VII JORNADAS SOBRE TECNOLOGÍAS DE LA INFORMACIÓN PARA LA MODERNIZACIÓN DE LAS ADMINISTRACIONES PÚBLICAS
A CORUÑA
15, 16, 17 y 18 de octubre de 2002
Palacio de Congresos y de la Ópera
www.tecnimap.com
Ponencias

TRANSFORMACIÓN DE LA ADMINISTRACIÓN PÚBLICA EN BASE A LAS TECNOLOGÍAS DE LA INFORMACIÓN EN LA REALIZACIÓN DE LOS CENSOS DE POBLACIÓN Y VIVIENDAS DE 2001

José Antonio Perea Yustres (INE)

Introducción

Los Censos Generales, como los que se han realizado durante fin del 2001 y principio del 2002, son la operacion estadística de mayor magnitud e importancia para el sistema estadístico nacional en tanto que determinan los parámetros fundamentales de los edificios, los locales, las viviendas y la población residente en el conjunto del país y para cada uno de sus ámbitos territoriales y administrativos, suministrando la información que será básica durante la próxima década.

Las informaciones que suministran dichos Censos son fundamentales para todas las Administraciones Públicas, en relación con sus tareas de planificación, gestión y seguimiento de servicios y políticas públicas, y que dichas informaciones son también esenciales para las empresas, los agentes sociales, los investigadores y los ciudadanos.

Los Censos de Población y Vivienda, por el volumen de información manejada, el numero de personas y Organismos implicados, y los requisitos de obtener resultados en los menores plazos posibles y con las menores molestias para los ciudadanos, son una operación que requiere una utilización intensiva de las tecnologías de la información y las comunicaciones mas avanzadas.

Para lograr los objetivos censales, el Instituto Nacional de Estadística ha realizado un gran esfuerzo utilizando las tecnologías más novedosas y eficaces, siendo estos Censos punteros en todo el mundo tanto en la utilización de tecnología de captura de cuestionarios en papel, como en facilitar a los ciudadanos su cumplimentación a través de Internet. Estos esfuerzos están en línea con las directrices estratégicas marcadas por el Plan INFO XXI para la construcción de la Sociedad de la Información en España.

Este proyecto se viene elaborando desde el año 1.991, fecha de realización del anterior Censo de Población y Viviendas, y fue presentado a través de una comunicación en el TECNIMAP celebrado en Palma de Mallorca con el título de "RECONOCE".

A continuación describiremos el proyecto informático que supone el mayor reto informático actual de captura y gestión de la información que existe a nivel mundial.

"Cuestionarios en papel: el mayor sistema de gestión documental avanzada del mundo"

Mediante este título se presenta el proyecto de producción censal que ha sido capaz de procesar mas información en menos tiempo de los proyectos de gestión documental avanzada que se conoce en el mundo. Al mismo tiempo es uno de los entornos informáticos de gestión documental más complejo en cuanto al volumen de datos a tratar en tiempo real y a las infraestructuras hardware, software, tratamiento de imágenes y sistemas de almacenamiento existentes.

Los objetivos del proyecto informático censal son los siguientes:

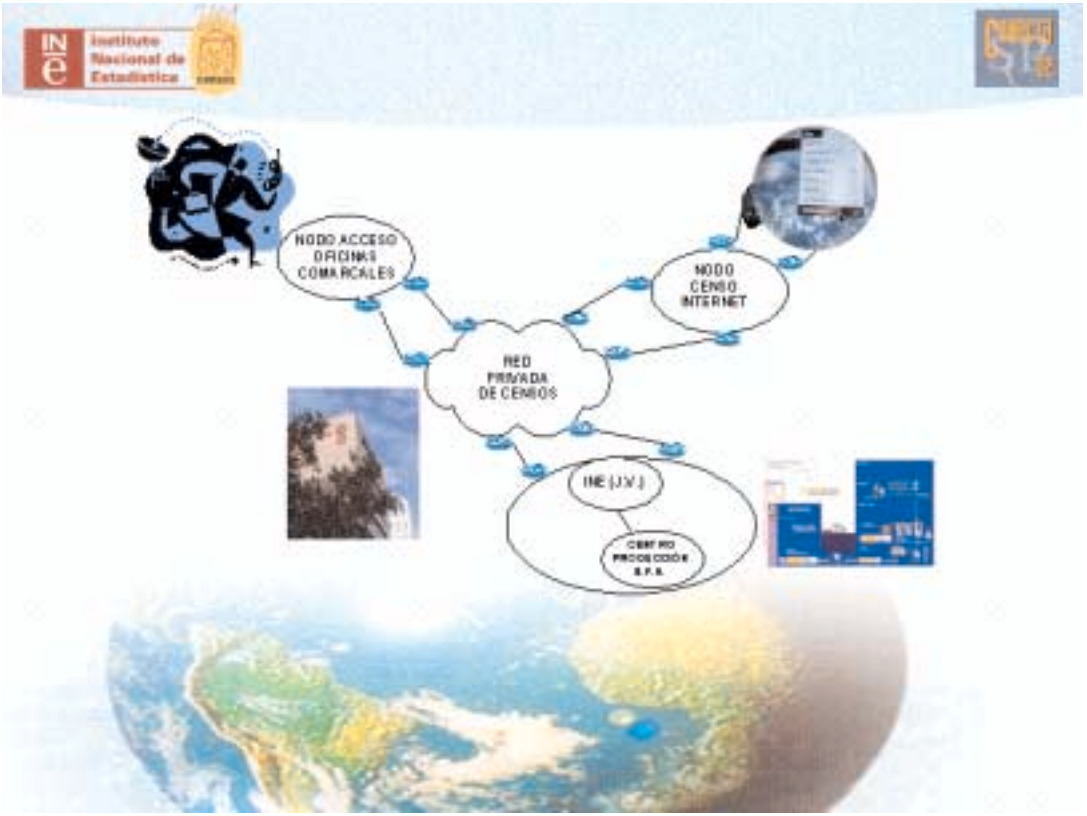
- Generar la información para realizar la preimpresión de los cuadernos de recorrido y de la información padronal
- Dar soporte a la gestión de la recogida (informatización y conectividad de la red de oficinas comarcales con el I.N.E)

- Realizar las contrataciones de personal eventual para la recogida y gestión del mismo
- Posibilitar la cumplimentación por Internet (de forma complementaria a la cumplimentación en papel)
- Realizar la captura y depuración básica de la información recogida en papel
- Realizar la difusión de la información de forma ágil y rápida
- Cumplir con los plazos previstos y con los niveles de calidad exigidos
- Garantizar la protección de los datos por ser de carácter personal (A.P.D.)
- Coordinar las actuaciones con las CCAA y ayuntamientos
- Apoyar las actuaciones en un conjunto de empresas especializadas seleccionadas por concursos públicos
- Envíar las imágenes padronales a todos los ayuntamientos de España.

La arquitectura de sistemas y de comunicaciones del Censo viene determinada de una parte por los sistemas disponibles en el propio INE, tanto en sus Servicios Centrales como en sus Delegaciones Provinciales y de otra por las contrataciones de los servicios de la cumplimentación de los censos vía Internet y la comunicación telefónica de las Oficinas Comarcales (500) con el INE. Estos sistemas se completan con el equipamiento necesario para realizar los procesos de Gestión Documental basados en el Reconocimiento Inteligente de Caracteres (ICR), y dotado de la infraestructura tecnológica mas avanzada del mundo.

Para asegurar que los intercambios de información se realizan en las debidas condiciones de seguridad y poder asegurar igualmente la confidencialidad de la información en el caso de ficheros con datos de carácter personal, se ha definido una extranet entre las Oficinas Comarcales y el INE, que se ha denominado Red Privada de los Censos (RPC). La RPC constará del conjunto de equipos físicos y lógicos y del sistema de comunicaciones del propio INE, complementado con un conjunto de redes locales, una para cada una de las Oficinas Comarcales.

