies to set up and manage systems of access to sensitive data.

Not all data archives are involved in the dissemination of government data: some only offer information and support services to users.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Table 2 provides evidence that (cont.):

In some countries, there is a tradition of direct and exclusive dissemination of anonymized data by NSIs.

Web-based tabulation is a widespread and fast-growing solution, often managed directly by government data producers; some NSIs are also using their websites to disseminate PUFs.

Growing access to confidential data (both onsite and remote) is mainly managed directly by government data producers, who might in future offer coordinated access through a European network.

⇒ Is a system of parallel structures emerging, potentially diverging from the network of CESSDA-mediated distribution?

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

4. Researchers' accreditation

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Diverse accreditation systems and limited data circulation across borders

Accreditation is the process of defining criteria to identify research and researchers, implementing criteria, and managing applications.

Multiple actors can be involved, e.g. National Statistical Institutes, Statistical Authorities, Data Archives, Research Institutions.

Requirements differ across countries: affiliation to an acknowledged institution, submission of a research project, experience, etc.

Recognition of researcher status and research purpose is often problematic for foreign researchers, even within Europe.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

5. Costs of access

	Provided by the producer, free of charge for users	Fees paid by end users	Charges are paid by a higher education or research institution, for all affiliated members	Charges are paid by Data Archive, Ministry, or National Research Council, for the whole scientific community
Public use files				
Scientific use files				
Extracts (subsets)				
Public tabulations				
Specific (bespoke) tabulations				
Secure remote access / execution				
On-site access through safe centres				

Table 3. Who covers costs of access to government data in Europe.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Table 3: Diversity of cost covering arrangements

Many countries have cut costs of access. Where costs subsist, they are sometimes offset by a Data Archive or national Research Council on behalf of the whole scientific community of a country.

Yet in several countries, researchers still pay fees individually.

Such disparities have repercussions on access to Eurostat data: there is currently no arrangement for a Data Archive or other central institution to cover costs for the wider research community.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

6. Access to Eurostat data

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Access to Eurostat (and other international) data

Existing modes of access: SUFs and an on-site safe center.

Eurostat negociates agreements with institutions, not individual researchers ⇒ Important role of research and higher education institutions as intermediaries.

No role for Data Archives (or other national institutions) so far.

Similar conditions prevail in other international organizations (e.g. OECD).

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

7. Conclusions

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Conclusions

The current situation suggests the need for a central infrastructure but a variety of directions might be taken.

In particular, there are signs of development of parallel structures, all the more so as NSIs are autonomously developing secure modes of access to confidential data.

This might lead to greater heterogeneity in the European data landscape and a possibly diminished role for Data Archives, with ensuing gaps in data documentation and other services.

Yet the renewed forms of cooperation that exist between a few Data Archives and NSIs set a promising example.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Conclusions (cont.)

Driven by technical change and growing demand for microdata, the new data landscape may lead to:

- a highly fragmented system, or
- a reconfiguration to build a multi-actor governance system.

The latter solution fits better with researchers' needs for open access, homogeneity of conditions, transparency, and high-quality documentation.

Ongoing work to upgrade CESSDA into a new European Research Infrastructure points to this direction.

Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Thank you!

Paola.Tubaro@ens.fr

Roxane.Silberman@ens.fr

cessda ppp