Historia y actualidad de dos muestras censales de población. Argentina, 1869 y 1895.

Diego Quartulli

IIGG – FSOC - UBA

Resumen

Esta ponencia detalla el proceso de realización y posterior digitalización de las muestras de los dos primeros censos de población de la Argentina, levantados en 1869 y 1895. Ambas muestras, selecciones aleatorias de los respectivos universos de cédulas censales del Archivo General de la Nación, fueron una de las actividades básicas Programa Población y Sociedad, que se desarrolló en el Centro de Investigaciones Sociales del Instituto Di Tella entre 1966 y 1970, bajo la dirección de Gino Germani y Jorge Somoza.

Más que en las muestras, el trabajo pone su atención en el proceso tecnológico con el cual se las realiza, usa y difunden. Particularmente se trata de mostrar, más allá del caso particular, la diversidad de problemas que implica la producción, conservación y difusión de bases de datos digitales y de su fuerte relación con la evolución de las tecnologías disponibles.

El estudio, debido a su amplia ventana temporal, puede servir como ejemplo empírico de como la evolución de las tecnologías hace cambiar el abanico de opciones institucionales viables que sin resentir la producción de bases de datos primarios, sean compatibles con un aumento de la difusión de ellas en la sociedad.

Introducción[1]

En la introducción de las narraciones es común presentar a los principales actores, en los tiempos y lugares en que transcurre la historia. En este sentido podríamos decir que esta historia comienza en el último tercio del siglo XIX, con la realización de los dos primeros censos nacionales de la Argentina, levantados en 1869 y 1895 por los gobiernos de Domingo F. Sarmiento y José Evaristo de Uriburu respectivamente.

Mucho tiempo después, la publicación del libro "Estructura Social de la Argentina" en 1955, es una referencia apropiada para introducir a Gino Germani, su autor, en esta historia. Este prestigioso sociólogo, tras casi una década como director e investigador del Instituto de Sociología y de la carrera de sociología en la Facultad de Filosofía y Letras de la UBA, decide migrar con sus investigaciones al ámbito privado, lugar supuestamente menos conflictivo que la universidad pública de esos años, y para ello crea en 1964 el Centro de Sociología Comparada en el Instituto Torcuato Di Tella, que luego, en 1966, se transformaría en el Centro de Investigaciones Sociales (CIS).

Germani pone en marcha un Programa de investigaciones denominado "Población y Sociedad" y en 1965 asocia al mismo al CELADE (Centro Latinoamericano de Demografía). Así pasa a compartir su dirección con Jorge Somoza, actuario-demógrafo argentino, que agregaba su amplia experiencia en investigación demográfica y docencia desarrollada en la División de Población de Naciones Unidas y, particularmente, en la CEPAL y el CELADE. Poco antes de la partida de Germani hacia la Universidad de Harvard a principios de 1966, se incorpora Alfredo E. Lattes, joven analista en demografía del CELADE, para

[1] *Una primera versión se presentó en las X jornadas de Sociología de la Universidad de Buenos Aires. Esta versión también se mejoró gracias a los comentarios de Alfredo Lattes, Manuel Riveiro y Alejandra Otamendi. Especialmente fueron importante los comentarios del primero acerca de algunos errores históricos del artículo anterior.*

coordinar las distintas actividades del programa bajo supervisión de Somoza y Germani, radicados en Chile y Estados Unidos respectivamente. Tiempo después, luego de una vinculación a tiempo parcial, se incorpora full-time al programa Zulma Recchini, demógrafa del CELADE y hasta entonces profesional del Consejo Nacional de Desarrollo (CONADE), para acelerar la investigación sobre migraciones en la Argentina; también se suma Ruth Sautu, recién regresada de Londres, para investigar sobre clases sociales, estructura ocupacional y desarrollo económico.

Hacia mediados de 1966, ya en el Centro de Investigaciones Sociales (CIS) del Instituto Torcuato di Tella, con Jorge García Bouza como director, se inician las tareas preliminares de las muestras censales a cargo de Lattes y Somoza. Junto a los investigadores ya mencionados, el grupo incluía a otros colaboradores, en particular, tres asistentes de investigación: María Cacopardo, María Muller y Raúl Poczter que desarrollaban distintas tareas en los proyectos del programa.

Unos años después entra en la historia Robert McCaa[2], historiador-demógrafo estadounidense, con quien Alfredo E. Lattes y Zulma Recchini comparten una estancia de posgrado en la Universidad de Pennsylvania en 1969-70. McCaa jugará un rol clave porque, con su intervención, hizo posible que las bases de las muestras censales sobrevivieran hasta hoy, posibilitando así que el autor de esta nota, las rescate, acondicione y las difunda para su aprovechamiento público, de ahora en más.

Otro personaje diferente, pero también clave en la historia fue, metafóricamente hablando, la tecnología, que casi nunca abandonó la escena. Es importante remarcar

[2] *En la actualidad Robert McCaa es muy conocido por ser uno de los fundadores de IPUMS (Integrated Public Use Microdata Series) la cual es una organización que almacena, edita y difunde a través de internet información demográfica de distintos países. Especialmente es conocida por los microdatos obtenidos a través de muestras de los censos originales de muchos países mediante convenios con diferentes países.*

la importancia del rol que juega el cambio tecnológico, tanto en el procesamiento de la gran cantidad de datos que implicaban las muestras como, más específicamente, su conservación y difusión. Desde otra perspectiva, esta historia también se relaciona con la problemática de la conservación y difusión de datos digitales en las sociedades modernas, en particular, en el ámbito científico. Hasta aquí, entonces, los principales actores e instituciones de esta historia.

Los dos primeros censos nacionales

Con la organización del estado argentino surge la necesidad de poseer información estadística sobre las características de su creciente población (Lattes, 1974) y la realización de los dos primeros censos nacionales es posible, justamente, por el progresivo desarrollo institucional del propio estado (Otero, 2007a). Los censos de 1869 y 1895 publican sus resultados tres años después del trabajo de campo, un tiempo corto considerando que no contaban con tabulaciones eléctricas[3] y que otros censos nacionales posteriores, por ejemplo el de 1947, demoraron muchos más años[4].

Estos comentarios son, simplemente, para asentar que los dos primeros censos nacionales y sus publicaciones son los primeros eslabones de una cadena de hechos que recorreremos en esta nota. Otros aspectos de estos censos, como sus dimensiones políticas, jurídicas e ideológicas, han sido atendidos y analizados por varios autores, entre otros (Mentz, 1991), (Otero, 1998),

[3] *Cabe recordar que Hernan Hollerith (uno de los fundadores de IBM) realizó la primera máquina eléctrica de tabulaciones para el censo de 1890 de Estados Unidos (Hollerith, 1889). De hecho la empresa que hoy se llama IBM (International Business Machines) desde 1911 hasta 1924 se llamó CTR (Computation Tabulating Recording Company) y la empresa original de Hollerith se llamó hasta 1911 TMC (Tabulating Machine Company).*

[4] *En esa época (1951) la oficina de censos de EEUU obtiene la computadora UNIVAC I (Universal Automatic Computer) que se considera la primera de las computadoras diseñadas para la venta. Hasta ese momento la oficina había usado distintas máquinas tipo "Hollerith".*

(Otero, 2007a), (Otero, 2007b), (Novick, 2004) y (González Bollo, 2010).

De los mencionados censos resultaron más de un millón de cédulas y los datos que ellas proveyeron posibilitaron la preparación de las tabulaciones censales[5] que luego se publicaron oficialmente (De la Fuente, 1872, y De la Fuente, Carrasco, & Martínez, 1898). En el caso de las muestras que nos ocupan, esas tabulaciones sirvieron también para evaluar algunas de las tabulaciones que resultaron de las muestras; en otras palabras, las tablas censales publicadas fueron los "parámetros" con los cuales se compararon algunas tablas de las muestras.

Cabe recordar que estos censos, como todos los censos de población, incurren en errores, por ejemplo, de omisión, mala declaración y otros derivados de su extensa organización. En las muestras aleatorias que se *extraen* de los mismos, se agrega el error muestral que, por definición, no poseen los censos. Todas estas cuestiones fueron tenidas en cuenta por Somoza y Lattes a la hora de evaluar la representatividad de las cédulas censales de los dos censos y de extraer las respectivas muestras (Somoza y Lattes, 1967, pp. 35-47)[6].

La visión de Germani

[5] *Cabe destacar que el origen de la palabra "tabulación" posee su raíz en la palabra "tabla" y esta es una degeneración de la expresión latina "tabula". Así, una tabulación es una forma de condensar en forma de "tabla" la información diseminada en millones de cédulas censales.*

[6] *Siguiendo el mismo léxico debe recordarse que casi por definición los censos, o sea la intención explícita de recabar datos acerca de determinado "universo" (conceptual) a través de operaciones empíricas (observar, contar, medir, etc.) efectuadas a toda su "población" (empírica), poseen un error no muestral mayor al que tendría una hipotética muestra aleatoria de esa misma "población". En otras palabras los errores muestrales son errores que implican los propios conceptos usados en la teoría del muestreo. En cambio en los errores no muestrales muchos son compartidos, aunque algunos, al ser función de la complejidad organizativa son mayores, ceteris paribus, en los censos que en las muestras de una misma población.*

En 1955, como señalamos antes, Gino Germani publicó "Estructura Social de la Argentina" (Germani, 1955), la investigación que, según Graciarena (1987, p. 7)[7] usara la mayor cantidad de datos censales hasta ese momento publicados. Efectivamente, utilizó los cuatro primeros censos nacionales, más datos inéditos del IV censo de 1947 y datos de censos provinciales.

Es muy probable que Germani, luego de analizar este enorme volumen de datos secundarios[8], desagregados por variables y unidades espaciales, y de tener que sufrir la demora de la publicación del censo de 1947 (ocho años) por lo que debió trabajar con tablas inéditas en sus originales a lápiz[9], haya fortalecido sus convicciones acerca de las potencialidades que tenía, para la investigación social, el libre acceso a los datos censales desagregados y, de ser posible, a muestras censales. De hecho, Germani fue en la región, un activo impulsor de estas actividades tal como lo refiere Alfredo E. Lattes (Lattes, 2010, p. 407).

Germani no quería reimprimir su libro "Estructura…." (Graciarena, 1987, p. 16) y esta negativa, no solo se explicaba porque Germani era consciente de algunas limitaciones del libro[10] sino porque se proponía extender el análisis hasta 1960 y, especialmente, profundizar esta investigación. Para ello inicia el Programa

---

[7] *El otro libro, quizá comparable en la utilización de fuentes censales hasta la época, fue "Una Nueva Argentina" de Alejandro Bunge* (A. Bunge, 1940).

[8] *Es importante destacar que la distinción usual entre datos primarios y secundarios gracias al avance de los "microdatos" y su posterior difusión por internet ha ido achicando la distancia entre ambos tipos de datos debido especialmente a la ampliación del dominio de lo posible con datos secundarios (McCaa, 2013). Por supuesto este comentario no elimina lo útil de su diferenciación sino que la relativiza en función de los cambios sucedidos. Aunque el comentario está dirigido principalmente a datos "cuantitativos" también cuenta para los "cualitativos" (Bishop, 2006)(Bishop, 2007).*

[9] *Graciarena comenta que se publicó sólo una tercera parte de los tablas que se habían procesado (Graciarena, 1987, p. 8). La recuperación de las tablas del censo de 1947 que nunca fueron publicadas, constituyó otra actividad del Programa Población y Sociedad, también a cargo de Alfredo E. Lattes.*

[10] *El mismo Germani dice en la introducción de su libro "En verdad el trabajo tiene todas las limitaciones características de un primer ensayo, de una exploración previa..."(Germani, 1955, p. 16).*

"Población y Sociedad" que incluía, además de éste proyecto suyo, otros estrechamente relacionados al mismo, tal como se puede leer en la propuesta e informes del referido Programa[11] y en la correspondencia que mantuviera entre 1966 y 1968 con Alfredo E. Lattes.

En lo que se refiere a las muestras de los dos primeros censos en particular, resulta muy evidente que Germani impulsaba el aprovechamiento de las condiciones favorables del momento, tales como: a) las cédulas censales estaban disponibles y aparentemente completas, en el Archivo General de la Nación; b) se contaba con avances en aplicaciones de la teoría del muestreo, y c) se disponía de nuevos desarrollos en computadoras (electrónicas) y en máquinas tabuladoras (eléctricas). Por todo esto Germani promueve el acuerdo entre el CELADE y el Di Tella, procura recursos financieros externos, acerca varios investigadores afines y, en general, genera las condiciones institucionales que le permitirían hacer las cosas que, diez años antes hubieran sido una quimera, pero que en ese momento, si bien trabajosas, eran posibles[12].

La realización de las muestras

Obtenido el financiamiento del Population Council, Somoza y Lattes, emprenden en junio de 1966 las tareas de las muestras censales. Aunque parezca paradójico, no detallamos aquí las cuestiones técnicas de cómo se realizaron las muestras dado que esto puede leerse en el texto original de Somoza y Lattes (Somoza y Lattes, 1967). Simplemente referimos que se realizaron dos muestras (una por cada censo) basadas en un muestreo

[11] *En efecto existe evidencia de que junto con Malvina Segre, Germani había diseñado y ejecutado un proyecto de investigación que incluía una muestra sobre el censo de 1960 (Di Tella, 1968a).*

[12] *Para una visión de Germani en donde se destacan sus cualidades oraganizacionales puede consultarse (Pereyra, 2010).*

aleatorio sistemático de las hojas de las cédulas censales con poco más de 100.000 casos en cada una[13].

