tecnimap Murcia 2004

23

SISTEMA DE INFORMACIÓN PARA LA DIFUSIÓN DE LOS RESULTADOS DE LOS CENSOS DE POBLACIÓN Y VIVIENDAS 2001

J. Adolfo Gálvez Moraleda
Jefe de Área de Planificación de Procesos Estadísticos
Instituto Nacional de Estadística

1. INTRODUCCIÓN

1.1 La difusión estadística

La difusión de resultados constituye para el órgano elaborador de una operación estadística la fase final de la misma y de las decisiones adoptadas en este punto depende, en buena parte, la utilidad que la citada operación pueda proporcionar a los usuarios interesados. Una mala estrategia de publicación de resultados (por el nivel de agregación presentado, por el uso de soportes que no permiten análisis de los datos,…) puede arruinar una operación estadística realizada con criterios de calidad en el resto de sus fases.

El número y variedad de usuarios de los datos estadísticos es creciente día a día. Por ello, en la difusión estadística actual se priman, además de valores tradicionales como la fiabilidad, el disponer de un buen nivel de desagregación, la rapidez en la publicación de resultados; otros no menos importantes como son la facilidad de uso, la posibilidad de realizar comparaciones en el tiempo y el espacio, información adicional sobre la operación (metodología, significado de las categorías que describen una variable…).

La calidad perseguida en cuanto a difusión estadística tiene como marco de referencia estas cuestiones, y para dar solución a todas ellas, el INE apoya la publicación de resultados en Internet como canal básico de difusión y usa distintas tecnologías que van desde ficheros no estructurados y notas de prensa con las que se publican avances de resultados de las distintas operaciones, a la difusión de datos detallados en ficheros estructurados; pasando por la difusión de series temporales del banco de datos.

Toda la información puesta a disposición del público en la web del INE se difunde bajo la marca INEbase que ayuda a contar con un interfaz "único" y flexible. La base técnica de este sistema flexible, para la gran mayoría de las operaciones estadísticas, es la combinación de recursos de una base de datos relacional y de ficheros estructurados de tablas estadísticas, un tipo de ficheros basado en un lenguaje "mark-up" anterior al XML denominado PC-Axis, desarrollado por la oficina estadística sueca y adoptado por un amplio conjunto de países europeos y fácilmente migrable al actual estándar XML.

Esto facilita a los usuarios la localización de la información al estar ésta agrupada temáticamente. Además se cuenta con un potente motor de búsquedas.

En cuanto a contenidos, la difusión estadística está esencialmente basada en un conjunto de tablas predefinidas en las cuales se da la información sobre los cruces más relevantes de las variables estudiadas en cada operación. Además se ofrecen las publicaciones de síntesis que abordan un aspecto concreto y en las que se ofrecen datos de varias operaciones relativas al tema tratado.

El siguiente escalón consiste en pasar de sistemas de difusión a sistemas de información que respondan a las consultas que el usuario formula. Una primera aproximación a este enfoque es dar al usuario la capacidad de crear sus propias tablas eligiendo las variables a presentar, su orden de anidación, la unidad de medida, el nivel de detalle, … Esto tiene particular relevancia en el caso de las operaciones estadísticas que permiten una mayor riqueza de información a distintos niveles territoriales, desde los más generales (total nacional, por CCAA,…) a los más concretos; esto es, descender a lo que se conoce como estadística de áreas pequeñas: secciones censales, unidades poblacionales, tramos de vía…En particular los Censos de Población y Viviendas permiten este tipo de explotación.

1.2 La difusión en los Censos de Población y Viviendas 2001

Los Censos de Población y Viviendas, realizados en la forma tradicional, constituyen la operación más voluminosa en cuanto a recursos, información, presupuesto …que la oficina central de estadística de un país ha de acometer.

En la ronda de Censos de 2001 el INE apostó por un fuerte componente de innovación tecnológica en las fases de preparación (preimpresión de datos padronales como base de los cuestionarios censales), recogida (posibilidad de cumplimentar los cuestionarios a través de Internet), captura de la información (digitalización de todos los cuestionarios censales). Algunas de estas innovaciones han sido presentadas en anteriores ediciones de Tecnimap.