Las comunicaciones entre dichas redes darán servicio tanto a voz como a datos a través de telefonía móvil, pasando siempre por los servicios centrales del INE.



Los tratamientos informáticos que se aplican a los datos censales están fuertemente condicionados por el enorme volumen de información a procesar y por la reducción sustancial del tiempo que los usuarios demandan para obtener los datos censales. Ambos factores confluyen en que los tratamientos censales, aparte de asegurar la calidad de los procesos, deben ser ante todo rápidos.

Los condicionantes del actual proceso informático de producción censal que nos encontramos son los siguientes:

- Procesamiento informático de mas de 60.000.000 de cuestionarios (más de 100.000.000 de imágenes entre anversos y reversos de los cuestionarios)

- 5 tipos de cuestionarios con 60 modelos[1] y un total de 120 imágenes diferentes para reconocimiento óptico (anverso y reverso de los cuestionarios):
  - Padrón
  - Vivienda
  - Hogar
  - Individual
  - Cuadernos de recorrido
- Reconocimiento inteligente de caracteres (ICR) manuscritos y marcas[2].
- Utilización de imágenes digitalizadas de los cuestionarios para realizar los procesos de gestión documental de producción y postproducción[3].
- Criticidad en tiempos: La producción se realizará en menos de 3 meses.

Como consecuencia, se puede afirmar que el proyecto de producción censal de los cuestionarios en papel, no es solo el mayor proyecto de almacenamiento y gestión documental avanzada[4] en España, sino que tampoco existen precedentes a nivel mundial de un proyecto de Gestión Documental de estas características y dimensión. No existe un proyecto de Gestión Documental que debido a su criticidad en tiempos, volúmenes de información y seguridad, requiera de un total de 34 TB procesados en menos de 3 meses y que cuente con volúmenes físicos y lógicos de hasta 4,5 TB. manteniendo a la vez alta disponibilidad en cluster.

1 *Tanto las hojas padronales como los cuestionarios censales tienen modelos bilingües para cada uno de los idiomas oficiales del estado (Castellano, Catalán, Gallego, Mallorquín, Valenciano y Vasco). Igualmente existen diferentes modelos según el número de personas que componen la unidad familiar.*

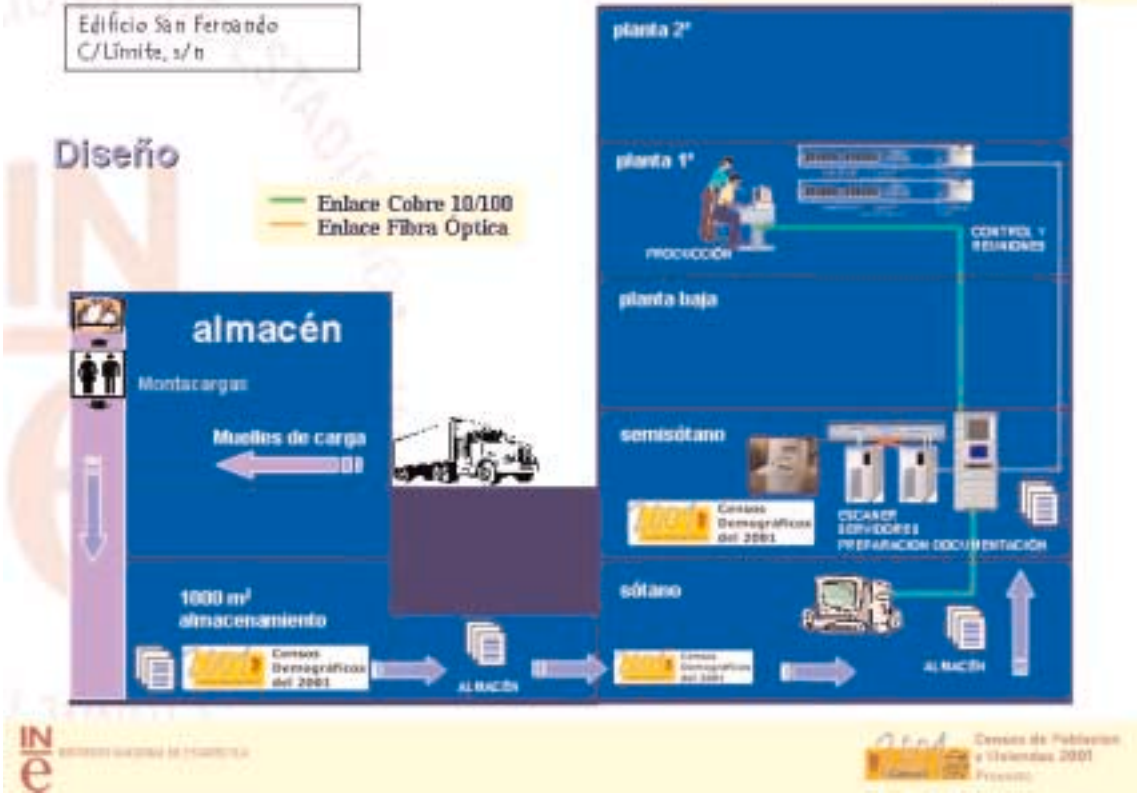
2 *El número de diferentes tipos de escritura a reconocer en los cuestionarios es igual al total de personas que los rellenan.*