Todas las actividades del trabajo que incluyen la evaluación de los materiales originales, selección de las muestras, verificación y transcripción de los datos a grandes planillas, confección de los manuales de códigos, perforación y verificación de las tarjetas y tabulaciones, fue realizado en 16 meses. Teniendo en cuenta la cantidad de personas involucradas (42) y el presupuesto disponible (U$S 21.000), el conjunto de la investigación resulta todo un record de eficiencia (McCaa, Haines, and Mulhare, 2001)[14].

Como nota de color cabe recordar que en ese momento los investigadores no poseían computadora electrónica o por lo menos un tabulador eléctrico. De esta manera lo usual era construir una "base de datos" en tarjetas perforadas (con sus procesos de codificación, perforación y verificación) y luego acceder a alguna gran computadora para su procesamiento. En este caso, Somoza y Lattes accedieron a una computadora electrónica en horario nocturno perteneciente a la compañía de seguros "La Continental" (fruto de los contactos del actuario Somoza) razón por la cual debían pensar con antelación y precisión cuales de todas las "salidas" posibles iban efectivamente a realizar[15]. Sólo "cargar" la base de

[13] *También se realizó una tercera muestra de más 21.000 casos (sobre una población de poco menos de 100.000 individuos) de las cédulas del censo de población de Buenos Aires de 1955 (Lattes y Poczter, 1968).*

[14] *En efecto, si uno agrega los casos de la muestra del censo de Buenos Aires el costo unitario por persona (con todas sus variables) sería inferior a 10 centavos de dólar.*

[15] *El sendero del uso de las computadoras electrónicas en las ciencias sociales de Argentina ya había sido iniciado años antes precisamente por uno de los fundadores del Instituto Di Tella. Al principio de la década de los 60' Torcuato Di Tella había tenido oportunidad de trabajar con "Clementina" la computadora electrónica ubicada en la Facultad de Ciencias Exactas de la UBA. En concordancia con esa experiencia era también interés del Di Tella, crear un banco de datos censales de Latinoamérica (García Bouza, 1967)(Cornblit & Mora y Araujo, 1967)(Di Tella, 1968a). Además, en esa época el CELADE ya contaba con su propio banco de datos (Morales, 1972).*

datos (tarjetas perforadas) en la computadora implicaba un trabajo de muchas horas[16].

Quizá sea ocasión de incluir un paréntesis metodológico. La realización de una muestra aleatoria sobre un registro de cédulas censales en principio otorgaría un tipo de "dato" similar si la muestra se ejecutaría el día posterior al censo, un siglo después o en la actualidad (siempre que los registros sigan existiendo y mantengan cierto orden)[17]. Lo que sí cambiaría serían las *técnicas de análisis* disponibles. En este último sentido la posibilidad de que en la actualidad se pueda tener acceso a los microdatos de las muestras abre nuevas alternativas de análisis que no eran posibles ni en la época de las publicaciones censales ni en la época de la realización de las muestras.

Muchos de los datos censales tanto en Argentina como en el resto del mundo no tienen la propiedad de ser *microdatos*[18] cuando se difunden, aún en el caso de los

---

[16] *Aage Sørensen en un trabajo crítico de la sociología empírica multivariada deja entender que los avances de las computadoras en velocidad de procesamiento, cantidad de variables y técnicas de análisis ha sobrepasado en mucho la capacidad teórica de los investigadores para establecer los mecanismos inobservables que relacionan las variables (Sørensen, 1998). Otras críticas similares, véanse en (Boudon, 1976)(Boudon, 1998)(Sørensen, 2009).*

[17] *Según nos refiriera Alfredo E. Lattes, personal del Archivo General, incentivados por la posibilidad de continuar trabajando, fuera de horario, en la extracción de muestras censales, organizaron una exhaustiva búsqueda de las cédulas del censo de 1914. El resultado fue el descubrimiento de unos pocos restos de las mismas que habían sobrevivido a las inundaciones que afectaron a los depósitos en que se encontraban. Pero la búsqueda tuvo un premio, aunque menor, porque se encontraron los cajones que contenían las cédulas del censo de la Ciudad de Buenos Aires de 1855, del que luego se extraería una muestra (Lattes y Poczter, 1969). Muchos años después Massé (1993) digitalizó todas las cédulas del censo de 1855 con apoyo del CELADE, sin embargo, esa base de microdatos no parece estar disponible para la investigación.*

[18] *La categoría de microdatos toma su importancia en el proceso de difusión y publicación ya que por definición los investigadores que trabajan con fuentes "primarias" suelen o por lo menos tienen la posibilidad de trabajar con microdatos. En este sentido sus datos publicados o difundidos son la agregación individual de todas sus unidades de análisis (al nivel de observación determinado por el diseño de investigación). Esto cambia para aquellos que trabajen con los datos publicados o puede cambiar para aquellos que trabajen con alguna base de datos difundida. Complementando lo dicho en la nota al pie n° 8 la difusión de una base de datos de "microdatos" reduce mucho la distancia, en términos de posibilidades de análisis, entre las fuentes "primarias" y las "secundarias".*

difundidos por medios digitales. En el caso de Argentina, por ejemplo, INDEC elaboró una base de datos sobre el censo de población del 2001 que, manteniendo la cobertura de toda la población, posibilita análisis de una gran desagregación geográfica (hasta radio censal). Esto fue un gran avance para una gran cantidad de investigaciones ya que redujo la distancia entre las posibilidades de análisis de trabajar con fuentes *primarias* y *secundarias*. Por ejemplo esos datos posibilitaban tabulados "a medida" para calcular diseños muestrales como también para tomar mejores decisiones a nivel departamental y municipal. Por otro lado tomando la menor unidad geográfica (el radio o segmento censal) y con ayuda de algunas técnicas estadísticas es posible asumir algunas hipótesis a nivel de los individuos a pesar de no contar con los microdatos[19].

Por otro lado, para la Argentina existen muestras de microdatos de los censos de 1970, 1980, 1991, 2001 y 2010 que, si bien con las limitaciones propias de las muestras, son aptas para realizar análisis al nivel de las personas. Cabe destacar que estas muestras se encuentran disponibles en el sitio del IPUMS.

Sin duda alguna para Germani la realización de las muestras censales respondía a varias razones, además de ser una contribución a la investigación social; básicamente, podría profundizar varios análisis ya realizados desde los tabulados de las publicaciones originales y desarrollar nuevos análisis. Si bien no encontramos evidencias de que tuviera un interés teórico en poner a prueba nuevas hipótesis para lo cual se requirieran microdatos, cabe recordar que en esa época

[19] *Si bien es verdad que puede existir una "falacia ecológica" en donde en base a datos agregados imputemos propiedades a los individuos (Robinson, 1950, Boudon 1963) también es verdad que en la actualidad existen procedimientos de análisis, que dentro de determinadas opciones, permiten realizar lo anterior sin caer en razonamientos ni falaces ni ecológicos . Dentro de las alternativas más conocidas puede citarse la famosa propuesta de Gary King (King, Rosen, & Tanner, 2004).*

no eran muy comunes los análisis basados en regresiones efectuadas sobre este tipo de datos[20].

En relación a lo anterior se señala que los microdatos brindan la posibilidad de diseñar nuevas investigaciones y de revisitar viejas hipótesis, desde ángulos diferentes y con técnicas estadísticas más flexibles y potentes.

Algunos eslabones de la larga cadena[21]

Antes de entrar en este punto parece apropiado hacer explícitas las definiciones que se han utilizado en relación al concepto de *dato*, aunque en ocasiones se hace uso del léxico propio de la biología evolutiva por considerarlo intuitivo.

Se entiende por *dato* la información contenida en algún *soporte*. *Dato digital* es un tipo de *dato* que tiene como característica algún *formato* discreto.

Por *formato* se entiende la forma en que se usan los bits para codificar un *dato digital*. En este artículo, principalmente importan los *formatos* de los *datos digitales* de *texto plano* aunque, en la última parte, también se comentan algunos *formatos* específicos para bases de datos, más complejos que los de *texto plano*. En general los sistemas de codificación han crecido, por un lado, en la cantidad de caracteres que permiten codificar gracias al aumento de los bits y por otro han aumentado la variedad de cómo se usan esos bits. Esto último es

[20] *De esa época es el famoso libro de Blau y Duncan "The American Occupational Structure" (Blau y Duncan, 1967). En él los autores hacen distintos análisis de regresión como el "path analysis" sobre microdatos censales analizados con tarjetas perforadas. Arthur Stinchcombe en alguna ocasión ha dicho que ese libro es el "trabajo destructivo más brillante de la historia de la sociología" (Stinchcombe, 1978).*

[21] *Esta sección presenta diferencias históricas con lo escrito en una versión anterior (Quartulli, 2013) gracias a los comentarios recibidos por Alfredo Lattes.*

notorio en el caso de las computadoras personales y en el auge de la industria del software.

Por *soporte* se entiende el sustrato físico en donde se almacenan los *datos*. En el pasado el *acceso* a los *datos* de distintos *soportes* solía incluir el uso de periféricos específicos de escasa *difusión* social y rápida obsolescencia tecnológica.

Por *archivo* (digital) se entiende a la unión de determinado *dato* con su respectivo *formato*. Si bien los *archivos* al igual que los *datos* pueden conceptualizarse como constructos (en efecto se pueden patentar), en la vida real siempre se encuentran en algún tipo de *soporte*.

Por *copia* al proceso de replicación de los *datos* aunque varíe el *formato* y/o *soporte*. En otras palabras, la *copia* es un evento clave en el proceso de *difusión* de los datos. Internet ha permitido que se puedan hacer *copias* a pesar de las diferencias espaciales.

Por *accesibilidad* a la condición de acceder a los *datos* grabados en algún *soporte*. Es una disposición de los *datos* que cambia con el tiempo por cuestiones exógenas (emergencia, difusión y extinción de la tecnología específica de cada *soporte*) como por cuestiones endógenas (los *datos* se extinguen por la propia degradación de su *soporte*). Internet ha permitido que se pueda *acceder* a *datos* grabados en servidores web a pesar de las diferencias espaciales.

Por *difusión*, al proceso social de propagación de los *datos*. De modo más directo por un mayor *acceso* a los *datos* (en algún *soporte* y *formato*) o por una mayor replicación de los mismos (*copias*). De modo más indirecto por una mayor propagación de la tecnología que permite operar sobre algún *soporte o formato*. Internet es una herramienta privilegiada para la difusión de *datos* sea tanto de forma directa (acceso y copia) como indirecta debido a su rápida expansión y su plausible continuidad en el tiempo.

Por *actualización* se entiende el cambio de *soporte* en el proceso de replicación de los *datos*. Antes el proceso de *actualización* era más costoso y trabajoso debido a la escasa difusión y rápida obsolescencia tecnológica de los periféricos.

Por *traducción* se entiende el cambio de *formato* en el proceso de replicación de los *datos*. Generalmente esto sucede desde un *formato* más antiguo pero más compatible a uno más moderno y funcional pero menos compatible. Este último punto al reducir la compatibilidad se convierte en ambiente-dependiente, en tanto que ante un cambio brusco de tecnología se corre el riesgo de la *extinción* de los *datos*.

Por *mutación* al proceso que durante una *copia* o la simple conservación, los *datos* cambian pero mantienen su *accesibilidad* con las respectivas tecnologías que operan con los *soportes o formatos*. En general los *datos digitales* presentan muchas menos *mutaciones* en sus *copias* que los datos analógicos.

Finalmente la *extinción* es el evento que produce el deceso de los *datos* (originales) sea por *mutación* de los mismos o por la propia degradación del *soporte*. La falta de *accesibilidad* por cuestiones exógenas (disponibilidad de la tecnología necesaria) podría ser considerada un caso de *extinción* (potencial) de los *datos* ya que si ellos son inaccesibles en el corto o largo plazo se *extinguen* por la propia degradación de su *soporte*.

Retornando a la historia que nos ocupa, entra en escena Robert McCaa, muy conocido por ser el incansable promotor del IPUMS. Lo que quizá pocos saben es que la idea básica del IPUMS, esto es, la concreción de acuerdos con entidades gubernamentales para la realización y difusión de muestras censales de microdatos se inspira, entre

otros antecedentes, en este trabajo fundacional de Somoza y Lattes[22].

La relación que se establece entre Lattes y McCaa posibilita varios intercambios y así, en los 80´ Lattes le envía una *copia* de las muestras en *formato* ASCII, grabada en computadora personal (PC) con ayuda de la tecnología "Bernoulli" que utilizaba un *soporte* similar a un disquete de 3,5 pulgadas.

Con el advenimiento de Internet Robert McCaa decide *actualizar* esos *archivos* en alguno de los *soportes* compatibles con un servidor web[23] (por ejemplo un disco rígido) y así favorece su *difusión y posibilita su accesibilidad desde cualquier* localización geográfica. Básicamente, fueron esos los *datos* que permitieron a este autor, generar el producto ofrecido en esta página web. En todas las *actualizaciones* y *traducciones* que se hicieron, prácticamente, no hubo pérdida o cambio de la información de los *datos* originales; en otras palabras, las *copias* de los *datos* no generaron *mutaciones* de los mismos.