Llegado el turno de la explotación y difusión de los datos de esta operación también se da relevancia a este aspecto tecnológico proponiendo un nuevo sistema de consultas que, basado en un sistema de almacenamiento datawarehouse proporciona flexibilidad en la forma de analizar la información al tiempo que se concede importancia prioritaria a Internet como canal de difusión por delante de las publicaciones tradicionales o electrónicas de operaciones censales anteriores.

Las características principales del nuevo sistema de difusión de datos censales que el INE ha dispuesto para los ciudadanos son las siguientes:

- Orientación de la difusión hacia Internet como canal básico de difusión.
- Almacenamiento de la información en estructuras multidimensionales orientadas a proporcionar un ágil entorno de consultas (no son ejemplos de gestión de espacio en disco, transacciones rápidas,…)
- Desarrollo de un sistema de explotación de la información que además de proporcionar a los usuarios las consultas requeridas dota de capacidades de navegación OLAP a través de los datos presentados en los informes. .

Se describen aquí los elementos básicos de este sistema informacional y la tecnología empleada en su desarrollo.

2. EL PROYECTO DE DIFUSIÓN CENSAL

2.1 Objetivos

La meta que se propone el INE alcanzar en la difusión de los datos censales es devolver al conjunto de la sociedad española el esfuerzo realizado en cuanto a presupuesto, colaboración a la hora de rellenar los cuestionarios,…en el valor de una información que dé servicio a cualquier ciudadano interesado. La práctica totalidad de las variables recogidas en la operación censal va a procesarse, depurarse y publicarse a través, principalmente, de Internet.

Dado el volumen de información a publicar se considera adecuado dar prioridad a que el sistema permita a los usuarios especificar qué consulta quieren obtener, en lugar de las publicaciones habituales basadas en un conjunto extenso de tablas previamente definidas por el órgano productor de la estadística. Para ello dispondrán de las variables generadas en la operación censal y podrán organizar la consulta tal y como deseen.

Los objetivos concretos que constituyen la definición del proyecto de difusión censal son dos:

- Alcanzar un buen rendimiento en las consultas al sistema teniendo en cuenta que en estas consultas será el usuario quien defina los informes, que existirá un enorme volumen de información disponible y que el canal de difusión elegido implica un entorno de usuarios con gran concurrencia en determinados periodos de actividad.
- Conseguir una aplicación de consultas con una amplia funcionalidad al tiempo que sea sencilla e intuitiva que permita atender las necesidades tanto de usuarios generales como especializados.

2.2 Elección tecnológica

Del análisis de soluciones realizado para alcanzar el primer objetivo propuesto se concluye que el centro de la arquitectura de este sistema de información sea un datawarehouse (DW) que permita la organización y almacenamiento de los datos para su procesamiento analítico en línea (On Line Analytic Processing – OLAP).

En esencia, un DW es un conjunto de datos orientado a la consulta que tiene como objetivo básico garantizar la coherencia de la información que contiene. Normalmente integra distintas fuentes de información operacionales para dar respuesta a las necesidades de información sobre determinados aspectos del negocio de una organización y basa la coherencia de la información que contiene en un conjunto de metadatos que define cada categoría de cada variable almacenada.

Cuando se integra nueva información ésta debe ser además de extraida de la fuente original, transformada de acuerdo a las especificaciones marcadas por los metadatos y cargada finalmente en la base de datos que constituye el DW. A esto se conoce como ciclo de Extracción, Carga y Transformación de la información (ETL) y son los procesos que alimentan el DW.

Disponer de un entorno en el que la integridad de la información está garantizada proporciona un soporte seguro sobre el que basar el proceso de toma de decisiones. Las herramientas de explotación de estos sistemas permiten obtener informes a medida que permiten capacidades de análisis del fenómeno que representan profundizando en los aspectos que más interesan, estableciendo comparaciones, representaciones gráficas…A estas capacidades se les conoce bajo el nombre genérico de análisis OLAP. .