3 *Las imágenes servirán para enviar a todos los municipios españoles sus respectivos cuestionarios padronales.*

4 *Por gestión documental avanzada se entiende aquel proyecto en el que se utilizan sistemas de digitalización de alta producción, reconocimiento inteligente de caracteres manuscritos (alfabéticos y numéricos) y gestión de imágenes.*

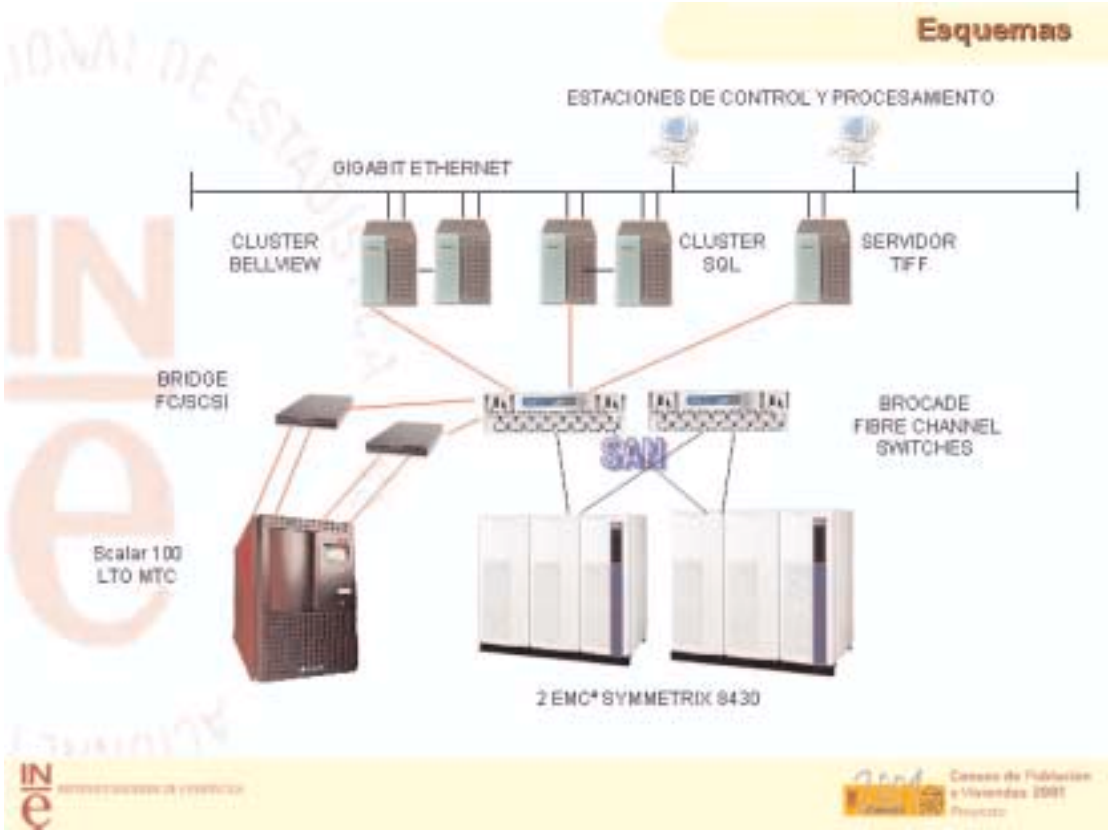
Contamos con el Centro de Producción Censal (CPC) situado en la localidad de San Fernando de Henares en Madrid (a partir de ahora SFH). En este macrocentro, que cuenta con más de 5.000 mts2, se ha realizado la Producción de todo el Censo de Población y Viviendas 2001, a excepción de los cuestionarios que se han cumplimentado vía Internet. Más de 1.500 personas han trabajado en la explotación de los datos censales. También, durante el tiempo de recogida de la documentación, se ha contado con 500 oficinas comarcales en toda España que han sido las encargadas de remitir los cuestionarios al centro de producción censal de SFH.

