CONFERENCE OF EUROPEAN STATISTICIANS
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 1.1
29-31 October 2019, the Hague, the Netherlands 18 October 2019

Harnessing the potentiality of microdata access risk management model

Natalia Volkow (National Institute of Statistics and Geography of Mexico)
NATALIA.VOLKOW@inegi.org.mx

Abstract and Paper

In 2008, Mexico promulgated a new National Statistics and Geography Information System that included an article regarding the provision of access to microdata for research purpose. Base in this Law, the National Institute of Statistics and Geography of Mexico, inaugurated in April 2013 its Microdata Laboratory. The services was organized following best international practices The demand for the service rocket but not only in number of applications but in the complexity of the research, users wanted to undertake. The challenge was interoperability among different statistical projects and other information sources by unit of analysis, safeguarding confidentiality and managing within an environment of scare resources. The Institute did not had capacity of response to make the cross roads among identifiers. The solution was to innovate how to use the technical configuration of 5's model, so users could carry out the linkage without breaching confidentiality. The organization of what we called virtual panel harness the potentiality of microdata available for research and of the statistical infrastructure of the country, with no impact on costs of operation.

Harnessing the potentiality of microdata access risk management model

Natalia Volkow*

National Institute of Statistics and Geography, natalia.volkow@inegi.org.mx

Abstract: The paper presents how it was harness the potentiality of risk management model to tackle the need of interoperability between different year statistical projects or data sets to support research safeguarding confidentiality and enhance the integral security of the process of serving microdata access.

1 Background

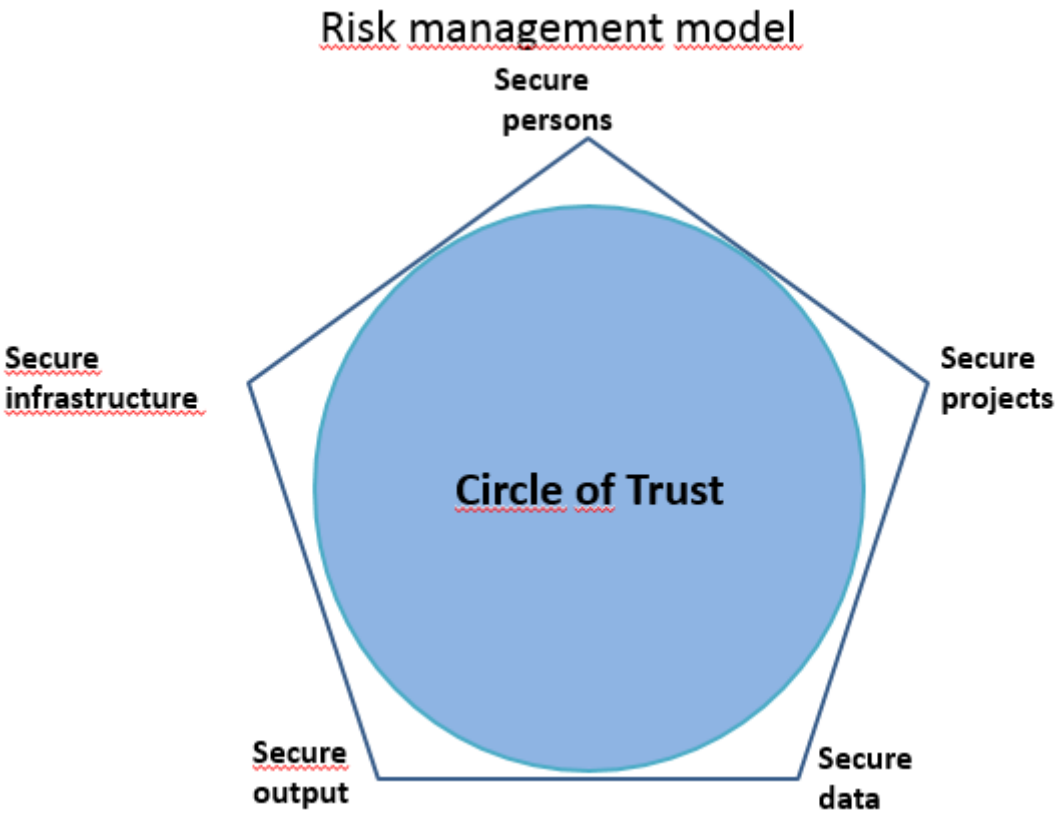
The National Institute of Statistics and Geography was created in 1983. The Law of the National System of Statistical and Geographic Information was published in April 16, 2008 provided autonomy to the Institute. The article 100 of this Law establishes that the Institute, following best international practices, will provide access to microdata. In the first session of the Board of Governors of the Institute, they authorized the terms and conditions in which the Institute will provide access to microdata of all statistical projects carried out by the Institute.

