tecnim@p Sevilla 2006
30 mayo a 2 junio
060.es
24x7x365

Comunicación

121

DEL LIBRO AL DATO

Armando de la Torre del Río
Jefe del Área Editorial
Instituto Nacional de Estadística

Palabras clave

Bases de datos, estadística pública, libros, publicaciones estadísticas, memorias estadísticas de organismos públicos, Internet, difusión de datos, aplicaciones informáticas de difusión de datos.

Resumen de su Comunicación

Las bases de datos estadísticas suelen contener información actual (la última información para un tema concreto), y en el mejor de los casos información que se remonta a las dos o tres últimas décadas, cuando se generalizó el tratamiento informatizado de la información.

En algunos casos alguien se ha encargado de construir (introduciendo los datos por procedimientos convencionales) series temporales más largas, que suelen limitarse, por lo laborioso de su elaboración a las informaciones mas relevantes. Ese es el caso de los bancos de datos de algunas instituciones con función estadística, el INE, el Banco de España, los Ministerios, los organismos autonómicos de estadística. ¿Qué hacer con las muy relevantes masas de datos, vertidas en libros de papel, anteriores a la época "informática"?

Son datos muy relevantes en estudios comparativos, en todas las tareas habituales de análisis de fenómenos demográficos y económicos en el medio y largo plazo.

El INE ha abordado dos líneas de trabajo complementarias para esa tarea: una ya en producción, el apartado "Historia" de su base de datos gratuita "INEbase", y otra, con un cierto componente experimental, basada en llevar al límite las posibilidades que hoy brinda la técnica OCR (Reconocimiento óptico de caracteres), orientada a recuperar para su base de datos numérica el contenido de tablas estadísticas solo existentes en libros, sin traslación hasta ahora al mundo de la difusión de datos estadísticos por Internet.

Sobre algunos aspectos de oportunidad, dificultad técnica y coste de esta última línea de trabajo versa esta comunicación.

El INE confía en poder ofrecer tanto los resultados de este trabajo, como la experiencia adquirida en el uso de este refinamiento de las tecnologías OCR, a cualquier organismo o entidad interesada en "desempolvar", hacer universalmente accesibles, sus colecciones de publicaciones y memorias estadísticas.

DEL LIBRO AL DATO

1. Antecedentes

Como primera aproximación al problema de sacar a pública difusión en Internet, los datos de informaciones estadísticas solo disponibles en libros se iniciaron una serie de análisis de alternativas en el INE.

Cuando decimos "sólo disponibles en libros" no nos referimos sólo a los casos en que eso es estrictamente cierto.

A veces se conservan los "microdatos" de partida, e incluso los programas con los que fueron tratados, programas de tabulación (no entraremos a considerar las fase previas de grabación, depuración e imputación, típicas de todo proceso estadístico).

En esos casos una alternativa es "resucitar" el proceso, modificándolo levemente para que dé como resultado matrices de "macrodatos" y tablas estadísticas propiamente dichas, que puedan ser cargadas por los procedimientos usuales en las bases de datos numéricas destinadas a la difusión estadística.
Pero suelen aparecer problemas, de coste, de oportunidad, incluso algunos tan humanos como el derivado de volver a hacer funcionar programas, por bien documentados que estén, si fueron hechos en un lenguaje algo antiguo, o para ordenadores hoy en desuso, o por personas que quizá ya no están disponibles. Estamos hablando entonces de información histórica, aunque pertenezca a la "era de la informatización".

Esto puede ocurrir con operaciones estadísticas incluso de la década de los ochenta, cuando ya se manejaban rutinariamente sistemas informatizados de tabulación.

Si miramos hacia atrás, a las décadas de los setenta y anteriores, y en el caso del INE a su larga historia, pues el primer Anuario de España corresponde al año 1858, encontramos, tirando por lo bajo, al menos dos centenares de miles de páginas, susceptibles de "republicación".

