El proyecto del Diccionario de Datos de la Gerencia de Informática de la Seguridad Social

Jorge Manrubia Díez, Eugenio Bezares Ruíz

Introducción

El proyecto del Diccionario de Datos es una iniciativa de la Gerencia de Informática de la Seguridad Social (GISS) que tiene como objetivo crear un modelo de conocimiento corporativo de la información manejada por las aplicaciones informáticas desarrolladas en la organización. El proyecto se enmarca dentro de las iniciativas de normalización llevadas a cabo por el Centro de Coordinación de la GISS, y facilita la adaptación de los sistemas de información a la ley 11/2007, en lo que respecta a la interoperabilidad y el intercambio de información electrónica entre diferentes órganos de la Administración Pública.

Para entender la motivación del proyecto hay que remontarse al año 2001, cuando se pone en marcha una iniciativa de descentralización de los desarrollos en la organización informática de la Seguridad Social. Esta iniciativa propone acercar los desarrollos, hasta entonces centralizados en la propia GISS, a su negocio de gestión. En el nuevo esquema, los Centros de Desarrollo de las Entidades Gestoras de la Seguridad Social pasan a depender orgánicamente de la GISS, donde se crea un nuevo Centro denominado Centro de Coordinación[1], que tiene como misión ejercer labores de coordinación entre los centros y de soporte al desarrollo.

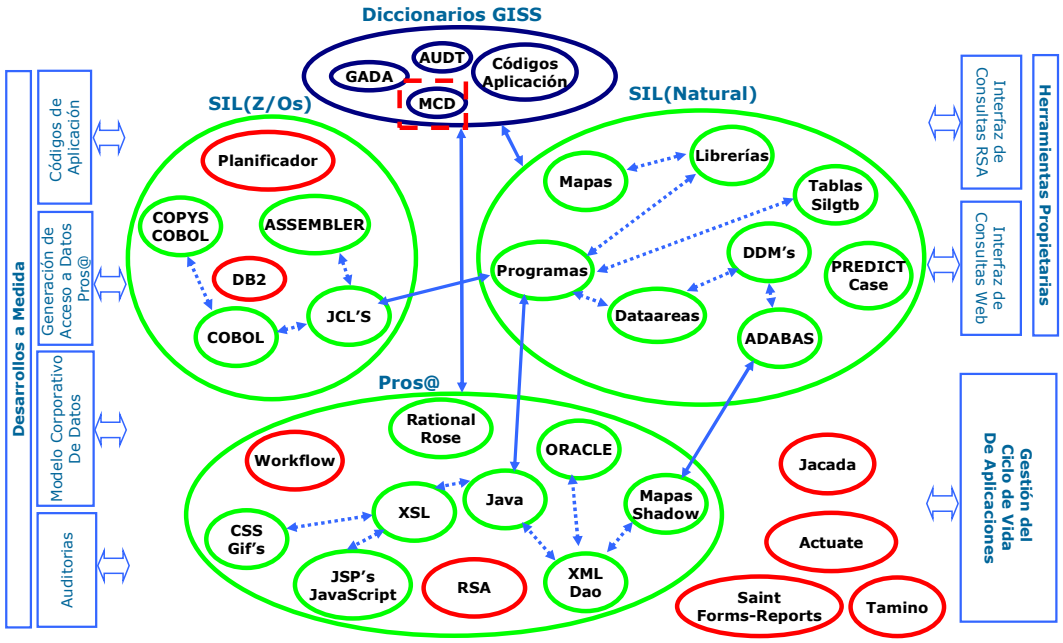

Figura 1. Repositorio Corporativo de la Seguridad Social

[1] Anteriormente denominado Centro de Proyectos

Como consecuencia de esta iniciativa, el Centro de Coordinación detecta la necesidad de un registro central de información sobre los diferentes activos informáticos manejados por los desarrollos. Para ello se inició el proyecto denominado Repositorio Corporativo de la GISS. Su objetivo es elaborar un almacén estructurado de metadatos sobre los activos informáticos utilizados por los desarrollos. Así, utilizando como soporte el producto Rochade de ASG, se creó una infraestructura para alimentar un almacén central de metainformación sobre multitud de artefactos: código fuente Java, estructuras PREDICT de ADABAS, esquemas Oracle, librerías NATURAL, etc (ver Figura 1. Repositorio Corporativo de la Seguridad Social).