La reconstrucción de esta historia puede ser vista como muy lineal, sin embargo esa linealidad sólo es la punta de un iceberg. Por ejemplo, no se explica por qué los *datos* no tuvieron mayor *difusión* y tampoco se explica porque los *datos* continuaron durante largo tiempo en *formatos* anticuados.

Es importante destacar en relación a la cuestión de la *conservación* y *difusión* de estos *datos* académicos, que en su momento se realizaron varias *actualizaciones* y

---

[22] *Esto se refleja en el artículo escrito por McCaa y otros, titulado "The first national Historical Census Microdata" en donde elogia el trabajo de Somoza y Lattes (McCaa, Haines, y Mulhare, 2001). Del mismo modo, en otro artículo también refiere la experiencia de la OMUECE (Operación muestras de Censos) como fuente de inspiración para IPUMS (McCaa y Jaspers Faijer, 2000).*

[23] *Si entrar en precisiones técnicas, un servidor web es cualquier soporte en donde se puedan copiar datos con la particularidad de que ese soporte se encuentra conectado a otras computadoras mediante el protocolo HTTP. En este sentido el servidor web posee su especificidad en el modo en que se conecta con otras computadoras y no en el tipo de soporte.*

traducciones. Efectivamente, una vez completadas las muestras y siguiendo los principios que las habían hecho posible, Alfredo E. Lattes trató de difundir estas bases de *datos* y de asegurar la *accesibilidad* a las mismas, antes de su anunciada *extinción*. Para ello distribuyó los *datos* en *soporte* de cinta magnética entre varias instituciones del país y del exterior y, también, generó *actualizaciones* y *traducciones* de las muestras; lo mismo hicieron Somoza, en el CELADE, y Robert McCaa.

Cabe aquí una reflexión; una serie de eslabones que se enganchan forman *una* cadena y para mantenerse como tal deben continuar unidos. Para conservar el *acceso* a estos datos se construyeron varios eslabones secuenciados (*traducciones*, *actualizaciones*, etc.) que podrían ser vistos como una cadena, sin embargo, en la práctica esto no ocurrió, excepto, en la única secuencia que consiguió mantenerse y que fue la que vinculó a Alfredo E. Lattes con Robert McCaa.

Es evidente que algunas de las *actualizaciones* que se hicieron trataran de ganar tiempo hasta que una nueva tecnología facilitara la *difusión, aunque t*ambién es posible que algunas de ellas se realizaran principalmente pensando en la *difusión* misma. De algunas de las *actualizaciones* realizadas se llegó a realizar fueron *copias,* sin embargo, casi todas se *extinguieron* antes de volver a replicarse.

El tema es que la tecnología (el otro actor clave de esta historia) evoluciona rápidamente y si bien su avance expande las posibilidades, para aprovecharla hay que disponer de las herramientas necesarias y del saber para reciclar viejas opciones. Quizá sirva, como ejemplo, esta analogía con el mundo musical.

El fonógrafo (*soporte*) *accedía* a la información musical desde un cilindro en cambio el gramófono (otro soporte) *accedía* a la información musical del primer disco plano de 78 revoluciones por minuto (RPM). Posteriormente los reproductores de discos (otros soportes) redujeron la

velocidad a la que podían *acceder* a sólo 33 RPM, implicando que los discos de 78 RPM dejaran de poder usarse en los nuevos reproductores. La misma suerte, en términos de incompatibilidad, tuvieron los de 33 RPM debido a la aparición de los reproductores de casete que, a su turno, cedieron el paso a los reproductores de CD y los DVD. Además, en los últimos dos casos la información musical dejó de tener un formato analógico para pasar a un *formato digital*. El problema es que cada una de estas tecnologías no es compatible, en el sentido de que no se puede *acceder* a la información musical de un disco desde un reproductor de CD ni viceversa. Lo que quizá podría hacerse, con más o menos trabajo según el caso, es *actualizar los soportes* donde se encuentra la información musical desde el cilindro del fonógrafo, a un disco de 78 RPM, de este a uno de 33 RPM, de este a un casete y de este a un CD y luego a un DVD. En el paso del casete al CD también habría un proceso de *traducción* del *formato,* desde el analógico al digital.

Es importante remarcar que los *datos* originales tienen más chance de sobrevivir si las *actualizaciones* se realizan a medida que *emergen* y *difunden* las respectivas tecnologías de cada *soporte*. Para precisar, si alguien quisiera hoy recuperar la información grabada para un fonógrafo tendría que tener acceso a todas las tecnologías utilizadas para hacer todas las *actualizaciones* intermedias, lo que implica tener a disposición desde un gramófono hasta un reproductor de CD, o construir desde cero un artefacto que permite la grabación desde un fonógrafo a un reproductor de CD.

Volviendo a nuestra historia, en 1967 podríamos llamar "base de datos" a las más de 200.000 tarjetas perforadas IBM80 (una por persona)[24]. Esto, aunque parezca muy

[24] *Suponiendo que en una caja típica de zapatos (30x18x12cm) entrarían unas 1500 tarjetas perforadas (18x8cm), para movilizar la "base de datos" se necesitaría mover físicamente más de 130 cajas de ese tamaño. En efecto, como se refiere más adelante, Lattes en la década del 70´ (debido a un proyecto de investigación junto a Raúl Nordio) intentó recuperar la base de datos desde*

lejano, fue un primer paso hacia el *dato digital*. En efecto los proceso de *codificación* y posterior *edición*, usuales en la jerga de las encuestas, no son otra cosa que una *interpretación* (y posterior control de ella) desde un lenguaje propio del mundo social (en el *soporte* que sea por ejemplo "papel") hacia un *dato digital* que puede ser interpretado por una computadora[25].

Si uno se abstrae del *soporte* en donde se graban los *datos* (digitales o no) quizá sea más fácil de entender que se quiere decir con *dato digital*[26]. En los tiempos que corren es posible que muchas personas supongan que los *datos digitales* tienen que ver con la computación. Si bien esto no es incorrecto cabe explicitar un poco más la idea de *digital* con ejemplos que no tengan que ver con la historia de la computación[27].

Lo importante del concepto *dato digital* es lo discreto del *dato*. En efecto antes de que surjan las "sorters" eléctricas como la de Hollerith (1880) o las más modernas computadoras electrónicas (después de 1940) existían artefactos que también exigían una digitalización previa de la información. Por ejemplo era usual en la industria textil "digitalizar" la información que explicita la forma en que se intercalan los hilos de cada tela. En una operación llamada *mise en carte* se realizaba una cuadriculación extrema de la tela mediante una previa ampliación del urdimbre y la trama original para

---

las mismas tarjetas perforadas en la Universidad Nacional de Córdoba, por lo que tuvo que transportarlas en un camión alquilado para ese fin.

[25] *De todos modos, se puede hacer un trabajo de codificación con la simple intensión de facilitar un análisis posterior y este último puede hacerse "a mano" sin la necesidad de ninguna computadora. Del mismo modo que se pueden codificar un conjunto de palabras (sentencia) en una sola palabra y no necesariamente en un numeral o en un caracter. Lo importante es que tanto el numeral (o carácter) como las palabras suelen ser símbolos (términos) que designan conceptos (predicados)(Kneale, 1972)(M. Bunge, 1974).*

[26] *Un ejemplo típico de un dato no digital es un dibujo a mano alzada sobre un papel. El dibujo en sí mismo sería el dato y el papel junto con la tinta sería el soporte.*

[27] *La relación entre computadora y dato digital es necesaria en tanto la primera es un artefacto que exige información discreta (+ voltaje – voltaje) y los datos digitales pueden cumplir este requisito. Lo inverso, o sea que los datos digitales se lean mediante computadoras es contingente.*

asignarle un valor (perforado/no perforado) a cada cuadrado. Luego estos *datos digitales* se volcaban a una tarjeta perforada (*soporte*) que luego se agregaban al telar (junto con los hilos por otro lado) y este interpretaba de forma mecánica esos datos como instrucciones para intercalar los hilos entre el urdimbre y la trama[28] (Essinger, 2007).

En el caso de las muestras censales, como era usual para la época, ese proceso de *interpretación* y *codificación* se realizó con el *soporte* estándar de esa época que eran las tarjetas perforadas IBM80[29]. Un soporte que parece haber influenciado la manera de conceptualizar en la metodología de las ciencias sociales de la época. En efecto, el concepto de matriz de datos difundido por Galtung (Galtung, [1966] 1973) parece tener un aire de familia con las tarjetas perforadas y, especialmente, con el modelo IBM80[30]. El mismo Galtung afirma:

- "Puesto que el enfoque de este libro es principalmente nomotético, puede valer la pena dar más sustancia a estas ideas, destacando lo que significarían en términos de tarjetas IBM" (1973, p. 20)
- "Esta es exactamente la forma en que aparecen los datos si hay una tarjeta perforada para cada unidad, una columna para cada variable y una perforación para cada valor, y se pasan las

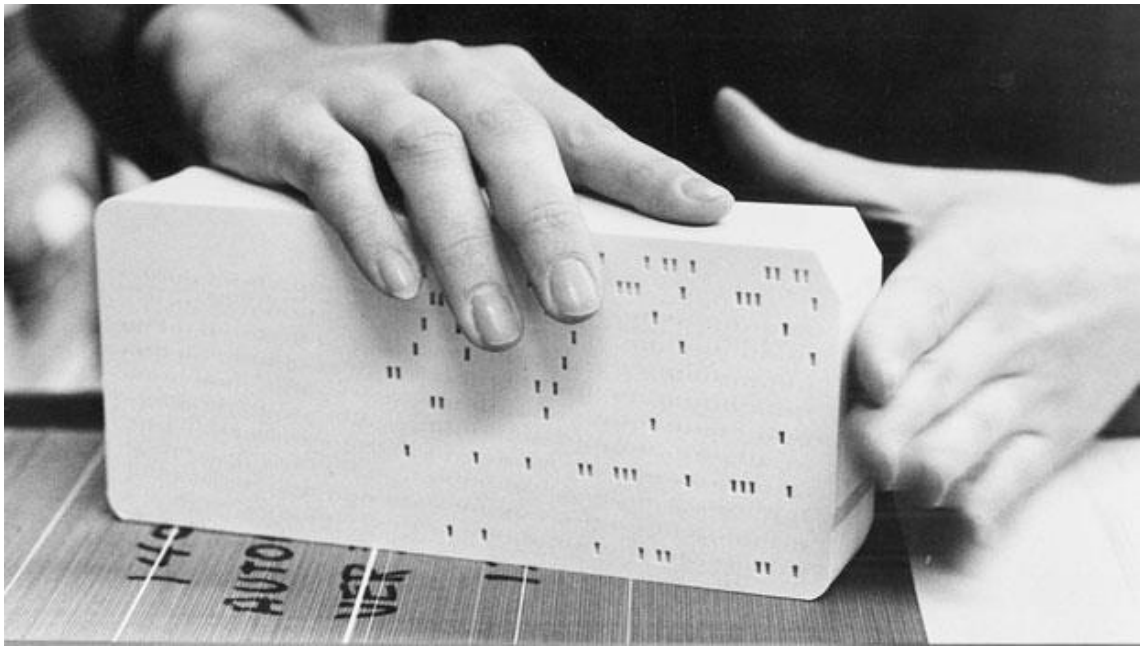
[28] *Específicamente si había un agujero en la tarjeta perforada, por él pasaba un gancho "bolus" que hilaba la urdiembre con la trama. En caso contrario el gancho se detenía y no se producía la hilada en ese "cuadrado" específico.*

[29] *La tarjeta perforada IBM80 (80 columnas), diseñada en 1928 (Strickland, 2012), a pesar de ser sólo un tipo de tarjeta perforada específica de una empresa (IBM) llegó a identificarse como sinónimo de la tarjeta perforada (la original de Hollerith era de sólo 45 columnas). Para consultas acerca del diseño y posibilidades de estas tarjetas puede consultarse (IBM, 1961)(IBM, 1970)(IBM, 1971).*

[30] *La influencia de las tarjetas IBM80 también parece haber llegado a través del diseño de los primeros programas para el análisis de datos sociales. En efecto, a pesar de no nombrar ni a Galtung ni a al concepto de matriz de datos en el libro fundacional de SPSS los autores explicitan porque diseñan como diseñan el programa siguiendo ideas muy parecidas a las de Galtung(Nie, Bent, & Hull, 1970).*

> tarjetas por una máquina que registra para cada tarjeta lo que está perforado en ella" (Galtung, 1973, p. 4)

Imagen 1. Tarjetas IBM 80.