2. La experiencia de "INEbase História"

Hace ahora dos años se procedió a escanear página a página una parte considerada especialmente relevante del fondo editorial del INE, todos los Anuarios, y todos los Censos de Población y Viviendas.

Se tomaron estos libros se sometió a escaneado y posterior proceso de OCR cada imagen de página, y con la ayuda de los textos capturados, y mediante una aplicación específicamente desarrollada al efecto, se consiguió un entorno favorable para realizar una catalogación, que permite finalmente al usuario navegar por contenidos detallados (básicamente a través del equivalente a los índices de los libros), con capacidades de búsqueda relevantes. Aun así, el objeto de esta base de datos es la "página".

En esta zona especial de INEbase, el usuario dispone de dos niveles de búsqueda, el primero a través de los índices, y cuando ya ha llegado a la página que contiene la tabla estadística de su interés, un segundo dentro de la propia página, ya que esta se suministra en el popular formato PDF, pero con el texto reconocido en la fase OCR incrustado en el propio archivo, lo que permite buscar textos (provincias, categorías) dentro de ella.

Aún así hay que reconocer que, al igual que en otras experiencias de bibliotecas públicas nacionales relevantes con presencia en Internet, el objeto de búsqueda y recuperación es la página, no la propia tabla estadística o matriz multidimensional con datos y "metadatos" que es el objeto que caracteriza a las bases de datos de difusión estadística.

Figura 1. Menú de Anuarios Estadísticos de INEbase Historia

Una cincuentena de Anuarios Estadísticos son ya accesibles, en lo que cabe considerar una infraestructura documental y educativa propia de la sociedad de la información.

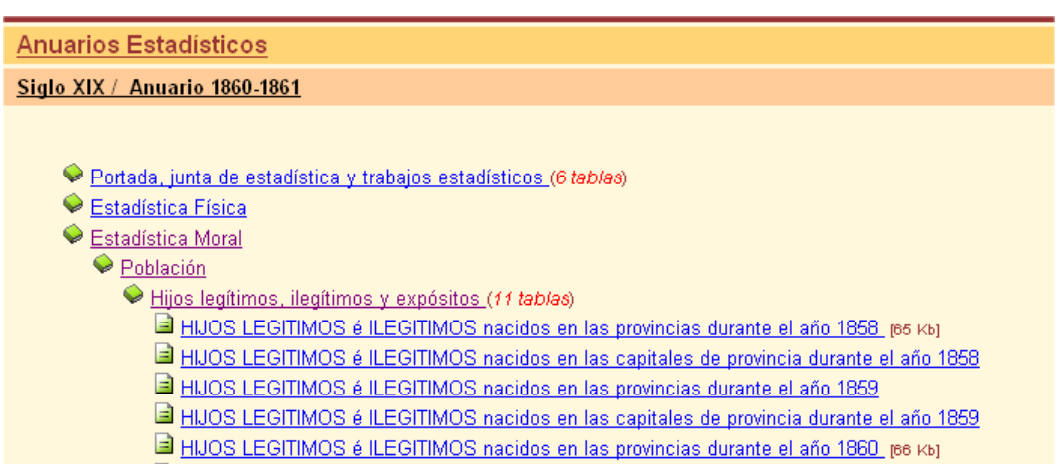
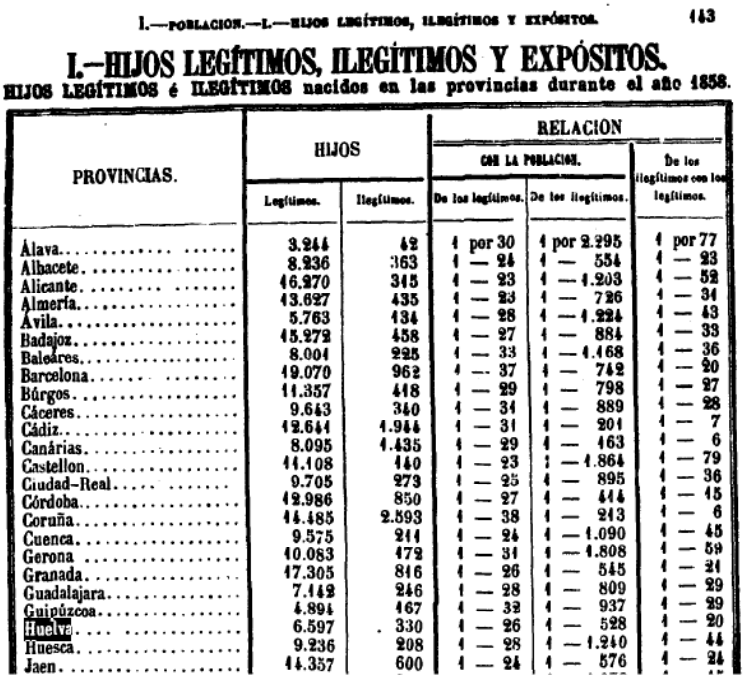