Las tecnologías de almacenamiento DW y OLAP forman parte de lo que ahora se conoce como soluciones de inteligencia de negocio: Business Intelligence (BI)

El segundo objetivo marcará la necesidad de adaptación de las herramientas propias de estos entornos dado que, en nuestro caso, no hay oportunidad de formar a los usuarios potenciales de la misma. Los usuarios deben realizar consultas sin ocuparse de cuál es la estructura física y lógica del modelo de datos y van a poder navegar por la información censal sin conocer ninguno de los conceptos de los entornos que proporcionan análisis en línea u OLAP (drill down/up/accross, rotate, sort…). Se llega a la conclusión de que es necesario un desarrollo de interfaz a medida que simplifique en la medida de lo posible todos estos aspectos.

Además, el INE debe dar servicio a todos los usuarios y en ese sentido, el sistema de consultas debe ser, en la medida de lo posible, independiente de la tecnología de base del cliente (navegador de Internet utilizado, velocidad en la comunicación…) lo que provoca que el des-

arrollo esté condicionado por estos aspectos (uso de etiquetas y funciones entendidas por la mayoría de navegadores, se evitan descargas de aplicación en el lado del cliente…).

2.3 Arranque del proyecto

El INE analiza distintas experiencias de otras oficinas de estadística con prestigio en temas de difusión y que para esta operación también quieren destacar en el sistema de publicación. Los sistemas de difusión de las oficinas de estadística de Holanda, Brasil e Italia son analizados en profundidad. En estos sistemas se ofrece tradicionalmente un gran volumen de información a disposición del público con herramientas avanzadas que permiten al usuario definir con precisión la información a conseguir. Además se analizan experiencias en empresas privadas, si bien, las características de sus sistemas de información son distintas de las necesidades descritas de la difusión estadística.

Generalmente, en los ambientes de BI tienen gran relevancia las actualizaciones periódicas de los datos. La orientación de las consultas suele estar bien definida y el número de usuarios, normalmente especializados, que acceden a las mismas es relativamente pequeño. En este caso, los datos que se cargan en el sistema de información no sufren actualizaciones periódicas, si bien el número de variables que el usuario puede consultar es muy alto y no existen apenas restricciones, salvo limitaciones impuestas por la necesidad de protección del secreto estadístico, a los posibles cruces de variables que pueden realizarse. Además el número y diversidad de usuarios potenciales del sistema y la restricción de considerar tiempos de respuesta en el entorno internet aceptables hacen que este proyecto, aunque enmarcado en las soluciones tecnológicas de BI/DW tenga características particulares y que son las propias de un sistema de información estadística.

Determinados los principales objetivos a conseguir y los condicionantes que pueden acarrear se decide, restringir el entorno tecnológico del proyecto a plataformas y productos consolidados en el mercado del Business Intelligence y conocidos, en cierta medida, por el INE. Se convoca concurso público para el suministro de herramientas y el desarrollo de la interfaz del sistema de información y, como característica a destacar, en el pliego de condiciones se establece la necesidad de presentar un prototipo del sistema a construir a empresas seleccionadas en función de la oferta técnica presentada.

Tras el análisis de las tecnologías y prototipos realizados, la solución está basada en tecnología SAS que proporciona soporte para un modelo de datos HOLAP y sobre el que se desarrolla la interfaz de usuario como aplicación que cumple estándares J2EE.

El desarrollo del sistema lo lleva a cabo una U.T.E. entre las empresas SAS Institute y Capgemini.

2.4 Componentes funcionales

Se identifican distintos componentes funcionales a la hora de abordar el proyecto:

- Técnica de sistemas
- Construcción y carga del modelo de datos.
- Desarrollo de la interfaz de usuarios.

En base a estos componentes se establece el equipo de proyecto con recursos especializados en cada uno de ellos.

3.- ELEMENTOS RELEVANTES DEL SISTEMA DE DIFUSIÓN CENSAL

3.1 Interfaz de usuarios http://www.ine.es/censo/es

3.1.1 Componentes tecnológicos

El sistema construido para la explotación de la información censal tiene como interfaz una solución basada en tecnología Java con estándares J2EE que permite a los usuarios definir sus propias consultas a través de una navegación asistida.