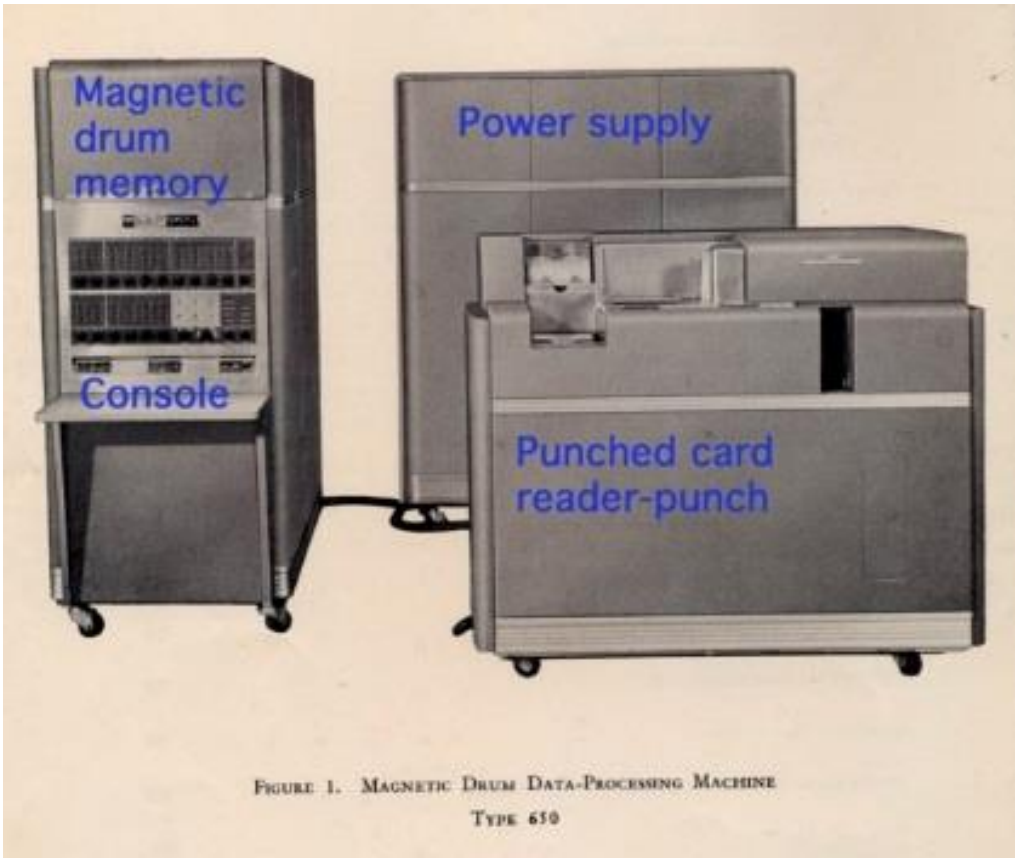
Por ese tiempo las computadoras *accedían* a los *datos* contenidos en ese *soporte* mediante un periférico específico que hacía pasar cada una de las tarjetas por una especie de tambor magnético ("magnetic drum") que era el *soporte* en donde los *datos* se almacenaban temporariamente en la memoria de la máquina. La codificación usual de esos *datos* era BCD (binary code decimal) de 6 bits que ofrecía la posibilidad de codificar 64 símbolos diferentes distribuidos entre numerales, caracteres (sólo mayúsculas) y símbolos especiales[31].

Una computadora usual de esa época, al menos en países como Argentina, era el main frame IBM 650 con algún periférico lector de tarjetas perforadas como el IBM 533

[31] Esta cantidad de símbolos era suficientes para programar en los lenguajes de la época como Fortran o Cobol y de allí también la norma que aconsejaba escribir los programas en letras mayúsculas.

o IBM 537 (IBM, 1955)[32]. Para fijar las ideas, auxiliada con una IBM 533, una IBM 650 podía acceder a 200 tarjetas perforadas por minuto (IBM, 1956, p. 6) lo que supone que leer cada muestra censal (100.000 tarjetas) demandaba un mínimo de 8 horas.

Imagen 2. IBM 650



Aparentemente en el momento de la realización de las primeras tabulaciones de las muestras (1967) en la computadora electrónica de la compañía de seguros "La Continental" la base de datos ya estaba *actualizada* en

[32] *Para este período las máquinas electrónicas exclusivamente estadísticas ("sorters") (IBM, 1958) derivación lejana de la máquina tabuladora de Hollerith si bien también leían las tarjetas perforadas ya estaban en descenso frente a las más versátiles computadoras electrónicas (como por ejemplo la citada IBM 650). Estas últimas requerían de una programación específica hecha por un programador para su uso. Las "sorters" no necesitaban un programador, pero sólo seleccionaban y contaban. En general, aunque no necesariamente, sólo se utilizaban para realizar tabulaciones simples de grandes cantidades de datos. Para un ejemplo de cómo calcular una "escala" con una sorter y sus tarjetas IBM80 puede consultarse* (Ford, 1950).

un *soporte* más moderno conocido como cinta magnética ("magnetic tape")[33]. Esto implica que la computadora utilizada poseía un periférico (ver imagen 3) que hizo posible *actualizar* los *datos* desde el *soporte* de las tarjetas perforadas al *soporte* de cinta magnética que, en ambos casos, hacía las veces de memoria externa de la computadora[34]. El uso de las cintas magnéticas como *soporte* para la introducción de *datos* eran una evolución de la original UNIVAC de la Remington Rand (Remington Rand, 1957)[35].

Imagen 3. IBM 727. Unidad de Cinta Magnética de 7 pistas

[33] *Un soporte intermedio de esa época era la cinta perforada. Esta ocupaba menos volumen y permitía acceder más rápido a sus datos en comparación con las tarjetas perforadas aunque en forma más lenta que con una cinta magnética. En comparación con estas últimas su costo era bastante menor aunque algo mayor que las tarjetas perforadas.*

[34] *En lo que sigue cuando nos referiremos a las cintas magnéticas serán a cintas magnéticas de datos y no de audio y vídeo que ya tenían una trayectoria anterior y por la época sufrieron grandes transformaciones.*

[35] *Como nota de color se destaca que la publicidad más recordada de la UNIVAC fue la realizada por la CBS en donde con una muestra (aleatoria) de sólo 1% de la población predijo en 1952 la victoria de Eisenhower (cuando no era favorito) basado en cómputos efectuados por la UNIVAC. Esto también muestra la co-evolución de la escuela de la encuesta ("survey"), de la muestra ("sample") junto a la emergencia y difusión de las primeras computadoras.*

Estas cintas magnéticas, generalmente de siete pistas[36] para la época eran algo difíciles de editar aunque eran casi plenamente regrabables[37]. Tenían la ventaja que ocupaban un volumen transportable para una persona y aceleraban los tiempos de futuros cálculos sobre esa base de datos[38]. El primer dato, acerca de su volumen, es sumamente importante para la *difusión* de los *datos*, ya que por primera vez era posible hacer una *copia* de los *archivos* y posteriormente entregársela en mano a otro investigador.

Con el instrumental adecuado, por esa época se podría *traducir* a un ritmo de 250 tarjetas por minuto[39]. Esto implica que para *traducir* las "bases" de la muestra desde el conjunto de las tarjetas perforadas hacia las cintas magnéticas de 7 pistas se requería más de 13 horas.

Es importante remarcar que las cintas magnéticas, a pesar de la reducción de tamaño, también fueron un problema para las instituciones que poseían una "cintoteca" que era el lugar en donde ellas se almacenaban. Entre otros problemas se tenía el de cómo guardar las cintas, una al lado de otra, sin que se magnetizaran entre sí[40]. Otro de los problemas

---

[36] *La cantidad de pistas de una cinta magnética tiene que ver con la cantidad de bits que puede contener la unidad de información que en el caso de los datos de textos es el carácter, el numeral y algunos signos especiales. IBM en 1952 lanza para su línea de productos la tecnología de 7 pistas (6+1 bits) y recién en 1965 con la IBM S/360 introduce la familia de arquitectura de 9 pistas (8+1 bits) manteniendo el soporte de las cintas magnéticas. Es en este momento histórico que se asume, convencionalmente, que un byte es igual a 8 bits.*

[37] *Las tarjetas perforadas tampoco lo eran en forma individual, pero las "bases de datos" construidas con miles de ellas sí en el sentido que se podía sacar la tarjeta específica que estuviera incorrecta y reponer por una nueva en el estado correcto.*

[38] *Al menos aceleraba considerablemente los tiempos de la época. Luego esta misma tecnología presentó un obstáculo fuerte en que los datos se leían en el orden en que fueron grabados y no en forma azarosa (RAM = Random Access Memory). Sirve de ejemplo pensar en la forma que se "adelantaban" los casetes de música para escuchar una canción específica.*

[39] *Por ejemplo desde una lectora de tarjeta IBM 714 o IBM 759 hacia una reproductora de cinta magnética como una IBM 727 (IBM, 1962). Otras reproductoras de cintas de 7 pistas eran las IBM 726, IBM 728, IBM 729 y IBM 7330.*

[40] *En las cintas magnéticas de audio este efecto se conocía como "efecto copia" o "efecto eco". Para evitarlo se debían guardar "de cola". Esto implicaba*

fundamentales de este tipo de almacenamiento era su precaria conservación en el tiempo. En efecto, los *datos* guardados en una cinta magnética tenían expectativa de vida variable que, según la temperatura y la humedad relativa, podía variar desde 2 hasta más de 60 años (Van Bogart, 1995, p. 35).

Aparentemente las *actualizaciones* realizadas desde las tarjetas perforadas hacia la cinta magnética no duraban lo esperado y de tanto en tanto, era necesario la realización de una *copia* antes que los *datos* quedaran *inaccesibles* debido a la degradación del *soporte*.

A principios de los 70', se realizó otra *actualización y traducción* desde la cinta magnética de 7 pistas a una de 9 pistas que hasta donde se sabe también incluyó los *datos* sobre la muestra del primer censo de la ciudad de Buenos Aires (1855). Esta operación parece haberse realizado en las instalaciones del CELADE en Chile, bajo la coordinación de Jorge Somoza, que por ese entonces, era investigador de esa institución. En esta operación se agregaron los *datos* de la muestra del censo de la Ciudad de Buenos Aires de 1855[41].

De esta operación, como en parte se puede observar en la imagen 4, se realizaron copias que fueron entregadas tanto al Instituto Di Tella como al INDEC. Este último organismo se incluyó fruto de la conexión entre Zulma Rechini, que trabajaba como asesora del organismo y el director del mismo, Carlos Noriega[42]. En la actualidad es probable o bien la falta de *accesibilidad* o bien la *extinción* de esos datos.

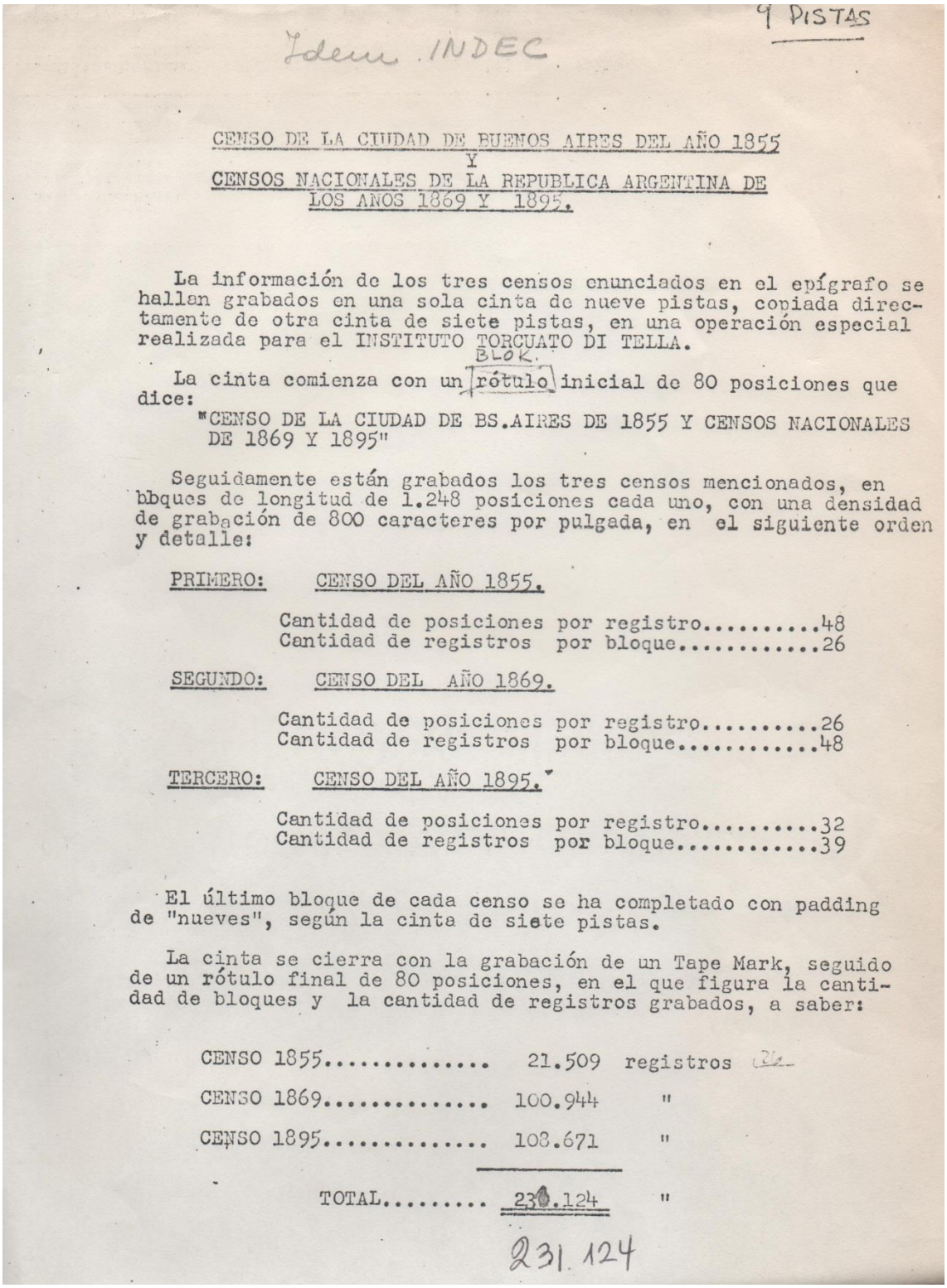
---

que las grabaciones dejaran una última zona "virgen" en la cinta y que luego, al almacenarse, todas las cintas quedaran con esa "cola" al aire.