This agreement established that access to household surveys, government censuses and the samples of the household and population censuses will be provided direct from the Institute web site by download of the files of the household surveys that contain raw microdata. These are probabilistic surveys, confidentiality is kept by limiting to municipality level the geographic location of microdata of the random sample.

This agreement also established that access to microdata from surveys and censuses of establishments and agricultural units and from the population and household census will be provided for free through a microdata laboratory and remote processing only for research purpose or to public servants that are defining, operating or evaluating public policies. The Institute also kept in operation the regular public service of processing specific tabulates, upon request with charge.

2 Microdata Laboratory

The Microdata Laboratory was inaugurated in April 2013, it was developed following the risk management model of the 5 secure elements - infrastructure, persons, project, data and output- developed by the UK Office on Nations Statistics. Putting in place these five secure elements creates a “Circle of trust” (OECD, 2014). Within this space users can have access to confidential data to do the processing they need for their research.



INEGI Microdata Laboratory in a secure enclave in its premises located in Mexico City. Only authorized personnel of the Institute can have access to this space. In 2019 it was inaugurated the first remote access Laboratory in an academic institution in Mexico City. Both laboratories operated upon the same rules and similar infrastructure. The logical security has been developed with the same technological platform than that of the Secure Data Services of the UK Data Archive. The platform allows the creation of virtual desks that are logical spaces allocated to specific research projects where the microdata requested by the user is loaded, with no direct identifier. The geographic

location of microdata varies for each statistical project, the most disaggregate that is possible is block level, when it is available. Precise location -address of coordinates- is considered a direct identifier.

To each virtual space only authorized users with specific privileges can have access through thin clients. The personnel of the areas that generate information can access to input the data need for a session of a specific user or extract the output from users processing to check they do not breach confidentiality, before it´s delivered to user.

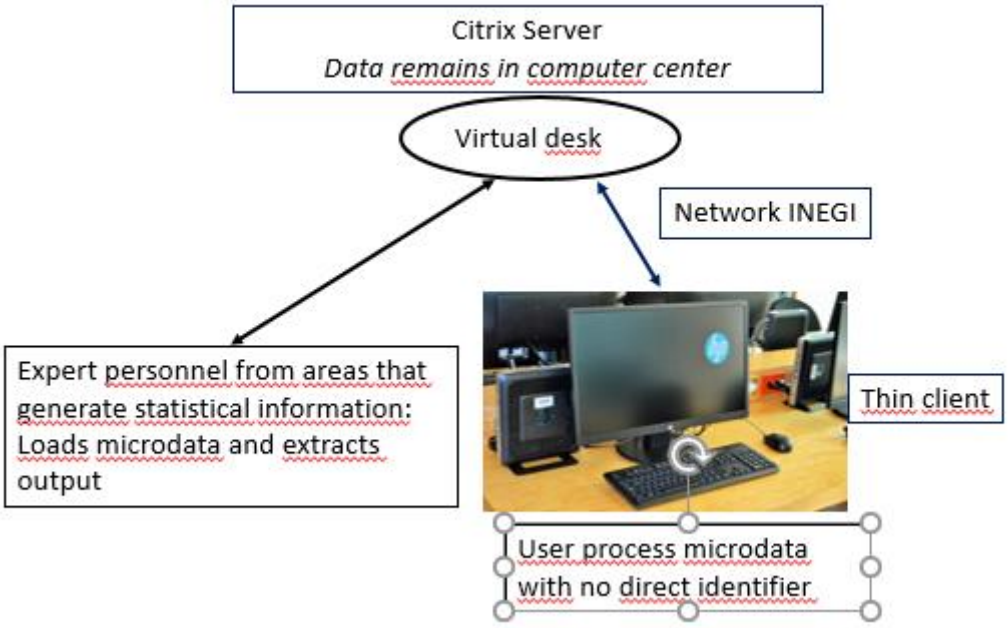
The main Laboratory has 21 terminals and the remote access has three. Both are in secure spaces, access is control through biometric systems and only authorized personnel can enter. They both has CCTV 24x7 hours surveillance.