La interfaz esencialmente está apoyada en páginas dinámicas que se construyen del lado del servidor (servlet y jsp) y nunca en el cliente por la razón, ya comentada en este documento, de evitar descargas de applets a través de internet que, si bien pueden incorporar elementos más dinámicos, hacen pesado y a veces inviable, el proceso para aquellos usuarios con menos recursos de comunicación con el servidor.

La herramienta de desarrollo AppDev Studio incorpora un entorno de desarrollo integrado Java J2EE con Java Beans que encapsulan la funcionalidad necesaria para la conexión a SAS y utilización de sus recursos

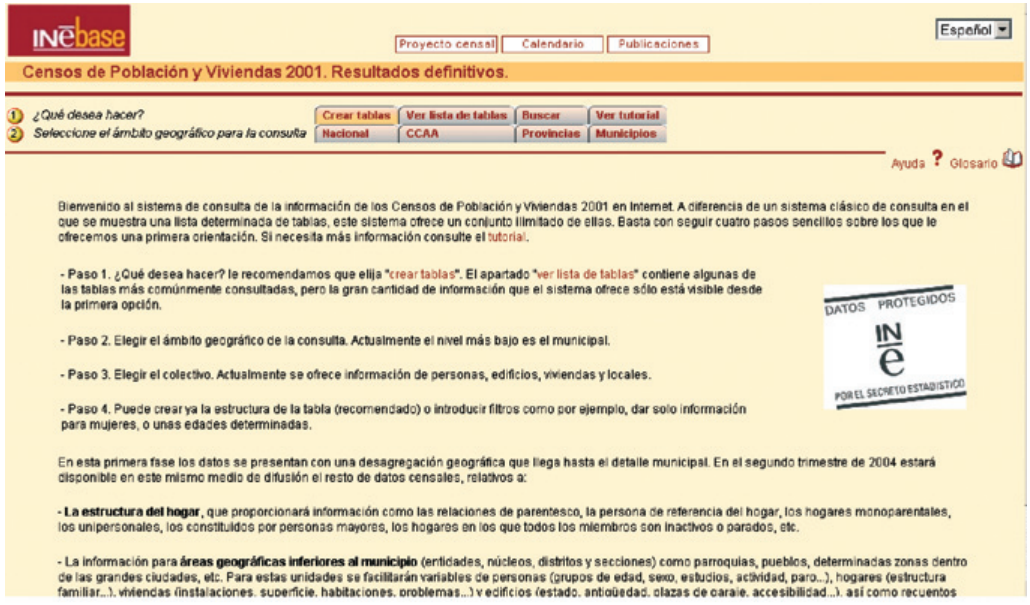
Como servidor de aplicaciones se utiliza Sun One Application Server sobre una máquina SunFire 280R que, momentáneamente, también es utilizado como servidor de datos, si bien la arquitectura de la aplicación permitiría separar en máquinas diferentes estos componentes.

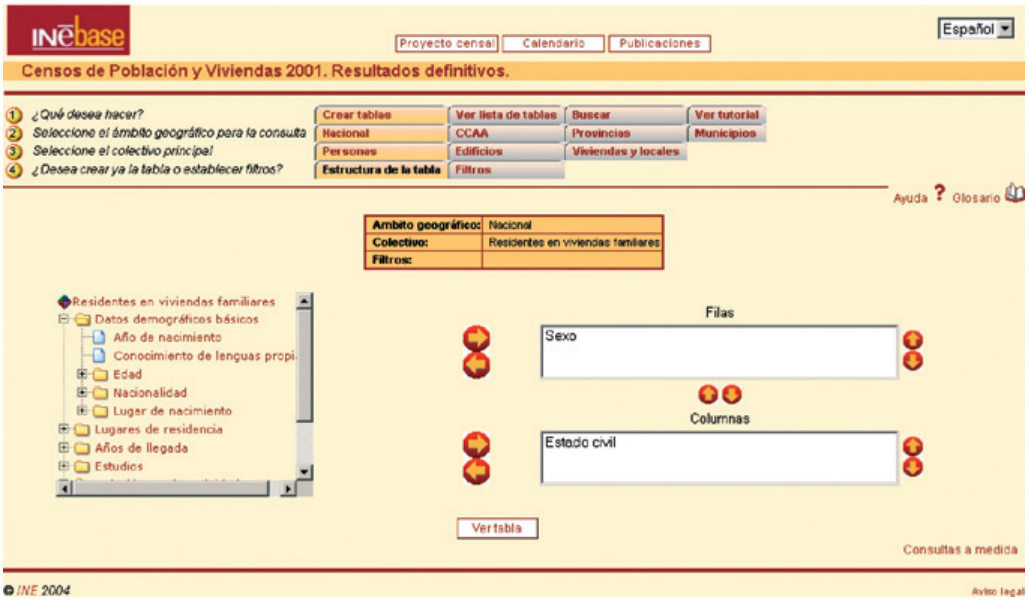
Para el mantenimiento de los usuarios registrados del sistema se utiliza un directorio LDAP.