[41] *Como se anticipó anteriormente el trabajo inicial realizado por Lattes y Somoza (Somoza y Lattes, 1967) tuvo su continuación en un trabajo hecho por Lattes y Poczter para la ciudad de Buenos Aires (Lattes y Poczter, 1968). Cuando se procesaron los datos de las muestras en 1967, no existían aún los datos digitales del censo de 1855 de Buenos Aires. Estos se actualizaron desde las tarjetas perforadas hacia una cinta magnética con una diferencia de algunos meses.*

[42] *Noriega fue director del INDEC desde 1973 hasta que en situaciones extrañas pasó a formar parte de la lista de desaparecidos en el verano de 1977.*

Imagen 4. Constancia de Traducción de 7 a 9 pistas



9 PISTAS

Idem INDEC

CENSO DE LA CIUDAD DE BUENOS AIRES DEL AÑO 1855
Y
CENSOS NACIONALES DE LA REPUBLICA ARGENTINA DE
LOS AÑOS 1869 Y 1895.

La información de los tres censos enunciados en el epígrafo se hallan grabados en una sola cinta de nueve pistas, copiada directamente de otra cinta de siete pistas, en una operación especial realizada para el INSTITUTO TORCUATO DI TELLA.

La cinta comienza con un BLOK rótulo inicial de 80 posiciones que dice:

"CENSO DE LA CIUDAD DE BS.AIRES DE 1855 Y CENSOS NACIONALES DE 1869 Y 1895"

Seguidamente están grabados los tres censos mencionados, en bbques de longitud de 1.248 posiciones cada uno, con una densidad de grabación de 800 caracteres por pulgada, en el siguiente orden y detalle:

PRIMERO: CENSO DEL AÑO 1855.

Cantidad de posiciones por registro.....48
Cantidad de registros por bloque.....26

SEGUNDO: CENSO DEL AÑO 1869.

Cantidad de posiciones por registro.....26
Cantidad de registros por bloque.....48

TERCERO: CENSO DEL AÑO 1895.

Cantidad de posiciones por registro.....32
Cantidad de registros por bloque.....39

El último bloque de cada censo se ha completado con padding de "nueves", según la cinta de siete pistas.

La cinta se cierra con la grabación de un Tape Mark, seguido de un rótulo final de 80 posiciones, en el que figura la cantidad de bloques y la cantidad de registros grabados, a saber:

CENSO 1855..... 21.509 registros
CENSO 1869..... 100.944 "
CENSO 1895..... 108.671 "

TOTAL..... 230.124 "

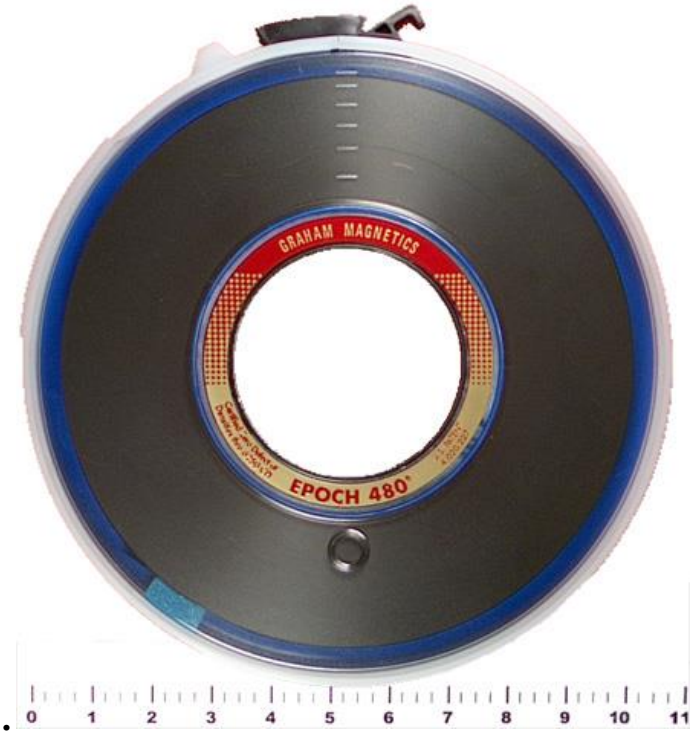
231.124

Casi con seguridad ambas cintas (la de 7 y la de 9 pistas) fueran las clásicas cintas magnéticas de 10.5 pulgadas de carretel o bobina y ½ pulgadas de espesor (*ver imagen 5*) usadas en los periféricos de los mainframe

de la época como por ejemplo la famosa IBM S/360[43]. La cinta magnética de 7 pistas utilizada como insumo en la operación posiblemente sea la utilizada en las computadoras de La Continental en 1967 y 1968.

Cabe destacar que las densidades comunes en las cintas de 7 pistas comenzaron en 100, ascendieron a 200, luego pasaron a 556 y finalmente llegaron a los 800 caracteres por pulgada. En cambio las cintas de 9 pistas tenían densidades que comenzaron en los 800, pasaron por los 1600, y llegaron a los 6250 caracteres por pulgada[44]. Como se aclara en la imagen 5 la *traducción* se realizó en 800 caracteres por pulgadas que era la densidad más baja de las cintas magnéticas de 9 pistas.

Imagen 5. Cinta Magnética de 9 pistas, carretel de 10.5 pulgadas y ½ pulgada de espesor



[43] *Este era el tipo de mainframe utilizado por el CELADE en su banco de datos a principios de los 70´. Por esa época esta institución utilizaba el mainframe de la Universidad de Chile (Morales, 1972, p. 130).*

[44] *Las densidades de las pulgadas fueron evolucionando gracias a los mecanismos cada vez precisos con los cuales funcionaban los periféricos que se encargaban de la grabación y lectura de los datos. De este modo con una cinta de un igual largo (generalmente 2400 pies o 730 metros aprox.) se podían grabar desde 2,5 megas (con una densidad de 100 c/p) hasta 140 megas (con una densidad de 6250 c/p) en una cinta magnética.*

Como se aclaró anteriormente el cambio de 7 a 9 pistas es un cambio en la cantidad de posibles bits pero no necesariamente produce un cambio en los *datos*. Específicamente las siete pistas provenían de la codificación BCD (Binary coded decimal) que era un *formato* apto para que los mainframe interpretaran los *datos* grabados en las zonas superiores (2) y en los dígitos (10) de las tarjetas perforadas. En el caso de las 9 pistas existían muchos formatos disponibles como EBCDIC, ASCII, CPIO, TAR, etc., aunque los dos primeros eran los más usados en contextos de *datos* con *formato* de texto plano. Estos *formatos* aprovecharon las ventajas de los 9 bits para crear códigos de caracteres de 7 bits (ASCII) y 8 bits (EBCDIC) que ampliaron la cantidad de símbolos que se podían codificar.

De todos modos si bien hubo mejoras apreciables que hacían posible una mejor *difusión* de los *datos* (principalmente gracias a la reducción de volumen) todavía existían problemas de conservación a un mediano plazo y lo que quizá sea fuera más problemático para su futuro era la necesidad de contar con un periférico físico para su efectiva *actualización* a un soporte más moderno. Por otro lado todavía en esa época la *difusión* de las computadoras personales (PC) era escasa (aunque en fuerte crecimiento) y se seguía necesitando de una fuerte ayuda institucional para poder *acceder* a los *datos* aún en el caso en que un investigador tuviera los *archivos* de las bases de datos en los *soportes* usuales de la época.

Quizá sirva como ejemplo de la importancia de contar con el periférico adecuado sirve la historia de las líneas siguientes. Por esos años aparece un eslabón de esta historia que finalmente parece no haber sobrevivido hasta el día de hoy.

Por la segunda mitad de la década de 70’, fruto de un acuerdo de investigación con Raúl Nordio (investigador

de la Universidad Nacional de Córdoba (UNC)), Alfredo E. Lattes envió las tarjetas perforadas originales que aún conservaba, para que se realizara una nueva *actualización*.

Como ya se comentó esto implicó un traslado físico de tal magnitud de tarjetas perforadas que demandó el alquiler de un camión para ese fin[45]. Como para esta época esos mismos *datos* ya existían en *soporte* de cinta magnética una hipótesis plausible es que se intentó volver a un *soporte* más antiguo en función de ciertas restricciones técnicas del instrumental de la UNC.

Es probable que en la UNC hubiera una IBM 1130 (*imagen 6*), que era el mainframe[46] más barato de IBM de esa época. Este modelo ofrecía la posibilidad de agregar periféricos que permitieran *actualizar* desde tarjetas perforadas a cintas magnéticas y se había lanzado al mercado en 1965 (Sweger, 1965)[47]. En función de lo aclarado en el párrafo anterior, es razonable pensar que la UNC carecía de los periféricos necesarios para *acceder* a *datos* en *soporte* de cinta magnética de 7 pistas. Al igual que las *copias* entregadas al INDEC en la actualidad es probable o bien la falta de *accesibilidad* o bien la *extinción* de esos datos.

---

[45] *Alfredo E. Lattes afirma en comunicación personal, que esto se debió a que también viajaron hacia Córdoba las tarjetas perforadas de una muestra del censo de 1960 hecho por la Fundación Bariloche.*

[46] *Un mainframe es una computadora grande, potente y costosa usada principalmente por una gran compañía para el procesamiento de una gran cantidad de datos. Antes del auge de las PC, los mainframe dominaban el mercado e inventaron el sistema de redes entre una computadora central y sus terminales. Luego con la difusión de la tecnología inalámbrica y la fuerte reducción en términos de conexión, los mainframe volvieron a destacarse. Esta vez ya teniendo como "WorkSation" o estaciones de trabajo a un tipo especial de PC (no simples "terminales") y en los casos en donde se precise "servidores" no sólo con mucha memoria sino con una velocidad de procesamiento elevada.*

[47] *Es probable que la máquina en donde se hicieron estas actualizaciones sea una IBM 1130 apodada "La Porota" de la Universidad Nacional de Córdoba de la Facultad de Ciencias Económicas. Esta máquina fue adquirida de segunda mano a la compañía Empresa Provincial de Energía de Córdoba (EPEC)(Arónica & Buraschi, 2011, p. 2).*

Imagen 6. IBM 1130 sin periféricos

Es razonable suponer que para finales de los 70’ y principios de los 80’ las grandes computadoras (en Argentina) fueran descendientes de la arquitectura de la IBM S/360 con la posibilidad de anexar periféricos específicos que *accedían* a *datos* en *soportes* como cintas magnéticas de 9 pistas y carreteles de 10,5 pulgadas o los primeros cartuchos (diskettes o floppy disks) de 8 pulgadas[48]. Estos últimos eran mucho más seguros aunque al principio no ofrecieran beneficios en cuanto a velocidad y capacidad de almacenar información.

Mucho de esta imagen comienza a cambiar con la *emergencia* y *difusión* de las computadoras personales (PC). Especialmente desde el punto de vista tecnológico

[48] *Más tarde llegaron los más conocidos para la mayoría como los diskettes de 5¼ pulgadas y los posteriores 3½. Esto es así porque era inviable la construcción de computadoras personales de tamaño reducido con diskettes de ese volumen.*

es importante el crecimiento del software y como consecuencia de esto los diferentes *formatos* de *datos* para cada tipo de software específico.

De este modo cada *formato* de *datos*, aunque comparta el *soporte*, supone un software específico que codifica (encode) *datos* a bits y decodifica (decode) bits a *datos* de determinada manera[49]. Una diferencia práctica con el pasado es que si bien estos nuevos *formatos* son diferentes y potencialmente incompatibles entre sí, el problema tiene más que ver con cuestiones de software que de hardware en el sentido que para *acceder* y/o *traducir* los *datos* de ahora en más se podrán hacer con el "programa" adecuado y ya no con el "periférico" adecuado.

Esto último, si bien puede parecer algo intrascendente, es pertinente para la problemática de la *conservación* y *difusión* de los *datos*, en este caso académicos. Esto permite que sea mucho más factible y menos costoso encontrar un software anticuado que un periférico de la misma época. Además, en caso que el software requiera un sistema operativo también anterior existen emuladores que pueden cumplir esa función en las pc más modernas.

Volviendo a la historia de los archivos de las muestras censales, una hipótesis compatible con la escasa evidencia es que en la segunda mitad de los 80' se haya realizado otra *actualización* que consistió en un cambio de *soporte* aunque no quede del todo claro si también hubo una transformación en el *formato* de los *datos*. En efecto el cambio supuso pasar de un *soporte* de cinta magnética con *datos* en *formatos* de 9 bits a otro *soporte* también magnético llamado "Bernoulli" que a grandes rasgos podría describirse como un diskette con una tecnología

[49] *El formato del dato puede identificarse muchas veces con la extensión del archivo (txt, doc, xlc, etc.) pero esta relación es contingente. En efecto, existen distintos formatos de datos que comparten una misma extensión de archivo.*

especial que permitía una mayor capacidad de almacenamiento[50].