As for the secure persons, only public servants, postgraduate students and researchers of an academic or research institutions, or personnel from international organizations can apply to the service of the Laboratory. Each institution must sign an agreement with INEGI to provide the accreditation of the people affiliate to it. Following the practice carried by the Secure Data Service and the UK Economic and Social Research Council, INEGI signed an agreement with the National Council of Science and Technology (CONACYT) by which the researchers and students that are receiving benefits from the Council will have its accreditation and will be able to use the Laboratory.

f

Users must fill an application format, provide copy of their Curriculum Vitae, document of proof of institutional affiliation and official identification. In the format, they must describe the aim of its project, methodology and justify the need to have access to microdata. The research project must serve public good, by supporting the definition or evaluation of public policy or else academic research to be published, so it is publicly available. The data required must be justified in terms of the research project aim. The output of the processing is checked by the personnel of the areas that generate information to validate they do not breach confidentiality.

The procedure was defined like a star, having one single point of contact with users that manages the relation with all the areas of INEGI that generate statistical information. It is the experts of each statistical project that load the microdata in the virtual desk and them who extract the output and check it. I it does not breach confidentiality, they cleared it and send it to the single point of entry that delivers it to the users.



Each research project has a key identifier (LM 1111) and is allocated to a virtual desk that is a logical space within a high security server located in the city of Aguascalientes where INEGI main quarters are. The data does not leave the secure environment of INEGI secure computer centre.

The users carry out their processing through a thin client that is connected to the virtual desk that is allocated to his or her research project. The virtual desk created in the Citrix server, has not connection to any other resource, it contains the following folders:

- **Processing**
    - **Inputs** - microdata requested by user is loaded here and it cannot be change, microdata can only have blind identifiers
    - **Work** – the user can manage this folder as he or she needs
- **Results** – in this folder the user save the outputs that will later request for clearance. He or she needs to save the results following the confidentiality guidelines, with the standard nomenclature and filling a structure format to describe what is what he or she is requesting for clearance, if it is an algorithm, results, log file or notes. If its results he or she must explain what the results are and how he or she obtained them. The word file with the structured descriptive format is in this folder.

Indicar·con·X·si·solicita·revisión·de:·¶	Carpeta·de·archivos¤		Archivo¤	
Nombre·del·archivo·o·carpeta·de·archivos·a·revisar¤	¤			
Indicar·con·X·si·solicita·revisión·de:¤	Resultados¤	Algoritmo¤	Notas¤	Logfile¤
Fecha·de·solicitud·de·revisión¤	¤			
Proyecto(s)·estadístico(s)·de·los·microdatos·utilizados(s)·en·el·procesamiento¤	¤			
Número·mínimo·de·observaciones·consideradas·en·los·resultados¤	¤			
Descripción·de·los·resultados¶				
¿Cómo·obtuvo·los·resultados?¶				

The nomenclature is standardised in such a way that when a file is download from the virtual desk anybody can recognize from which research project comes from. For the one point of entry it helps identify to which area of INEGI must be sent for clearance.

The folder name has a prefix that is also used in the descriptive format file and the file requested for clearance. The prefix has the number of research project, the acronym used for the statistical project and the date starting year, month day so when listed the most recent will be at the top. In this example CE stands for Economic census.

Folder¤	LM·614-·CE-2018-02-27¤
Output·file¤	LM614-CE-2018-02-27-·descriptivos¤
Description·format·file·¤	LM·614-CE-2018-02-27-·Leer-descriptivos.docx¤

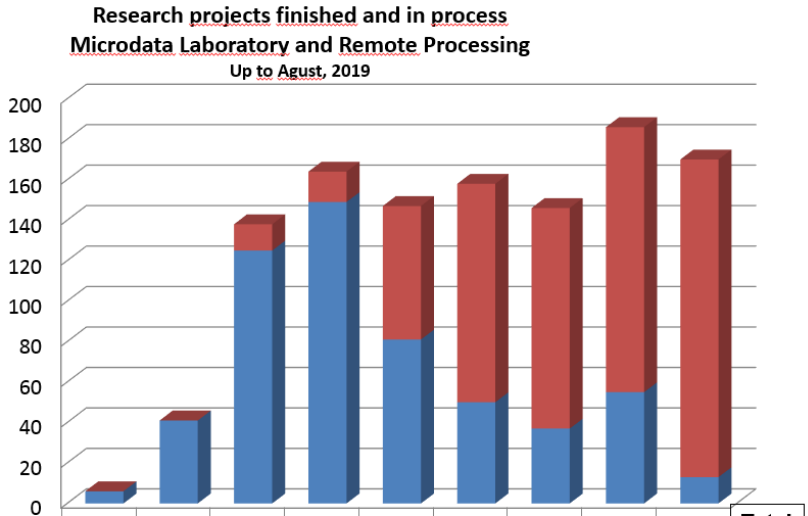
Training.