Las descargas de peticiones a medida y peticiones batch se sirve por protocolo FTP en un servidor al efecto.

3.1.2 Funcionalidad

Se muestran algunas pantallas de esta interfaz.





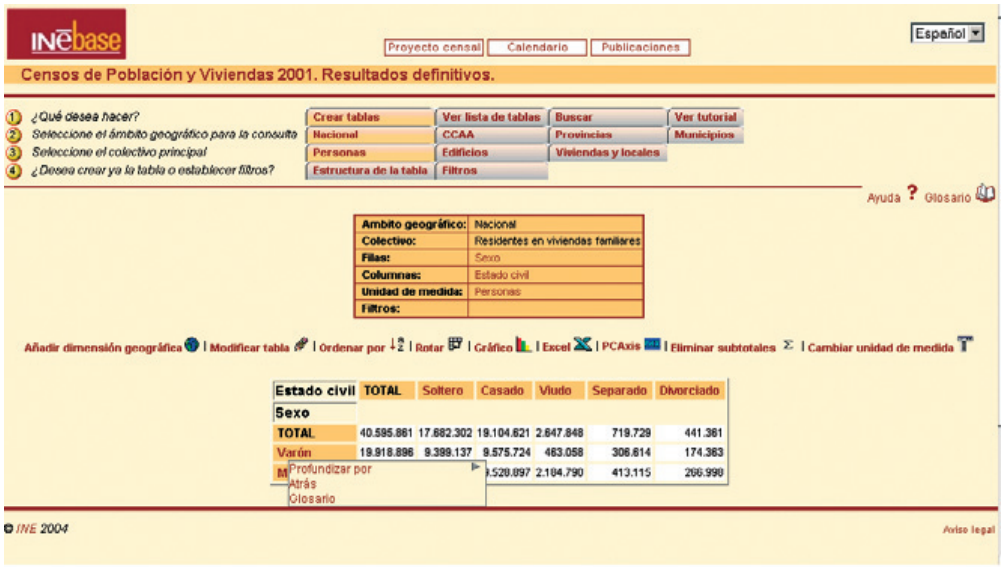
Básicamente el usuario realiza la elección del ámbito territorial que quiere consultar y del colectivo del que quiere obtener la información. En los censos de población y viviendas, se estudian además de las características de las personas, las de sus viviendas y edificios y las características de los hogares. Se incluye también información sobre los locales en la fecha de referencia censal.

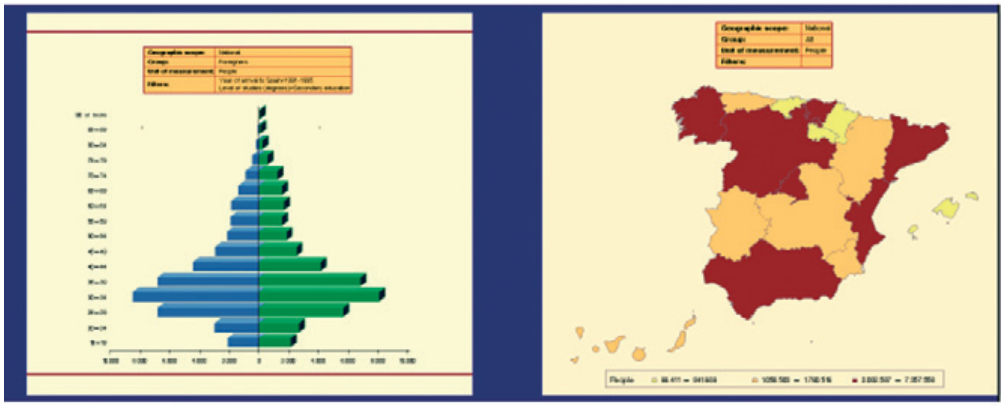
Como limitación a las posibles consultas se permiten un máximo de tres variables en la cabecera del informe y otras tres variables en el ladillo del mismo. Esta limitación se establece en orden a la visualización de los datos de la tabla y no a restricciones técnicas de la solución.