Esquemas



La realización de los Censos de Población y Vivienda, encomendada al Instituto Nacional de Estadística, es una operación enormemente compleja en la que han participado más de 40.000 personas, que durante un periodo de unos dos meses y medio han recorrido 21 millones de direcciones postales para recoger información sobre los edificios, las viviendas ocupadas y las personas que habitan en ellas. Se han visitado unos 13 millones de hogares y recogido información de aproximadamente 40 millones de personas.

Los equipamientos físicos y lógicos necesarios para realizar el Censo de Población y Viviendas 2001, conforme al modelo de procesamiento establecido, se representa en el gráfico siguiente:



Las unidades de almacenamiento masivo que dispone el INE para este proyecto tiene una capacidad de 34 TB (equivalentes a 34.000 colecciones de la Enciclopedia Británica) y están dotadas de las medidas de seguridad informática mas avanzadas que existen hasta el momento (Raid 0+1, Cluster, fibra óptica, SAN...).

los 34 TB. planificados se requieren para realizar el almacenamiento de las imágenes digitalizadas y poder realizar la explotación de la información recogida a través de procesos de reconocimiento inteligente de caracteres, control y depuración estadística de los datos obtenidos a través de las imágenes anteriormente citadas, tablas, mejora de literales, codificaciones automáticas... no olvidemos que, por ejemplo, la base de datos contiene un registro por cada ciudadano de nuestro país (40.000.000).

Para ello contamos con 2 sistemas de almacenamiento centralizado de última generación con 221 discos de 181 Gbytes (25 TB en total) y cada uno de ellos cuenta con 4 puertos fibra óptica FC SW (switch fabric).

Debido a las características especiales de tiempo que se dispone para realizar la Producción informática (2,5 meses de producción masiva) y del volumen de información a procesar (digitalización de 60.000.000 de cuestionarios ,lo que hace unas 100.000.000 de imágenes escaneadas), hemos tenido la necesidad de conformar unidades lógicas (volúmenes físicos de información) de 4,5 TB cifra récord a nivel mundial que sobrepasa los límites conocidos en entorno Windows hasta el momento.