Before going into the Microdata Laboratory for their first session, every user must to undertake a training session the objective is to create consciousness about the importance of confidentiality. The content of the course is:

- Team work of what it is confidentiality and its importance for the statistics infrastructure of a country i
- Best international practices

- Context in which we live and the threats they represent to confidentiality.
- Procedures, nomenclatures and rules of the service.

Since the Microdata Laboratory was put in operation the demand has increased steadily. The service is provided for academic research on a free border basis and free of charge. To enter the Laboratory, users need to have an accreditation of their institution that must sign an agreement with INEGI, and he or she must sign some terms and conditions and undertake a training session. We have singed 61 agreements for accreditation, 10 with international organizations, 23 with foreign academic institutions and 28 Mexican academic institutions.



*Only remote processing	2011 *	2012	*2013	2014	2015	2016	2017	2018	2019	Total
In process	0	0	13	15	66	108	109	131	157	599
Finished	6	41	125	149	81	50	37	55	13	557
Received/year	**6**	**41**	**138**	**164**	**147**	**158**	**146**	**186**	**170**	**1156**

This growth has not only been in number of research project but in complexity as well.

Year	Research Projects	Remote Processing	Microdata Laboratory
2011	6	6	-
2012	41	41	-
2013	138	127	11
2014	164	143	21
2015	147	102	45
2016	158	88	70
2017	146	62	84
2018	186	41	145
2019*	170	75	96
Total	**1156**	**685**	**471**

*August 31

The regular size of a virtual desk is:

- 40 GB storage
- 4 GB RAM memory
- 2 processors

But we have cases in which have had to increase the capacity up to

- 40 GB storage
- 56 RAM memory
- 16 processors

This is just an example, but it shows the use of a big volume of data from different statistical projects. Users started requiring access to microdata to one specific statistical project or only by year, but soon they start asking the possibility of linking microdata from different statistical projects or different years of same statistical project.

From 2009 onwards INEGI created for establishment a unique identifier, so from that date on the linking is done through this unique identifier. For previous years

and for linking with external data sets the data linkage needs to be done, case by case, and using direct identifiers. INEGI did not have resources to do the worked required to do the linkage.

The only way to sort this need was to take advantage of the risk model that was operating in the Microdata Laboratory so users could do the linking themselves but safeguarding confidentiality. With the same configuration of a virtual desk we create a panel desk, with different rules to the virtual desk. No variables are allowed and the data sets that would be linked had to have the direct identifiers and a blind identifier. Users cannot request the clearance of any file from this type of panel desk.

The cross road or table of equivalences of blind identifiers must be done by the user in the panel desk. Once he has integrated them, he or she must ask INEGI personnel to pass the cross road of the blind identifiers to his or her virtual desk, explaining, what he did and what he or she is asking to be transferred to his virtual desk. If he or she is working with an external data sets, he or she must request INEGI personnel to load the dataset with blind identifiers into his or her virtual desk, and then he follows the regular process. The cross road that have been integrated are pass to the National Business Register and can be share to other users. But when they a shared to another user he or she must include the reference of the paper of the user that did the linkage.

In this way INEGI supports research, enhances the potentiality of the national statistic infrastructure safeguarding confidentiality, gives credits to the user that built the cross roads and support the efficient use of social researcher's time.

We are now trying to migrate the personnel of the areas that generate information to virtual desks to enhance information security and to standardize the statistical software. Nowadays the areas have older versions of the software that creates problems of accessibility when they need to cleared output. We ask users to save them in different versions depending of the statistical project they are working.

The personnel of the areas that generate information work with the databases in their own laptops, the virtual desks are in a secures servers with all the security of INEGI computer centre. When all the personnel of the areas that generate statistical information are migrated to this technological configuration, no microdata will be out of security firewall of the Institute. Information security is an essential element for safeguarding confidentiality. This working arrangement will also support that microdata access would be served only from the Datawarehouse, as one sole point of exit to microdata services, contributing to the integrity and security of the statistical information.

The linkage between household and establishments surveys, or linkage of persons or households cannot be done. There is still no institutional procedure to have access to administrative registers from public institutions for research purpose. For INEGI the microdata access services have made evident the need to work data architecture and to try to recuperate historical data to harness the potential of research. We have just achieved a first step, of a 1000 miles journey…

References

Ley del Sistema Nacional de Información Estadística y Geográfica, (2008) (Diario Oficial de la Federación México.

Ritchie, Felix. *Secure access to confidential microdata: four years of the Virtual Microdata Laboratory*, Economic & Labour Market Review May 2008, Volume 2, Issue 5, pp 29–34.

Organisation for Economic Co-operation and Development (OECD) (2014). *Microdata Access for OECD Research and Policy Analysis*, OECD: Paris France.