En una consulta, el usuario puede introducir condiciones sobre la información presentada en base al conjunto de variables presentadas u otro diferente.

Existen excepciones a esta regla general que son consecuencia de las medidas adoptadas para la protección de datos individuales de los informantes a la que el INE está obligado por la Ley de la Función Estadística Pública.

Una vez obtenida la consulta realizada, el usuario tiene distintas opciones que le permiten exportar la información a distintos formatos, cambiar el orden de las variables, cambiar la unidad de medida de la tabla, representar la información obtenida en gráficos estadísticos o mapas… y además la interfaz proporciona capacidades de navegar por la información presentada que permite el análisis particular del conjunto de información presentado. Estas capacidades se presentan como enlaces en los elementos de cabecera de las columnas y filas del informe. Permiten profundizar en una característica determinada por cualquier variable de interés para el usuario o cambiar el detalle de las variables consultadas...





En las consultas se ofrece como unidad básica la frecuencia absoluta -número de personas, vivivendas, … - que corresponden al cruce de variables presentado. Sin embargo es posible mostrar la tabla en porcentajes respecto a las filas o columnas o respecto al total además de un conjunto de unidades de medida de interés: tasa de ocupación, edad media, superficie media de las viviendas, condición socioeconómica media. Estas medidas permiten analizar desde distintas perspectivas una misma consulta.

Señalados ya los aspectos generales y funcionales de la interfaz se describen ahora unas cuantas soluciones adoptadas que constituyen diferencias con las consultas generales:

1- Límites a las consultas en función del tamaño:

1.1- Consideración sobre el tamaño máximo de una consulta

El INE presta un servicio al público y el recurso del que se dispone para cualquier proyecto es limitado. Es por tanto necesario garantizar que el canal de servicio está abierto y no es utilizado o aprovechado en exclusiva por algún ciudadano. Con el fin de evitar una sobreexplotación del sistema por parte de un ciudadano en detri-

mento de otros usuarios de la aplicación, las consultas que devolverían más de un millón de celdas no son procesadas.

La estimación del tamaño de la respuesta se realiza mediante un evaluador que considera la cardinalidad de las variables implicadas y los filtros existentes. Si la consulta pasa los límites establecidos se recibe un mensaje de advertencia al respecto.

1.2- Consideración sobre el tamaño máximo de una consulta que puede servirse en línea:

Con un doble fin se establece el límite de las consultas que se proporcionan en línea en 10000 celdas de información. Por una parte la capacidad de proceso del sistema no debe verse afectada en las horas de máxima concurrencia de usuario por el hecho de que las consultas que éstos efectúan supongan una mayor carga en cada una de ellas. Por otra parte, el coste en tiempo de comunicación de una consulta con un gran número de celdas es muy alto para los usuarios y puede hacerlos desistir por lo que resulta más conveniente procesar en una ventana de ejecución diferida todas estas consultas y permitir al usuario descargarlas en un fichero en el momento que crea más conveniente. Se dota al sistema de un servidor para la transferencia de estos ficheros y se controla el tiempo en que éstos permanecen almacenados para garantizar la disponibilidad de espacio.

2- Límites a las consultas en función de su complejidad

El sistema desarrollado permite asignar un coste a cada consulta en función del tamaño del objeto en el cual ésta ha de resolverse así como de las condiciones que conlleva en su definición. Este coste se asimila al tiempo en que la consulta va a poder ser resuelta en el sistema de almacenamiento y en función de él la consulta se resolverá en línea o avisará al usuario de su disponibilidad tras la ejecución en el sistema.