Esta tecnología fue fundamental porque permitía el paso a otra compatible con una computadora personal (las disqueteras eran compatibles con los puertos de la CPU de una PC) aunque la *actualización* en sí misma es posible que se haya desarrollado en algún sitio institucional debido a que había que tener un periférico que pudiera acceder a los datos de una cinta magnética con un carretel de 10.5 pulgadas lo que suponía un periférico genéricamente asociado a un mainframe.

De este modo la llegada a un *soporte* compatible con un PC otorgó una mayor importancia a las *traducciones* por problemas de *formato* que a las *actualizaciones* por problemas de *soporte*.

Imagen 4. Disquete y Disquetera sistema Bernoulli

Por el lado del *formato* de los *datos* la *traducción* parece haber consistido desde algún *formato* que fuera

50 *La tecnología Bernoulli consistía en un disquete (un plástico fuerte por fuera con un fino y elástico material magnético en su interior) al cual se podía acceder a sus datos mediante una disquetera compatible con una computadora personal mediante una interfaz SCSI (Small Computer System Interface). En un primer momento ofreció almacenaje para 5, 10 y 20 megabytes (Bernoulli I) y en un segundo momento capacidades de 20 hasta 250 megabytes (Bernoulli II). Surgió en 1983 en el mercado informático de la mano de la empresa Omega.*

compatible con 9 pistas (EBCDIC o ASCII) hacía un formato que casi con seguridad sea ASCII (7+1 bits)[51].

ASCII básicamente es un *formato* que permite codificar todos los caracteres de la lengua inglesa aunque dejando fuera otros descendientes de las lenguas latinas. Tampoco incluye símbolos como los musicales y los encontrados en escrituras orientales. Debido a que la totalidad de los datos de las bases de las muestras consisten en números esta *traducción* no trajo problemas de *mutación*.

Por otro lado también es posible que en esta *actualización* los *datos* hayan sufrido otra *traducción* para lograr una reducción de tamaño y que de esa manera se pudieran grabar en ese *soporte*.

Es probable que esta operación haya sida ejecutada a pedido por Alfredo E. Lattes quien años después, aprovechando los beneficios para la *difusión* de un archivo "Bernoulli" le entrega una *copia* a Robert McCaa. De este modo, el receptor de la *copia* sólo tenía que tener una disquetera "Bernoulli" para *acceder* a los *datos* de las bases muestrales y un programa que interprete el formato "zip" para descomprimir (decodificar) los datos y tener acceso a los datos en formato ASCII[52].

Si tenemos en cuenta la fácil conservación por un lado y la factible disposición de los *programas* obsoletos en comparación a los *periféricos* obsoletos la ventaja se torna evidente. Y precisamente McCaa desde un principio parece haber sido consciente de esta diferencia y esta es una de las razones por la cuales el proyecto IPUMS

[51] *Esta suposición se basa en que si bien el formato EBCDIC fue diseñado por la misma IBM, ya en los 80' la misma IBM (y por ende todos sus clones) había adoptado el formato ASCII en sus IBM-PC (las computadoras personales de IBM). Por otro lado grandes proyectos como el OMUECE también utilizaron la codificación ASCII como formato para sus datos (McCaa & Jaspers Faijer, 2000).*

[52] *Hasta la actualidad (2014) sólo se pudo acceder a los datos de las muestras censales de 1869 y 1895. Todavía es una incógnita que ha sucedido con los datos provenientes de la muestra del censo de 1855 de Buenos Aires.*

hubiera sido de una naturaleza muy diferente si se hubiera intentado hacer en otra época[53].

En la actualidad las bases de datos digitales son *archivos* digitales codificados bajo determinados *programas* (*formatos*) que pueden grabarse bajo diferentes *soportes* en una pc. Así también pueden *conservarse* y *difundirse* a través de internet.

En algún momento de la década de los 90' es probable que McCaa haya *actualizado* los *datos* a algún tipo de *soporte* como un disco rígido de una PC[54]. Es claro como ahora comienzan a emparentarse los conceptos de *actualizar* y *copiar* ya que en el lenguaje cotidiano a esta operación se la suele llamar *copia*, del mismo modo que se dice que alguien *copió* tal archivo desde un disco rígido (un *soporte*) a un pendrive (otro *soporte*).

Luego, el auge de internet permitió *copiar* los *archivos* en un *soporte* compatible con un servidor web y así permitir, en principio, una mayor *difusión* al estar disponibles para ser *accedidos* y *copiados* desde cualquier PC con conexión a Internet que supiera de su localización precisa.

Es posible que por el tipo de *datos*, en lo referente al tipo de respuestas que contiene y la franja temporal que abarca (1869-1895) estas bases muestrales se hayan dejado de ofrecer oficialmente en el sitio de IPUMS ya

---

[53] *Paradójicamente se podría suponer que el proyecto de IPUMS, sin las ventajas de la evolución tecnológica, se podría parecerse más al proyecto caduco del Instituto Di Tella con respecto a las bases censales latinoamericanas. Para tener una idea de la diferencia se puede leer de los documentos originales "…para cada país se consignará en fichas IBM la información más relevante a los objetivos teóricos antes señalados…" (Di Tella, 1968a, p. 4) o "…cada variable standard, documentada como un "átomo de información" tendrá una tarjeta IBM…" (Di Tella, 1968b, p. 6).*

[54] *Parece haber evidencia de que IPUMS desde 1993 hasta 1995 suministraba información en cinta magnética (aunque posiblemente se trate de disquetes) de los archivos censales que poseía por lo que es posible que por esa época McCaa haya realizado traducciones de los archivos entregados por Lattes (Thomas & McCaa, 2002, p. 310). De todos modos, en la versión en inglés del texto citado utilizan el vocablo "tape" que efectivamente refiere a cinta y no el vocablo "floppy disk" o "diskette" que refieren a disco flexible o disquete.*

que en términos relativos sus *datos* son menos comparables que el resto de los disponibles en el sitio[55].

De todos modos los *archivos* siguieron estando hospedados en el servidor web de la página personal de Robert McCaa aunque con una escasa *difusión* debido a que si bien poseían una dirección web específica, al menos al momento en que se encontraron, estos se encontraban en un sector discreto de la página web. Es posible que esto se deba a actualizaciones posteriores de la página web que priorizaron otros aspectos de la misma frente a los archivos de las muestras censales[56].

Esta escasa *difusión* hizo posible que casi no se encuentre en la Argentina un uso explícito de estos datos en los últimos 30 años, a pesar de haber sido muy utilizados en el pasado[57]. Aunque parezca paradójico se han encontrado también utilizaciones de las muestras en el exterior (Baten, Manzel, y Stolz, 2010)(Droller, 2012)(Stolz y Baten, 2012).

Desenlace

El final de esta historia todavía se encuentra abierto pero en parte nos vuelve al pasado, porque nos reencontramos con el nombre de Gino Germani.

Luego de conversarlo con Alfredo E. Lattes, se concluye que sería adecuado ofrecer al Instituto de

---

[55] *Existe evidencia de que al menos para el año 2001 se contemplaba como oficial las bases de las muestras censales de 1869 y 1895 (Odinga & McCaa, 2001, p. 3).*

[56] *Por ejemplo en una publicación de 2001 (McCaa et al., 2001) se cita un enlace caduco pero en cambio ya para 2005 y 2007 se cita el enlace específico (por ahora) correcto (McCaa & Esteve, 2005)(McCaa & Esteve, 2007). Esto se suma a los escasos metadatos del archivo que hace difícil su indexación mediante los buscadores tradicionales.*

[57] *Los trabajos realizados con estas bases de datos o con tabulados originados en ellas, provienen principalmente de investigadores del Instituto Di Tella, luego del CENEP. Ejemplo de ellos son (Somoza, 1967)(Somoza, 1973),* (Lattes, 1968)(Lattes, 1970)(Lattes, 1974)(Lattes, 1975)*, (Recchini de Lattes &Wainerman, 1979), Kritz (Kritz, 1985) y Pantelides (Pantelides, 2006).*

Investigaciones Gino Germani (IIGG), específicamente a través de su Centro de Documentación e Información (CDI) la posibilidad de hacer un sitio web en donde no solo se informara sobre la existencia de las bases sino que se diera libre acceso a las mismas. Cabe aclarar que el IIGG aparte de llevar el nombre de Germani desde 1992, es la institución que sucede al Instituto de Sociología fundado en 1940 y del cual Gino Germani fuera su director entre 1957 y 1966.

Desde un lado más técnico, luego de los chequeos de consistencia y comprobar la escasa *mutación* de los *datos* frente a los análisis anteriormente publicados (principalmente Somoza & Lattes, 1967) se decidió iniciar un proceso de mejora en algunos puntos de los *datos* junto con un proceso de mayor *difusión* de los mismos. Esto estuvo motivado básicamente por el supuesto que una mayor *difusión* indirectamente también previene la *extinción* de los *datos* aparte de un mejor aprovechamiento académico de los *datos* disponibles.

Entre las mejoras que se realizaron a las bases se encuentran los siguientes:

- Se creó una variable que identifica a cada caso en forma individual para permitir la creación de códigos o sintaxis específicas que corrijan casos individuales.
- Se creó una base apilada que contenga los datos comparables de las bases individuales de cada censo.
- Se crearon ponderadores y expansores específicos para cada base que devuelven con aceptable precisión los valores poblacionales respectivos.
- Se creó una variable más detallada de la ocupación de los individuos mediante un proceso de “deconstrucción” de las variables existentes en las bases junto con anotaciones inéditas encontradas en la biblioteca del CENEP.

- Se *tradujero*n los *datos* disponibles a *formatos* (de texto plano) más compatibles y generales (Unicode–UTF-8) y otros más difundidos y funcionales (SAV-SPSS).
- Se digitalizó una serie de trabajos académicos relacionados con las muestras que facilitan y contextualizan a los usuarios de las bases[58].
- Se construyó un sitio web específico para su alojamiento, difusión e intercambio entre los usuarios de la base.
- Se crearon sintaxis específicas que replican los resultados originales de la publicación de 1967 (Somoza y Lattes, 1967).

Si bien esto ha sido un avance todavía falta:

- Hacer un control más cuidadoso de los casos perdidos de cada variable en pos de realizar alguna razonable imputación[59].
- Corregir los típicos problemas de la calidad de la declaración de la edad[60].
- Compartir más sintaxis entre los usuario que enriquezcan la calidad y cantidad de variables de la bases y permitan una mejor replicabilidad de los resultados.
- Errores menores que se descubran gracias el uso intensivo de las bases.

Por último, pero no por ello menos importante, este texto ilustra un modo de hacer una tipo de investigación

---

58 *En especial se destaca la digitalización del cuaderno de trabajo N° 46 del Instituto Di Tella en la que se detalla con bastante precisión el proceso de muestreo realizado (Somoza y Lattes, 1967).*

59 *Algunas variables poseen casos imputados pero es razonable suponer que las técnicas actuales de imputación son más adecuadas que las de los 60'.*

60 *Esto es un problema del censo mismo que se arrastra a la muestra pero de todos modos es deseable su corrección en función de algún modelo que suavice razonablemente la atracción de los dígitos "0" y "5", especialmente, entre las mujeres.*

poco común, que consiste en aprovechar la información cristalizada en archivos de papel sobre censos o registros antiguos y que vía un proceso de codificación se la digitaliza. De esta manera se convierten en *datos digitales* útiles para la investigación académica y, por su misma condición *digital,* permiten un mejor análisis de los mismos junto con una mayor *difusión* lo cual, a su vez, ayuda a evitar su *extinción* en el tiempo[61].

[61] *En Argentina puede nombrarse algunos proyectos de Darío Cantón que han obtenido financiamiento de la Agencia Nacional de Promoción Científica y Tecnología (PICT) y de la UBA (UBACyT) para investigar la relación entre el voto y migración a través de digitalizaciones previas de registros en papel (Cantón, Abdala y Acosta, 2008)(Cantón, 2010). Fruto de este proyecto es el libro "Una hipótesis rechazada"* (Cantón, Acosta, & Jorrat, 2013).

⌽*Epílogo*⌽

Datos, instituciones, investigadores y reglas de juego.

…igual que quien enciende su vela con la mía,
recibe luz sin que yo quede a oscuras.
…Las Invenciones entonces no pueden,
por naturaleza, ser objeto de propiedad.
(Jefferson, 1813, p. 6)

El conocer y poner en uso una máquina que no es completamente empleada,
aprovechar la experiencia de alguien que puede ser mejor utilizada,
o tener conocimiento de artículos sobrantes que pueden aprovecharse
durante una interrupción del abastecimiento es socialmente tan útil
como el conocimiento de mejores técnicas alternativas.
(Hayek, 1945, p. 522)

Como parece demostrar la historia académica existen situaciones en que información relevante para una sociedad, por una u otra razón, deja de ser *accesible* para las próximas generaciones de posibles usuarios.

En este sentido, pueden relacionarse circunstancias tan disímiles como el incendio de la biblioteca de Alejandría o la *extinción* de datos académicos modernos. En ambos se trata de información *cultural* considerada importante e irrepetible que se encuentra codificada en distintos lenguajes y soportes.