Figura 2. Índice de tablas. Anuario de 1860-1861. INEbase Historia



I.—POBLACION.—I.—HIJOS LEGÍTIMOS, ILEGÍTIMOS Y EXPÓSITOS. 143

I.—HIJOS LEGÍTIMOS, ILEGÍTIMOS Y EXPÓSITOS.

HIJOS LEGÍTIMOS é ILEGÍTIMOS nacidos en las provincias durante el año 1858.

PROVINCIAS.	HIJOS		RELACION		
			CON LA POBLACION.		De los ilegítimos con los legítimos.
	Legítimos.	Ilegítimos.	De los legítimos.	De los ilegítimos.	
Álava	3.244	42	1 por 30	1 por 2.295	1 por 77
Albacete	8.236	363	1 — 24	1 — 554	1 — 23
Alicante	16.270	345	1 — 23	1 — 1.203	1 — 52
Almería	13.627	435	1 — 23	1 — 726	1 — 31
Ávila	5.763	134	1 — 28	1 — 1.224	1 — 43
Badajoz	15.272	458	1 — 27	1 — 884	1 — 33
Baleares	8.004	225	1 — 33	1 — 1.168	1 — 36
Barcelona	19.070	962	1 — 37	1 — 742	1 — 20
Búrgos	11.357	418	1 — 29	1 — 798	1 — 27
Cáceres	9.643	340	1 — 31	1 — 889	1 — 28
Cádiz	12.644	1.944	1 — 31	1 — 201	1 — 7
Canárias	8.095	1.435	1 — 29	1 — 163	1 — 6
Castellon	11.108	140	1 — 23	1 — 1.864	1 — 79
Ciudad-Real	9.705	273	1 — 25	1 — 895	1 — 36
Córdoba	12.986	850	1 — 27	1 — 414	1 — 15
Coruña	14.485	2.593	1 — 38	1 — 243	1 — 6
Cuenca	9.575	211	1 — 24	1 — 1.090	1 — 45
Gerona	10.083	172	1 — 31	1 — 1.808	1 — 59
Granada	17.305	816	1 — 26	1 — 545	1 — 21
Guadalajara	7.142	246	1 — 28	1 — 809	1 — 29
Guipúzcoa	4.894	167	1 — 32	1 — 937	1 — 29
Huelva	6.597	330	1 — 26	1 — 528	1 — 20
Huesca	9.236	208	1 — 28	1 — 1.240	1 — 44
Jaen	14.357	600	1 — 24	1 — 576	1 — 24

Figura 3. Mostrando una tabla. Anuario de 1860-1861. INEbase Historia

Se espera difundir todos los Anuarios en lo que queda del presente año, y continuar catalogando y difundiendo Censos de Población, que existen, como tal concepto estadístico desde 1900.

3. Un paso más: del libro al dato

En este proyecto, actualmente en desarrollo, se aprovecha el esfuerzo previo de escaneado de páginas, tratando de amortizar doblemente su coste, pero ahora se trata de llevar la técnica de OCR hasta el límite actual de dicha tecnología, que ha avanzado notablemente en los últimos años.
Ahora se están tratando técnicas (solo parcialmente automáticas) que permiten discriminar en cada página la "matriz" de datos, tratando de extraerla con la menor pérdida de información.