3- Límites establecidos para la protección del secreto estadístico

No toda la información puede obtenerse al mismo nivel de detalle. En función del tamaño territorial mínimo elegido en la consulta el sistema ofrece distinto nivel de detalle de la información y además se limita el número de variables a consultar.

4- Otros elementos de la interfaz

4.1- Consultas a medida.

La información se ha estructurado en colectivos de interés para la investigación de características sociodemográficas y el sistema ofrece para cada colectivo un conjunto de variables relevantes para el mismo (no se ofrece, por ejemplo, las características de los desplazamientos al trabajo o lugar de estudio en el colectivo de toda la población sino en los correspondientes a estudiantes y ocupados). Sin embargo, existe la posibilidad de que algunas de las consultas no queden suficientemente cubiertas con este enfoque. En un segundo nivel de la interfaz se ofrecen todas las variables de cada colectivo para que el usuario pueda detallar su consulta.

4.2- Tablas predefinidas.

Para dar continuidad a la forma de interactuar con el sistema que los usuarios del INE ya conocen se han extraido del sistema un conjunto numeroso de tablas pre-

definidas que constituyen también un elemento importante del sistema de difusión. Estas tablas se organizan jerárquicamente lo que permite localizar fácilmente la información que el usuario quiere obtener. Además existe un motor de búsquedas que identifica las tablas que contienen determinada información de interés.

4.3- Registro de usuarios

Mediante el registro, los usuarios pueden disfrutar de funcionalidad añadida. Pueden almacenar sus consultas, definir sus propias agregaciones en los datos o guardar filtros o condiciones en la información presentada que pueden utilizar de forma repetida. El registro es gratuito y se requiere del usuario la información mínima necesaria para garantizar el funcionamiento correcto de los servicios ofrecidos.

3.2 Modelo de datos

3.2.1 *Características*

El modelo de datos que soporta el sistema de difusión censal tiene características particulares que marcan la diferencia con entornos habituales de DW. Se señalan a continuación algunas de estas características y sus implicaciones en distinto sentido a la hora del diseño y carga del modelo de datos.

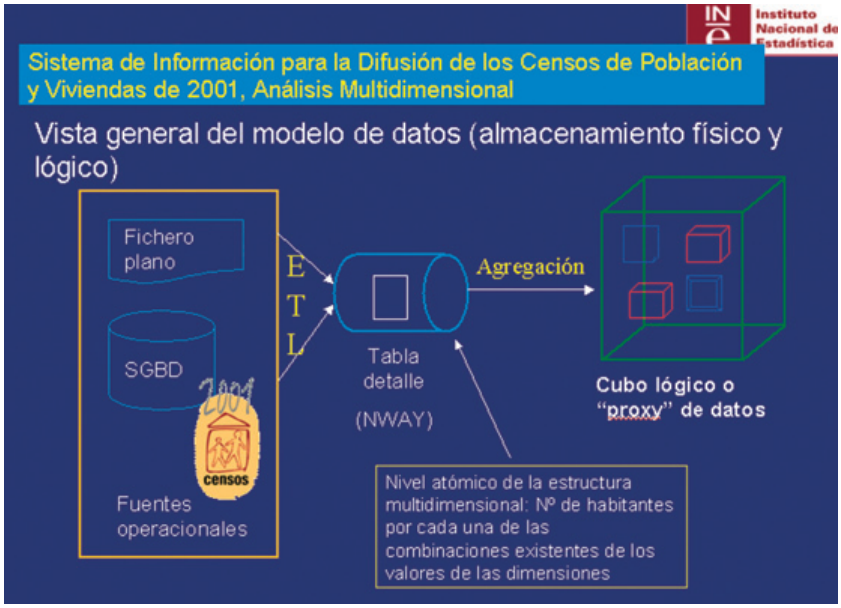
- Fuente de datos única que contiene en ficheros planos los datos de difusión de la operación censal. Esto nos permite ahorrar esfuerzos de integración de distintas fuentes de información.
- No se producen actualizaciones periódicas de la información. Los procesos de extracción, transformación y carga de la información no son críticos en cuanto a tiempo de ejecución dado que sólo se ejecutan una vez.
- El gran número de dimensiones consideradas así como el detalle y volumen de información inicial y el hecho de que la interfaz de consulta no establezca prioridades en la selección de variables o métricas hace que los procesos de agregación de la información para dar satisfacción a las posibles consultas sean muy pesados
- El sistema de consulta a través de internet acarrea también que se eleve el número de procesos de agregación para conseguir tiempos de respuesta adecuados en el canal de difusión elegido. Como consecuencia el almacenamiento requerido es superior.