El proyecto del Diccionario de Datos tiene como misión integrar en el Repositorio Corporativo un activo de especial relevancia: los datos manejados por las aplicaciones. Para entender la motivación del proyecto hay que entender el volumen de información manejada por los sistemas de la Seguridad Social. Tomando como métrica el número total de tablas gestionadas, existen más de 50.000 tablas Oracle y más de 2.000 ficheros ADABAS.

Este volumen de información manejado por diferentes Centros de Desarrollo crea una serie de necesidades en los procesos de desarrollo de la organización:

- **Necesidad de normalización**. Cuando en un desarrollo se desea manejar un tipo de dato, ¿qué formato se le debe asignar? El manejar la misma información con diferentes formatos dificulta la interoperabilidad de las aplicaciones entre sí.
- **Necesidad de reutilización**. Cuando en un desarrollo se desea manejar una información ¿cómo puedo saber dónde puedo consultar dicha información? Ante la ausencia de este mecanismo por lo que se opta es por duplicar los datos, creando un problema de calidad de la información.
- **Necesidad de consulta y trazabilidad de la información**. Cuando se produce un cambio que afecta a un tipo de datos (por ejemplo, el Código de Cuenta CCC), ¿cómo saber qué aplicaciones hay que modificar? ¿qué estructuras físicas se verán afectadas?

El proyecto del Diccionario de Datos busca satisfacer estas necesidades a través de un conjunto de líneas de actuación que tienen una doble dimensión:

- **Dimensión organizativa**. La implantación de una cultura de administración de datos en los Centros de Desarrollo de la GISS.
- **Dimensión técnica**. La elaboración de un almacén estructurado de metadatos y de las herramientas que permiten su alimentación y explotación.

Los destinatarios del proyecto son, en primer lugar, los propios Centros de Desarrollo. Ofrece a los desarrolladores herramientas para poder consultar el Diccionario de Datos y elaborar los modelos de sus aplicaciones teniendo una visión global de los datos manejados por la organización. Es también beneficiario cualquier actor interesado en explotar la información agregada sobre los datos de las aplicaciones (por ejemplo, para la realización de informes sobre calidad y auditorías). Además, cabe señalar como beneficiaria del proyecto a la propia Organización, que se dota de una herramienta para transformar el conocimiento acerca de la información que maneja en verdadero conocimiento corporativo.

En esta comunicación se abordará, primero, una descripción de alto nivel del proyecto. A continuación se abordarán los aspectos organizativos y técnicos del mismo.

Descripción del Proyecto

Descripción funcional

Puede pensarse en el Diccionario de Datos como un almacén central de información. Dicha información consiste en metadatos acerca de los datos que gestionan las aplicaciones. Existen una serie de actores que explotan el Diccionario consumiendo y cargando información. Aunque sus responsabilidades se verán con detalle más adelante, a continuación se resumen las funcionalidades que les ofrece el proyecto (Figura 2):