La arquitectura informática está diseñada en SAN (Storage Area Network) en la que múltiples servidores comparten un sistema de almacenamiento de forma segura vía protocolo Fiber Channel) y cuenta con:

- Servidores en cluster de la de 4 y 8 vías en Rack (armarios especiales que reúnen a todos los procesadores) con procesadores Xenon a 700 MHz con 2 MB de memoria cache y 3 GB RAM. Sistema de Backup LTO que cuenta con una capacidad de 15 TB, y un ratio de transferencia de 15 MB/s con una velocidad de copia de 324 Gb/h. Mas de 400 unidades con 1 y 2 procesadores Pentium III a 1 GHz y 256 y 512 MB RAM respectivamente
- El Sistema operativo empleado es Windows 2000 Advance Server, la base de datos Microsoft SQL 2000.
- Contamos con un sistema de copias en CD con el que se ha enviado a cada municipio de España (8.000 aprox.) las imágenes correspondientes a su Padrón Municipal.
- Como elemento fundamental del sistema de Producción hemos contado con 12 Sistemas de digitalización de alta producción de última generación que nos ha permitido un procesamiento de la imagen en tiempo

real. Tiene una resolución de hasta 400 dpi Velocidad de 120 páginas por minuto o 240 imágenes/minuto en modo duplex (por las dos caras). La velocidad de producción real debido a paradas, atascos... ha sido de 80 ppm (duplex). La producción se ha realizado durante 24 horas diarias 7 días a la semana (24 x 7).

- Consola única de gestión de rendimientos, configuración, alarmas, topología y capacity planning de cualquiera de los elementos de la SAN.
- Los programas de reconocimiento óptico ayudado con sistemas de mejora de literales, codificación automática, diccionarios... Permitirá realizar niveles de reconocimiento superiores al 80% de los procesado.

La operación de captura se realiza mediante un sistema de reconocimiento óptico de caracteres, que incorpora procedimientos de codificación automática, de control de rangos y de coherencia intra e inter registros, consta de los siguientes procesos:

- Digitalización mediante escáneres ópticos de alta producción.
- Control de cobertura de la digitalización
- Reconocimiento inteligente de caracteres manuscritos
- Sistema de mejora de literales y codificación asistida
- Videocorrección asociada al reconocimiento y a los controles de coherencia
- Control del flujo de trabajo
- Control de calidad
- Gestión documental

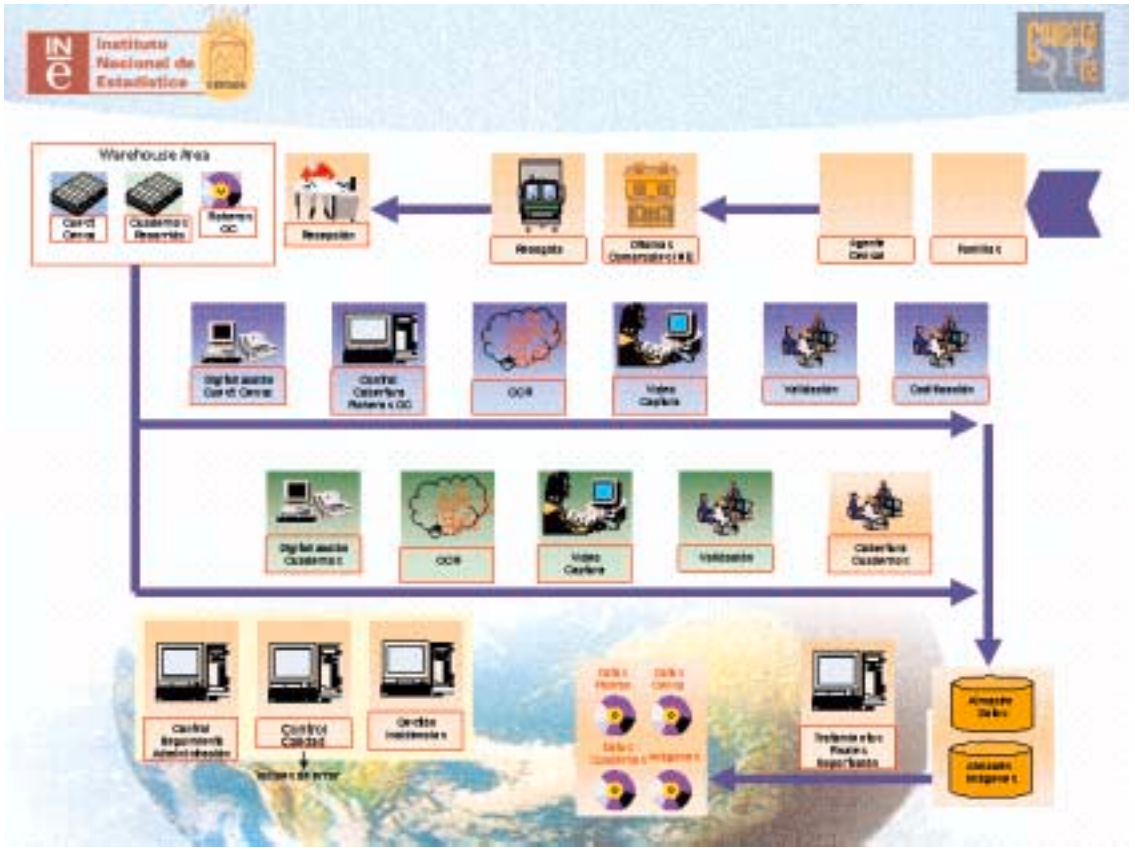
Algunos datos que reflejan la envergadura del proyecto son los siguientes:

- Número total de cuestionarios: 60.000.000 A4 doble cara (100.000.000 de imágenes) 3 meses (ICR)

- Recepción diaria de 32 palets de documentación: 3900 cajas conteniendo cerca de 1.800.000 cuestionarios.
- Velocidad de digitalización en producción 24 horas x 7 días: 6000 cuestionarios/hora (doble cara. Aprox. 100 ppm).
- Producción diaria con 12 escáner: 1.800.000 cuestionarios (doble cara)
- Jornadas de digitalización 40.

El esquema de producción censal del INE se divide en las siguiente áreas de gestión:

- Área A - Gestión de entrada / salida de la documentación censal
- Área B - Gestión de la digitalización de la documentación censal.
- Área C - Gestión del sistema de Videograbación censal.
- Área D- Gestión de Validación de datos censales.
- Área E - Gestión del sistema de tratamiento de los cuadernos de recorrido censales.
- Área F – Control de backup.
- Área G - Gestión del control informático.
- Área H - Gestión del control y administración, almacenamiento, comunicaciones y seguridad física y lógica de los ficheros de imágenes y datos censales.
- Área I - Control de calidad.
- Área J - Gestión del sistema de incidencias.
- Área K - Gestión del control, seguimiento y administración de la red general de producción censal del INE.



Hay que hacer mención a la base de datos de producción masiva que ha ofrecido los siguientes resultados:

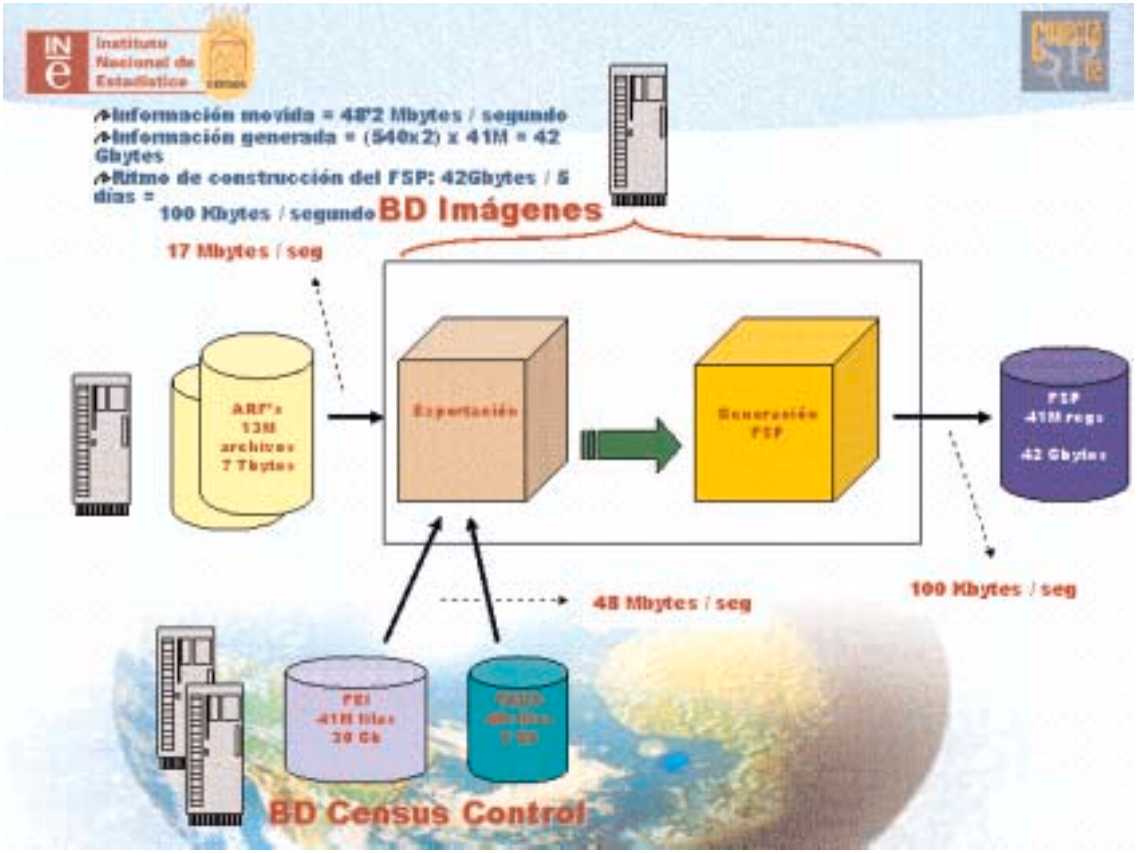
1. DbCensusControl:

- 300 Gbytes en RAID 0+1 ‡ Total de 600 Gbytes

- Tabla LotesPadron : 125.000 registros (1 por lote) ‡ 25 Mbytes (datos) + 100Kb (índices)
- Tabla SobresPadron: 16.000.000 registros (1 por familia o sobre)‡ 10 Gbytes (datos) + 1 Gbytes (índices)
- Tabla FEI: 41.000.000 registros (1 por individuo) ‡ 30 Gbytes (datos) + 9 Gbytes (índices)
- Tabla FAUX: 4.000.000 registros ‡ 5 GBytes (datos) + 3 Gbytes (índices)
- Tablas FEIS: 41.000.000 registros ‡ 40 Gbytes + 10
- Operaciones realizadas:
    - 14.000.000 inserciones a partir de diskettes de oficinas comarcales
    - Más de 50.000.000 de búsquedas para localizar información de sobres en tablas de proceso: ‡ 16.000.000 x 3 (cobertura, validación, exportación) + redundancias
    - Más de 100.000.000 de búsquedas para localizar información de individuos: ‡ 41.000.000 x 2 (validación, exportación)

2. DbImagenes:

- 300 Gbytes en RAID 0+1 ‡ Total de 600 Gbytes
- Operaciones realizadas:
- Exportación: Generación de 39 millones de registros
- Generación de FSP: Cruce de información entre FEI y grabación ‡ más de 1.000 Billones de operaciones



Podemos indicar que los objetivos informáticos pretendidos se han cumplido consiguiendo disminuir los tiempos dedicados a capturar y procesar informáticamente la información, pasando de mas de un año para realizar la captura de los datos a solo 3 meses. Igualmente se ha favorecido la utilización de las nuevas tecnologías, consiguiendo un proyecto de envergadura mundial reconocido con el premio al Mejor Proyecto de Tecnologías de la Información 2001 otorgado por COMPUTING España y VNU Business publications España.

Al mismo tiempo se pretende rentabilizar la inversión en infraestructura reaprovechando los sistemas una vez finalizado el proyecto de producción censal.