La tecnología SAS elegida permite establecer un modelo de datos híbrido HOLAP que puede trabajar tanto con estructuras relacionales como multidimensionales. El entorno de diseño del modelo de datos viene dado por la herramienta Warehouse Administrator que, en un entorno gráfico e intuitivo, permite tanto la definición de las estructuras de datos como la especificación de los procesos ETL que las alimentan. Esta herramienta además genera el código necesario para la creación de los diferentes objetos, los procesos de carga y proporciona una completa documentación del sistema.

En el diseño del modelo se utilizan técnicas de particionamiento por atributos y técnicas de análisis de cardinalidad para determinar el nivel de agregación adecuado al objetivo de rendimiento con las restricciones de almacenamiento que se contemplan. De modo genérico estas téc-

nicas consisten en realizar agregaciones de cruces de variables en aquellos casos en que el resultado de la agregación reduzca significativamente el tamaño del conjunto a partir del cual dicha agregación es calculada.

El componente OLAP SERVER se encarga de actuar como un proxy para las consultas presentando a la capa de aplicación un único objeto multidimensional virtual –cubo lógico- al que se dirigen todas las peticiones.



En la implementación de este modelo, dado el volumen de información y el hecho de que no todas las variables tienen sentido en todos los casos (ej. tiempo de desplazamiento al trabajo/estudio en personas que no trabajan y/o estudian) se adoptan las siguientes decisiones:

1- Establecer diferentes conjuntos lógicos de información (se tratan como diferentes estructuras HOLAP en terminología DW) que corresponden a los diferentes colectivos/subcolectivos mostrados en la interfaz.
2- El gran número de dimensiones consideradas (en algunos colectivos más de 300 variables) hace inviable el precálculo de todas las posibilidades de cruce de información entre ellas. En su lugar se ha optado por realizar una agrupación previa de variables en función de su semántica y posibles relaciones jerárquicas y proceder al cálculo de agregados a partir de dichos grupos.

La reducción del número de agregados a calcular se traduce en un menor tiempo para los procesos de agregación aunque como contrapartida dispondremos de una agregación más burda de lo que llegaríamos a obtener de haber considerado las variables iniciales.

Esta agrupación es revisada tras la primera agregación en aras a conseguir mejores rendimientos del sistema. El cálculo de agregados se realiza así, de forma evolutiva, reduciendo el tiempo de despliegue desde que se dispone del conjunto de datos para la difusión y teniendo presente el afán de mejora continua de servicio. Se han desarrollado como complemento ele-

mentos de monitorización de rendimiento del sistema para facilitar el análisis de su comportamiento.

3.2.2- *Componentes tecnológicos*

Toda la arquitectura software sobre la que se apoya el modelo de datos construido es de SAS Institute. Aunque algunos aparecen citados más arriba componentes más relevantes son:

- SPDS: Motor de base de datos de altas capacidades. Permite particionamiento de objetos y proceso
- OLAP SERVER : Proporcionan buenos rendimientos en consultas a datos
- Warehouse Administrator: Herramienta de desarrollo del modelo de datos
- SAS/ACCESS: Permite el acceso a distintas fuentes de datos para su integración en el DW
- SAS/BASE: Núcleo del sistema SAS.

La plataforma sobre la que está instalado este modelo de datos se describe a continuación:

- Dos servidores SunFire 450 con configuración de cluster (Veritas) activo/pasivo y sistema operativo Solaris 2.8
- Sistema de almacenamiento compartido de EMC. Espacio de almacenamiento 2 TB.

Las comunicaciones se integran en el nodo internet del INE.