Algunas veces la información está codificada en alguno de los lenguajes *naturales* de diferentes civilizaciones. Otras veces, se encuentra codificada en lenguajes *artificiales* diseñados para fines específicos (por el ejemplo el binario). Todas ellas, sin excepción, necesitan de algún tipo de intérprete que decodifique el código inicial y así le otorgue significado a la información.

En el caso de las civilizaciones antiguas el lenguaje *natural,* mientras fuera compartido convencionalmente por una serie de individuos, era interpretado por el cerebro de cada uno de ellos. En el caso específico del lenguaje artificial digital se tiene una serie de intérpretes intermedios como el código fuente, los periféricos (más antiguamente) y los software (más contemporáneos).

Es claro que cuantos menos intérpretes se necesiten mayor será la probabilidad de replicar la información codificada, porque se es menos dependiente del (cambiante) ambiente de la decodificación.

En este sentido y con una visión temporal amplia, uno de los mayores encantos de las bibliotecas de "papel" es que necesitan menos intérpretes aunque esto no quiera decir que se encuentren exentas de las variaciones del ambiente. También lo son y el caso de Alejandría es un ejemplo.

Existen casos en donde los *datos digitales* se tornan *inaccesibles* por el deceso de los individuos o por diversas discontinuidades de las instituciones que poseen los *datos*. Por otro lado existen casos en que los propios *datos* se pierden por problemas de precaria conservación de los *soportes* o por falta de *actualización* de los mismos que los vuelve obsoletos.

En todos los casos la sociedad, sea por uno u otro mecanismo, pierde información relevante sobre su pasado y su identidad. Lo importante para el mundo académico es que el análisis o disección de los mecanismos permite comprender lo específico de cada proceso y su eventual control y/o cambio a través de alguna política pública y/o modificación de las reglas de juego de los agentes intervinientes.

Siguiendo el léxico utilizado anteriormente, la información *digital* presenta una serie de características distintivas en cuanto a su producción, *conservación* y *difusión* en comparación a la información analógica.

En este sentido podría existir un conjunto de políticas complementarias entre sí que no sólo no serían contradictorias (ya que apuntan a mecanismos diferentes) sino que podrían complementarse positivamente entre ellas.

La determinación de las reglas de juego eficientes en este campo depende en modo crítico de la tecnología disponible. En la actualidad la conjunción de la tecnología digital por un lado y la difusión de internet cambiaron críticamente la gama de opciones viables (Rothenberg, 1995)(Hedstrom, 1998)(Suber, 2012).

Esto hace que sea recomendable guardar los *datos primarios* académicos, en especial aquellos que por buenas razones deban preservarse, en algún *formato* digital. Por otro lado las *copias* digitales son extremadamente fieles por lo que la información no suele *mutar* en cada réplica fácilmente. Asimismo en la tecnología digital suele haber una notoria diferencia entre el costo de *producción original* y su *replicación* haciendo viables estrategias de *difusión* que en papel no eran posible[62].

Por otro lado los costos de los *soportes* en donde se guardan los *datos* también poseen costos decrecientes en el tiempo (en especial si se cuenta el espacio físico necesario frente a otros tipos de información). En cambio, un punto negativo es el problema de las *actualizaciones,* aunque, la creación de emuladores puede ser una solución parcial[63].

Con respecto a este último punto los datos guardados en un soporte tipo "papel" poseen dos grandes ventajas. Primero que bajo condiciones óptimas pueden durar siglos y segundo que alcanza con hacer réplicas de siglo en

---

[62] *Quizá algo parecido puede decirse de la imprenta. La reducción significativa de los costos de replicación de un libro (aún frente al pequeño aumento de la construcción del primero por la confección y configuración de los tipos móviles) expandió fuertemente la frontera de lo posible en términos de la difusión del conocimiento.*

[63] *La idea de emulación hace referencia a la posibilidad de recrear el ambiente en donde los datos puedan ser interpretados. Esta alternativa es particularmente efectiva para los diferentes formatos de datos aunque sólo para los soportes más modernos. En otras palabras, mientras que los soportes sean relativamente modernos, existen emuladores que de forma exitosa permiten acceder a formatos antiguos.*

siglo (sin cambios de formato ni de soporte[64]) debido a que la tecnología necesaria para acceder a los datos es el propio organismo humano y no una tecnología más cambiante en el tiempo.

Teniendo en cuenta estas características es importante que en el diseño de reglas de juego tanto las instituciones como los investigadores posean incentivos para favorecer la *conservación* y d*istribución* de los *datos* sin desincentivar por ello la *producción* de los mismos. Esto último es especialmente importante en el caso de los *datos primarios*. El caso de los *papers* parece tener su propia particularidad[65].

Posiblemente pocos duden de los beneficios de tener buenos *datos primarios*. El tema es que si su *producción* es costosa para los mismos productores importa quien corre con esos costos. La *producción* de *datos primarios* cuando es ejecutada por investigadores, aún en el caso que el proyecto se encuentre totalmente financiado por agentes estatales, de todos modos conlleva costos personales como tiempo y problemas de logística para los investigadores[66].

Es posible que si se diseñaran reglas de juego que obligaran de forma inmediata (por ejemplo 6 meses) a ceder los derechos sobre esos *datos primarios* a la comunidad entera la esfera de la *producción* de esos *bienes* vería acrecentada algunos de los problemas usuales de los *bienes públicos*.

---

[64] *Esta es la razón por lo que algunos monjes copistas no sabían ni leer ni escribir a pesar de dedicarse a hacer réplicas de libros enteros. Lo que se necesitaba era saber imitar diferentes tipografías.*

[65] *El caso de los papers es diferente porque suelen existir incentivos específcos como la cantidad de publicaciones, la cantidad de citas recibidas que junto con el escaso costo de su construcción (frente a grandes bases de datos primarios) hacen de que su producción no corra tanto riesgo aún en contexto de fuertes políticas de difusión de acceso abierto irrestricto.*

[66] *Cuando estos datos primarios son ejecutados por instituciones públicas el problema es diferente porque el investigador se acerca a la categoría de asalariado en el sentido de no poseer los derechos de propiedad ni derechos sobre beneficios residuales de los datos primarios. Por otro lado la institución misma rara vez posee un ambiente tan competitivo como el académico para los propios investigadores.*

Específicamente parece razonable suponer que aumentarían los *costos de oportunidad* de los investigadores productores. Esto puede llegar a ser así porque una cantidad no despreciable de investigadores (potenciales productores de datos primarios) podrían esperar a que otros investigadores corran con los costos personales mientras ellos aprovechan el tiempo para *analizar datos primarios producidos por otros y continuar publicando papers de su autoría*. Estas actividades si bien también implican costos personales para los investigadores también poseen una estructura de incentivos que parece asegurar su *producción*.

El problema es que dependiendo de la cantidad de investigadores adopten este último comportamiento dependerá cuanto se resienta la *producción* misma de *datos primarios*[67]. En ambientes altamente competitivos como el científico y aun en el caso que se asuma que los investigadores poseen un altruismo por el conocimiento mayor a la media de la sociedad, es razonable suponer que su impacto será apreciable.

Pueden existir políticas que logren atenuar esta especie de *free-rider académico* sin resentir la *producción* misma de los datos primarios? Quizá sea posible atenuar este efecto y sin embargo, en comparación a la actualidad, expandir la *difusión* de los *datos primarios* académicos aumentando la externalidad positiva de los mismos[68].

Por ejemplo si se otorgara un tiempo de exclusividad razonable de los datos (por ejemplo 5 años) quizá esto

---

[67] *Al menos, aquella parte del conjunto de los datos primarios que cae bajo la órbita de investigadores que deciden qué y cómo investigar a través de proyectos en busca de financiamiento.*

[68] *En este apartado se deja de lado el problema de la replicación inter-científicos de las investigaciones científicas. En este caso parece obvio que una más rápida difusión (por ejemplo 6 meses) pueda promover mayores replicaciones de los análisis de los datos entre los investigadores. De todos modos, este punto podría ser mantenido informalmente (esto es, sin derechos y obligaciones lagales) por la norma que implica el "comunismo" en la ciencia (Merton, 2002, pp. 642-644).*

incentive al investigador a correr con los costos personales de la *producción* de *datos primarios* porque en parte después podrá contar, de forma exclusiva por 5 años, de los beneficios de realizar publicaciones sobre ellos.

Pasado ese tiempo, se encontraría en la *obligación* (y no ya en su *derecho*) de ofrecer un acceso abierto a sus datos de una manera institucionalmente acordada. En la actualidad, este último punto, no parece ser el comportamiento más usual. Y las razones de ello, pueden encontrarse en las preferencias de los investigadores como la falta de incentivos a la misma (en un contexto de fuertes costos de oportunidad) pero también en las limitaciones de las instituciones en las cuales están insertos.

Es por eso que también es importante que las instituciones del sistema de producción científica (Institutos de investigación, universidades, etc.) también aporten lo suyo para que cuando el investigador tenga que ofrecer sus datos a la comunidad tenga herramientas para hacerlo de manera simple e útil. Si esta parte también se cumple entonces la sociedad como un todo, podría beneficiarse porque se aumentaría la *difusión* de los datos primarios sin resentir su *producción*.

Por otro lado se podría mejorar el sistema de cita. Más allá de comentarios ocasionales o agradecimientos informales como pie de página, las bases de datos primarias, rara vez se encuentran citadas junto con el resto de la bibliografía. En ese caso se podría incentivar tanto a las instituciones que ofrecen un reservorio como a los investigadores que producen los datos primarios, diseñando un sistema de citas que reconozca a ambos.

Esto podría ser efectivo aun el caso en que ni las instituciones ni los investigadores puedan sacar algún rédito económico más o menos directo. En la ciencia

también importa el reconocimiento de los pares (Merton, 1970) y en algunas instancias los incentivos económicos presentan efectos no deseados en los comportamientos de individuos con marcadas preferencias sociales[69].

En este sentido el estado argentino, en línea con los epígrafes del comienzo del texto y a través de la ley 26.899 inició el camino del acceso abierto para las actividades financiadas con ayuda de recursos públicos. Allí se regula que en un plazo no mayor a 6 meses para el caso de las publicaciones y 5 años para los *datos primarios* ambos tipos de información deben estar disponibles en reservorios online con acceso irrestricto.

Aquellas *personas* e *instituciones* que se consideren que no cumplan su parte de la regulación se tornarán candidatos no elegibles para el financiamiento público de futuras investigaciones.

[69] *Por ejemplo pagar por sangre suele reducir en algunos casos las donaciones (Mellström & Johannesson, 2008) o incluir multas por llevar tarde a los chicos al jardín aumenta el impuntualismo (Gneezy & Rustichini, 2000).*

La importancia de una lengua franca en los datos primarios digitales

Los formatos patentados pronto se convierten
en formatos heredados, cuya edad, dependencia de sistemas,
idiomas o hardware los hace difíciles,
costosos y a veces imposibles de traspasar
(Thomas & McCaa, 2002, p. 312)

Unicode marca el avance más significativo
en sistemas de escritura desde los fenicios
James j. O'donnell

Es sabido que no se puede hacer muchas predicciones acerca del futuro tecnológico. De todas maneras al día de hoy algunas conjeturas parecen más plausibles que otras. Entre las últimas puede que se encuentre aquella que afirma que la tecnología *digital* parece haber llegado para quedarse. Desde su *emergencia* ha tenido una constante *difusión* y lo que es más importante para nuestro interés es que muchas de las tecnologías de la información que se están proyectando parecen seguir manteniendo su vínculo con el formato *digital*.

Pero las conjeturas quizá no avancen mucho más de allí. Cuál de todos los formatos *digitales* se usará en el futuro de las bases de datos parece más una profecía que una predicción. En este caso, suponiendo una fuerte aversión a la *extinción* de los mismos, es razonable fomentar la *difusión* de formatos simples, abiertos y escalables. Esto podría convenir aún en presencia de formatos sumamente funcionales para las bases de datos actuales pero cuyos derechos de propiedad son privados.

En caso de poseer preferencias por una mayor *difusión* (y quizá de forma indirecta ayudar a prevenir una eventual *extinción*) los formatos más funcionales de la actualidad parecen ser una razonable opción. Esto último tiene el punto negativo que nada asegura que la tecnología que en la actualidad sea la más difundida en el futuro también lo siga siendo. En efecto, a largo plazo quizá la historia de la tecnología sea un cementerio de tecnologías obsoletas.

La importancia de realizar una *traducción* de los *datos* desde un formato ASCII hacia algunos de los *formatos* de Unicode (especialmente UTF-8) es que permite que un conjunto de otros caracteres también tengan su codificación manteniendo una compatibilidad absoluta con ASCII. Esto es importante para asegurar una mejor *difusión* de los datos en tanto que quizá en próximas ediciones de los *archivos* surjan *datos* que no puedan ser codificados de modo unívoco con 7 bits y que en caso que se use 8 bits sea una codificación internacionalmente aceptada[70].

[70] *Unicode es un esfuerzo colectivo por poder codificar de forma consensuada todos los caracteres, numerales y símbolos especiales de todas las lenguas conocidas de la especie humana. Más información en www.unicode.org.*

Bibliografía:

Arónica, S., & Buraschi, M. (2011). *Análisis de las políticas de incorporación tecnológica en la Facultad de Ciencias Económicas de la UNC*. Córdoba: Universidad Nacional de Córdoba.