Como primer paso se prevé acometer la conversión de los datos del Censo del año 1981, el inmediatamente anterior a los ya disponibles en Internet (el de 2001, al que se accede a través de una zona especial de INEbase dotada de características Data Warehouse , y el de 1991, accesible desde la zona general de dicha base. (Del resto de los censos existen también algunas tablas principales y resúmenes).

El volumen de datos que se está tratando es de:

- 18.917 archivos TIFF con tablas.
- 3.400.000 celdas totales

Se describen a continuación algunos de los procesos más relevantes seguidos en las aplicaciones que se están construyendo.

Un primer programa se dedica a la eliminación de las líneas de división, verticales y horizontales, típicas de los cuadros de tablas estadísticas. Se trata en definitiva de ponérselo un poco más fácil al motor o motores OCR.

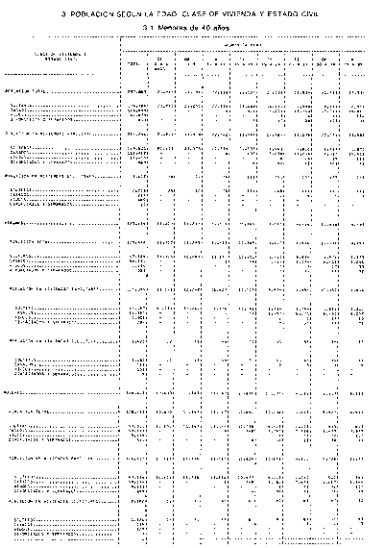


Figura 4. Una típica imagen de partida. Censo de 1981. Del libro al dato.

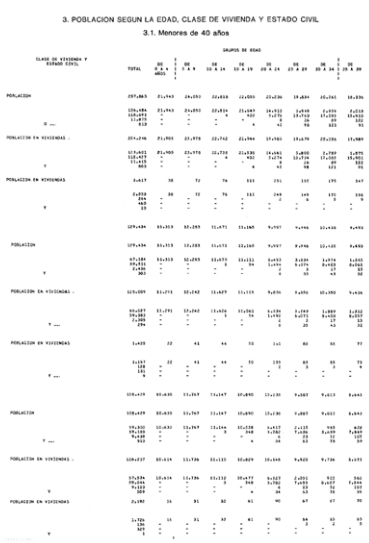
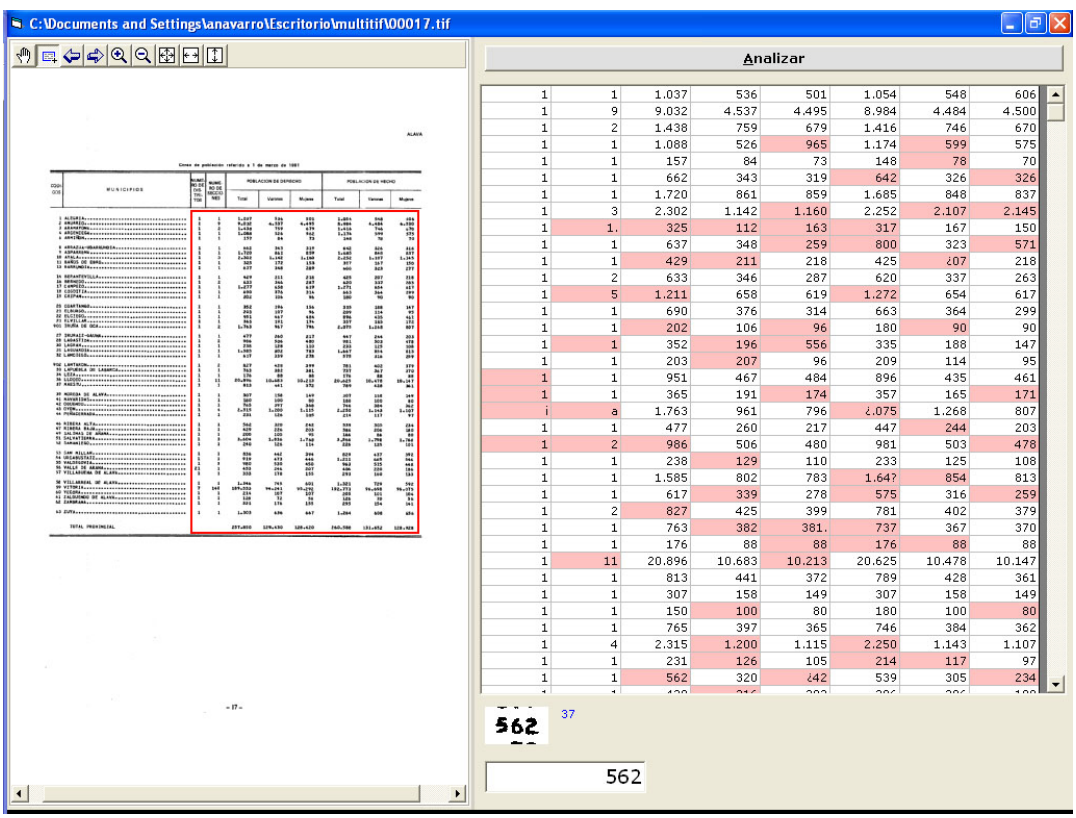


Figura 5. Una tabla sin líneas. Censo de 1981. Del libro al dato.

Como puede verse, de cada página tratada en este proceso resulta un nuevo fichero, adecuado aún para permitir la visualización tanto de la imagen original en fases posteriores.

Es ahora cuando se necesita la intervención de un operador (...por ahora, las tablas pueden ser tan distintas, que es difícil definir un patrón geométrico que permita automatizar este paso).

Dicho operador definirá visualmente, con una herramienta de selección por rectángulo, la tabla o grupos significativos o "fragmentos de tabla", que son zonas de la imagen que contienen matrices de números alineados en filas y columnas



El sistema localiza automáticamente el número de la página, que servirá de guía en posteriores comprobaciones, traspasa los caracteres numéricos encontrados a una rejilla contenedora, en la que ya se hace operar al motor o motores OCR y se colorean las celdas que presentan un nivel de calidad del reconocimiento inferior al fijado como objetivo. En el ejemplo mostrado se pone un objetivo alto, una parte significativa de las celdas se colorean para su verificación y en su caso grabación manual.

En esta fase se concentra la mayor dificultad técnica del proceso. Cuanto más se "eduque" al motor o motores de reconocimiento en la tipografía usada en ese libro en concreto, mejores resultados se obtendrán. Precisamente, algunos loables "ahorros" de costes, en la edición de la serie de libros de aquel Censo, están creando ahora dificultades en el proceso: se utilizaron impresoras láser convencionales, con 150 ppp (tecnología de la época) en lugar de sistemas fotográficos de filmación de textos originales, mucho más costosos, cualquier especialista conoce el salto de resolución y costes entre uno y otro sistema.

Enseñanza evidente: si un humano lee más cómodamente un original, también un sistema OCR tendrá más fácil su trabajo.

Se están haciendo esfuerzos en "educar" a los sistemas de reconocimiento de patrones de caracteres, para que venzan las dificultades del proceso: caracteres con poca resolución, originales con tipografías "cegadas" por exceso de intensidad de "toner", dificultades especiales en la distinción entre ciertos caracteres (el 5 y el 6 por ejemplo).

Curiosamente, pruebas con libros más antiguos pero en los que se usaron sistemas de composición y filmación convencionales, dan resultados mejores resultados, que estos "recientes" libros de la década de los 80.