Baten, J., Manzel, K., & Stolz, Y. (2010). Convergence and Divergence of Numeracy: The development of Age Heaping in Latin America, 17th to 20th century. Presentado en Historical Patterns of Development and Underdevelopment, Montevideo.

Bishop, L. (2006). A proposal for Archiving context for secondary analysys. *Methodological Innovations Online*, *1*(2), 10-20.

Bishop, L. (2007). A reflexive account of reusing Qualitative Data. Beyond Primary/Secondaty dualism. *Sociological Research Online*, *12*(3).

Blau, P., & Duncan, O. (1967). *The American Occupational Structure*. New York: Wiley.

Boudon, R. (1976). Comment on Hauser´s Review of Education, Opportunity, and Social Inequality. *The American Journal of Sociology*, *81*(5), 1175-1187.

Boudon, R. (1998). Social mechanisms without black boxes. En *Social mechanisms. An analytical approach to social theory*. Cambridge: Cambridge University Press.

Bunge, A. (1940). *Una nueva Argentina*. Buenos Aires: Guillermo Kraft Ltda.

Bunge, M. (1974). *Treatise on Basic Philosophy. Semantics I. Sense and Reference* (Vols. 1-VIII, Vol. I). Dordrecht: Reidel.

Canton, D. (2010). *Rasgos de los extranjeros que se inscriben en 1934 para votar en los comicios municipales porteños* (UBACYT 20020090100418). Buenos Aires: Universidad de Buenos Aires.

Canton, D., Abdala, F., & Acosta, L. (2008). *Migraciones internas de ciudadanos argentinos y voto en la capital federal y el conurbano alrededor de 1946 y de 1930-1934* (PICT 2008-1974). Buenos Aires: Universidad de Buenos Aires.

Canton, D., Acosta, L., & Jorrat, R. (2013). *Una Hipótesis rechazada. El rol de los migrantes según Gino Germani en los orígenes del peronismo*. Buenos Aires: Librería Hernandez.

Cornblit, O., & Mora y Araujo, M. (1967). *Proyecto para la creación de un archivo de datos sobre América Latina* (Documento de trabajo No. 41). Buenos Aires: Instituto Torcuato Di Tella.

De la Fuente, D. (1872). *Primer Censo de la República Argentina verificado los días 15, 16 y 17 de setiembre de 1869*. Buenos Aires: Imprenta del Porvenir.

De la Fuente, D., Carrasco, G., & Martínez, A. (1898). *Segundo Censo de la República Argentina, mayo 10 de 1895*. (Vols. 1-II). Buenos Aires: Taller Tipográfico de la Penintenciaría Nacional.

Di Tella, T. (1968a). *Banco de datos censales de América Latina: Lineamientos generales del proyecto* (Documento de trabajo No. 42). Buenos Aires: Instituto Torcuato Di Tella.

Di Tella, T. (1968b). *Metodología para un banco de datos socio-políticos de América Latina* (Documento de trabajo No. 43). Buenos Aires: Instituto Torcuato Di Tella.

Droller, F. (2012). *Migration and Long run Economic development: Evidence from settlements in the Pampas*. Providence: Brown University.

Essinger, J. (2007). *Jacquard´s Web*. Oxford: Oxford University Press.

Ford, R. (1950). A rapid Scoring procedure for Scaling Attitude Questions. *The Public Opinion Quarterly, 14*(3), 507-532.

Galtung, J. (1973). *Teoría y Método de la Investigación Social* (Vols. 1-II, Vol. I). Buenos Aires: Eudeba.

García Bouza, J. (1967). *El futuro desarrollo de los archivos de datos en Ciencias Sociales en América Latina* (Documento de trabajo No. 41). Buenos Aires: Instituto Torcuato Di Tella.

Germani, G. (1955). *Estructura Social de la Argentina. Análisis estadístico*. Buenos Aires: Editorial Raigal.

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies, XXIX*, 1-17.

González Bollo, H. (2010). Sobre la amenazante mayoría de dos provincias y una ciudad. Los tres primeros censos demográficos y su impacto político. *Estadística Española, 52*(174), 311-331.

Graciarena, J. (1987). Estudio preliminar. En G. Germani, *Estructura Social de la Argentina. Análisis estadístico.* (pp. 1-17). Buenos Aires: Ediciones Solar.

Hayek, F. (1945). The use of knowledge in society. *The American Economic Review, XXXV*(4), 519-530.

Hedstrom, M. (1998). Digital preservation: A time Bomb for Digital Libraries. *Computers and the Humanities, 31*, 189-202.

Hollerith, H. (1889). An electric tabulating system. *The Quarterly, X*(16), 238-255.

IBM. (1955). *The 650 magnetic drum data processing machine*. New York: IBM.

IBM. (1956). *IBM 650 Data processing System with 355 Random Access Memory and 838 Inquiry Stations*. New York: IBM.

IBM. (1958). *IBM 101 Electronic Statistical Machine*. New York: IBM.

IBM. (1961). *Form and Card Design*. New York: IBM.

IBM. (1962). *IBM Magnetic Tape Units*. New York: IBM.

IBM. (1970). *IBM 29 Card Punch*. New York: IBM.

IBM. (1971). *129 Card Data Recorder. Machine Description*. New York: IBM.

Jefferson, T. (1813, agosto 13). Thomas Jefferson to Isaac McPherson.

King, G., Rosen, O., & Tanner, M. (Eds.). (2004). *Ecological inference. New methodological strategies*. Cambridge: Cambridge University Press.

Kneale, W. (1972). Numbers and numerals. *The British Journal for the Philosophy of Science, 23*(3), 191-206.

Kritz, E. (1985). *La formación de la fuerza de trabajo en la Argentina: 1869-1914*. (Documento de trabajo No. 30). Buenos Aires: CENEP.

Lattes, A. (1968). *Evaluación y ajuste de algunos resultados de los tres primeros censos nacionales*

de población (Documento de trabajo No. 51). Buenos Aires: Instituto Torcuato Di Tella.

Lattes, A. (1970). Algunos indicios de migración interna diferencial en Argentina antes de 1869. En *Unión Internacional para el Estudio Científico de la Población* (pp. 558-562). Mexico D. F.

Lattes, A. (1974). Perspectiva histórica de la evolución de la población. En Z. Lattes & A. Lattes (Eds.), *La población de Argentina* (pp. 21-28). Buenos Aires: Committe for International Cooperation in NAtional Research in Demography.

Lattes, A. (1975). El crecimiento de la población y sus componentes demográficos entre 1870 y 1970. En A. Lattes, *La población de Argentina* (pp. 29-66). Buenos Aires: CICRED.

Lattes, A. (2010). La contribución de Germani al conocimiento de las migraciones. En *Gino Germani. La sociedad en cuestión* (pp. 402-409). Buenos Aires: Instituto Gino Germani - CLACSO.

Lattes, A., & Poczter, R. (1968). *Muestra del Censo de Población de la Ciudad de Buenos Aires de 1855* (Documento de trabajo No. 54). Buenos Aires: Instituto Torcuato Di Tella.

McCaa, R. (2013). The big data revolution: IPUMS-International. Trans-Border access to decades of census microdata samples for three fourths of the world and more. *Revista de Demografía Histórica, XXX*(I), 69-87.

McCaa, R., & Esteve, A. (2005). El proyecto IPUMS Internacional: Microdatos censales para investigadores argentino, latinoamericanos y del resto del mundo. En *Seminario Internacional de Población y Sociedad en América Latina*. Salta.

McCaa, R., & Esteve, A. (2007). El proyecto IPUMS - International: Microdatos censales para investigadores argentino, latinoamericanos y del resto del mundo. En Boleda & Mercado Herrera (Eds.), *Seminario Internacional de Población y Sociedad en América Latina 2005* (Vols. 1-II, Vol. I, pp. 51-74). Salta: GREDES.

McCaa, R., Haines, M., & Mulhare, E. (2001). The first national Historical Census Microdata. En *Handbook*

of International historical microdata for population research (pp. 13-22). Minnesota: The Minnesota Population Center.

McCaa, R., & Jaspers Faijer, D. (2000). The Standardized census sample operation (OMUECE) of Latin America 1959-1982 [1992]: A project of the Latin American Demographic Center (CELADE). En P. Kelly, R. McCaa, & G. Thorvaldsen (Eds.), *Handbook of International historical microdata for population research* (pp. 287-302). Minnesota: The Minnesota Population Center.

Mellström, C., & Johannesson, M. (2008). Crowding out in blood donation. Was Titmuss right? *Journal of the European Economic Association*, *6*(4), 845-863.

Mentz, R. (1991). Sobre la historia de la estadística oficial argentina. *Estadística española*, *33*(128), 501-532.

Merton, R. (1970). Behavior Patterns of Scientist. *Leonardo*, *3*(2), 213-220.

Merton, R. (2002). La ciencia y la estructura social democrática. En *Teoría y Estructura Social* (pp. 636-647). Mexico D. F.: Fondo de Cultura Económica.

Morales, J. (1972). El Banco de datos de CELADE. *Demografía y Economía*, *6*(1), 123-135.

Nie, N., Bent, D., & Hull, H. (1970). *SPSS. Statistical package of the social science*. New York: McGraw Hill.

Novick, S. (2004). *Aspectos jurídico políticos de los censos en la Argentina: 1852-1995* (Documento de trabajo No. 39). Buenos Aires: Instituto de investigaciones Gino Germani.

Odinga, A., & McCaa, R. (2001). *Statistical Confidentiality and the construction of anomized public use census samples: draft proposal for the Kenyan Microdata for 1989*. Minnesota: Minnesota Population Center.

Otero, H. (1998). Estadística censal y construcción de la nación. El caso argentino, 1869-1914. *Boletín del Instituto de Historia Argentina y Americana «Dr. Emilio Ravignani»*, (16-17), 123-149.

Otero, H. (2007a). Censos antiguos: 1869, 1895, 1941, 1947. En S. Torrado (Ed.), *Población y bienestar en*

la Argentina del primero al segundo centenario. Una historia social del siglo XX (Vols. 1-II, Vol. I, pp. 187-213). Buenos Aires: Edhasa.

Otero, H. (2007b). *Estadística y Nación. Una historia conceptual del pensamiento censal de la Argentina moderna (1869-1914)*. Buenos Aires: Prometeo LIbros.

Pantelides, E. (2006). *La transición de la fecundidad en la Argentina 1869-1947* (Documento de trabajo No. 54). Buenos Aires: CENEP.

Pereyra, D. (2010). Los científicos sociales como empresarios académicos. El caso de Gino Germani. En D. Pereyra (Ed.), *El desarrollo de las ciencias sociales. Tradiciones, actores e instituciones en Argentina, Chile, México y Centroamérica* (pp. 35-54). Costa Rica: CLACSO.

Quartulli, D. (2013, julio 6). *Digitalización de muestras de los 2 primeros censos de población Argentinos (1869- 1895). Haciendo visible (algunos) eslabones de una larga cadena*. Presentado en X Jornadas de Sociología de la UBA, Universidad de Buenos Aires. Buenos Aires.

Recchini de Lattes, Z., & Wainerman, C. (1979). *Empleo femenino y desarrollo económico. Algunas evidencias*. (Documento de trabajo No. 6). Buenos Aires: CENEP.

Remington Rand. (1957). *UNIVAC Scientific Computing System. Model II03A*. Minnesota: Sperry Rand Corporation.

Rothenberg, J. (1995). Ensuring the longevity of digital information. *Scientific American*, *272*(1), 42-7.

Somoza, J. (1967). *Nivel y diferenciales de la fecundidad en la Argentina en el siglo XIX* (Documento de trabajo No. 45). Buenos Aires.

Somoza, J. (1973). La mortalidad en la Argentina entre 1869 y 1960. *Desarrollo Económico*, *12*(48), 807-826.

Somoza, J., & Lattes, A. (1967). *Muestras de los dos primeros censos nacionales de población, 1869 y 1895* (Documento de trabajo No. 46). Buenos Aires: Instituto Torcuato Di Tella.

Sørensen, A. (1998). Theoretical mechanisms and the empirical study of social process. En P. Hedström & R. Swedberg (Eds.), *Social mechanisms. An analytical*

approach to social theory. Cambridge: Cambridge University Press.

Sørensen, A. (2009). Statistical models and mechanisms of social processes. En P. Hedström & B. Wittrock (Eds.), *Frontiers of sociology* (Vol. II). Leiden: Brill.

Stinchcombe, A. (1978). *Generations and cohorts in social mobility: Economic development and social mobility in Norway* (Memorandum No. 18). Oslo: Institute of applied social research.

Stolz, Y., & Baten, J. (2012). Brain drain in the age of mass migration: Does relative inequality explain migrant selectivity? *Explorations in Economic History*, *49*(2), 205-220.

Strickland, J. (2012). Who invented the 80 column, rectangular hole punched card? *Volunteer Information Exchange*, *2*(17), 2-3.

Suber, P. (2012). *Open Access*. Massachusetts: Massachusetts Isntitute of Technology.

Sweger, K. (1965). *IBM 1130 Computing System User´s Guide*. New York: IBM.

Thomas, W., & McCaa, R. (2002). Preservación de archivos con documentos y microdatos censales y aumento de los grupos de gestión. *Notas de Población*, *75*, 303-320.

Van Bogart, J. (1995). *Magnetic tape storage and handling. A guide for libraries and archives*. Washington: National Media Laboratory.