Una vez sometidos los datos presentados en las citadas rejillas a la visualización y grabación y verificación de cada celda dudosa, ahora por operadores humanos, se pasan nuevos programas que producen para cada libro o unidad de tratamiento, un fichero de texto, que contiene la secuencia de matrices de datos del libro, precedida cada matriz de unos identificadores de Página, fragmento de tabla o tabla, fichero de imagen origen, etc.

Se ha tenido que confeccionar también un programa que permite concatenar, vertical u horizontalmente los fragmentos de tabla, y eliminar líneas o columnas de subtotales innecesarios, para construir matrices que tengan una lógica compatible con los metadatos, pues lo que tiene sentido impreso no siempre lo tiene en una presentación informatizada.

¿Qué se hace con esas matrices?.

Algo a lo que la unidad de Difusión del INE está bien acostumbrada, unirle los correspondientes metadatos (cabecera, ladillo, notas, título) hasta formar una tabla estadística en el formato PC-Axis ®, un popular formato, predecesor, en el caso de los "objetos" estadísticos del actual lenguaje XML, que fue desarrollado, hace bastante años por Statistics Sweden, y que ha sido profusamente usado en el INE.

La confección de las metadatos de las matrices, que parecería muy trabajosa, no lo es tanto, gracias al uso de editores especiales, que permiten capturar de cualquier formato de archivo o de base de datos de clasificaciones, una sola vez, las listas de valores de variables temáticas, geográficas o temporales, y luego combinarlas de forma sencilla, según el orden de presentación y anidamiento de cada tabla. Un técnico bien entrenado, usando uno de estos editores, como el llamado "PX-Make ®", produce fácilmente en un día uno o dos centenares de esas "tablas en blanco" o metadatos de tabla, que solo esperan que un sencillo programa, también de uso frecuente en el INE, case cada tabla en blanco con su correspondiente matriz de datos.

Las tablas así construidas, son verificadas por técnicos del servicio promotor (la Subdirección General de Censos y Padrón del INE), y se vuelcan, como ocurre con otros tantos productos, en la base de datos de difusión estadística INEbase. El usuario no notará ninguna diferencia, con otras tablas cargadas desde otras aplicaciones, mucho más sencillas, precisamente de eso se trata.

Ahora, ya en INEbase, los datos se pueden filtrar, anidar, exportar a Excel o al propio formato PC-Axis, representar gráficamente, o en su caso con ayuda de mapas, en fin, dejan de ser "imagen de tabla" para ser plenamente "tabla". Del libro al dato, como queríamos demostrar.

Obviamente, los procesos complementarios de control de calidad, similares a los usados en el caso de que se hubiera procedido a una grabación convencional, forman parte del proceso de aseguramiento de la fiabilidad de lo publicado, y constituyen una parte, no por menos llamativa, menos importante del proceso. Algo importante, el coste. El contrato se ha licitado a un precio de 0,0088235 euros por celda reconocida,

de forma que el coste de traslación de un grupo de libros a ficheros publicables se descompone en este coste unitario por celda de datos, mas unos gastos evaluables en cuatro meses/persona, del equipo que ha de preparar y asociar los metadatos.

En cuanto a plazos, el trabajo de reconocimiento y exportación a ficheros de matrices, para la cantidad de imágenes de páginas antes indicada es de seis meses, por lo que cabe esperar que la nueva publicación de dicho Censo, ahora en Internet, pueda hacerse en el primer trimestre de 2007.

Solo insistir una vez mas en que ahora el objeto de publicación no es la página, de modo que el usuario no deberá ser particularmente consciente de que está accediendo a una publicación histórica, porque se le presenta con la misma funcionalidad que en los contenidos de la "época informática", no se le muestran páginas de publicaciones, sino tablas estadísticas.

El INE confía en poder ofrecer tanto los resultados de este trabajo, como la experiencia adquirida en el uso de este refinamiento de las tecnologías OCR, a cualquier organismo o entidad interesada en volver a publicar, en Internet sus colecciones de publicaciones y memorias estadísticas.

Referencia en Internet (del proyecto " INEbase HISTORIA")
http://www.ine.es/inebaseweb/welcome.do?language=0