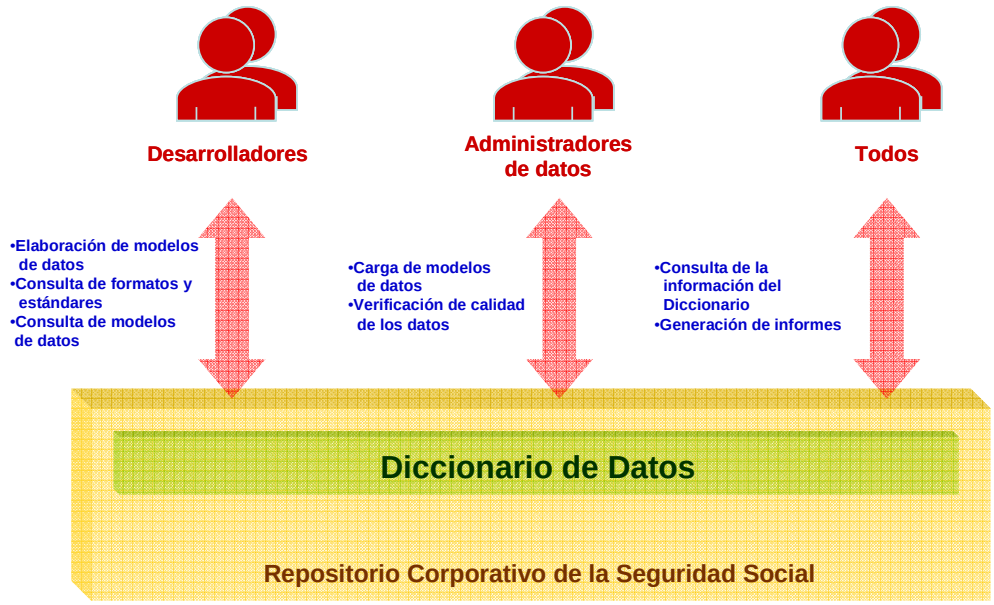

Figura 2. Descripción funcional del proyecto

- Los **desarrolladores** se apoyan en el Diccionario de Datos para elaborar sus modelos de datos. De una forma integrada con el entorno de desarrollo, pueden elaborar los modelos de datos de las aplicaciones apoyándose en la información existente en el Diccionario. Por ejemplo, pueden consultar los dominios estándares de la organización y utilizarlos en sus modelos. O pueden consultar los modelos de datos de cualquier aplicación que se haya documentado en el Diccionario para reutilizar sus artefactos.
- Los **administradores de datos** se encuentran en los Centros de Desarrollo y tienen la misión de velar por la calidad de la documentación de los modelos de datos. En última instancia, son los responsables de cargar los modelos en el Diccionario de Datos.
- Finalmente cabe señalar que el acceso en modo consulta al Diccionario de Datos es **público**. Su información está disponible para cualquier actor interesado.

Estructura

El proyecto del Diccionario de Datos construye un almacén de metadatos sobre la información manejada por las aplicaciones de la GISS. Nótese que se habla de metadatos, esto es, datos acerca de los datos (no se almacenan los datos en sí mismos). El almacén presenta una arquitectura en capas tal y cómo se recoge en la Figura 3.

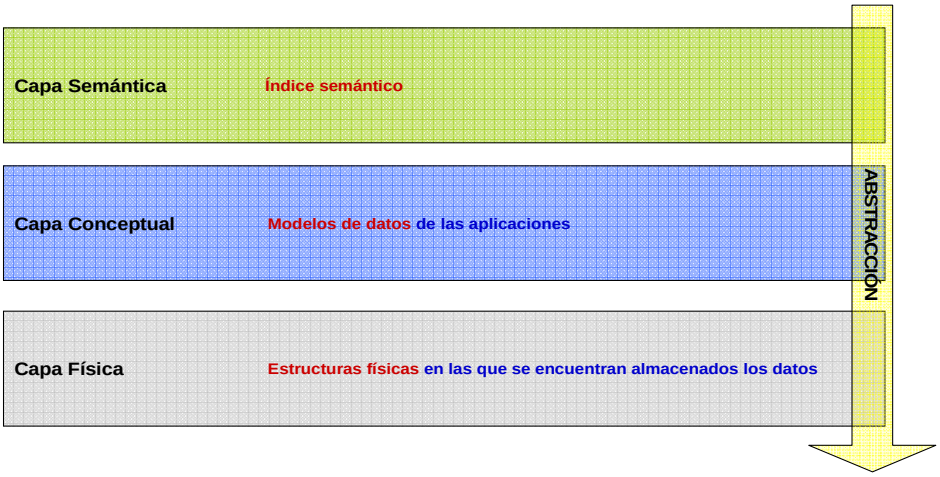

Figura 3. Estructura del Diccionario de Datos

Cada capa contiene un tipo de metadatos que caracteriza su naturaleza:

- **Capa Física**. Contiene los *esquemas de las tablas y ficheros físicos* de los Sistemas de Gestión de Bases de Datos de la Seguridad Social (ADABAS y Oracle).
- **Capa Conceptual**. Contiene los *modelos lógicos de datos* de las aplicaciones. Éstos vienen dados por los modelos Entidad-Relación (E-R) utilizados por los sistemas informáticos de la Gestión de la Seguridad Social. Estos modelos representan la información gestionada por la aplicación desde un nivel mayor de abstracción que el nivel físico. Los elementos que se almacenan en esta capa son entidades, atributos, relaciones entre entidades, dominios, etc.
- **Capa Semántica**. Contiene un *índice de términos y conceptos* que permiten acceder a los elementos de las capas inferiores relacionados con los mismos.

Debe señalarse que las diferentes capas están enlazadas entre sí. La trazabilidad de información entre capas es un aspecto esencial de cara a explotar la información del Diccionario de Datos como conocimiento corporativo. En la Figura 4 se recoge un ejemplo de dicha trazabilidad que sirve además para ilustrar el tipo de elementos que se almacenan en las capas del Diccionario de Datos.

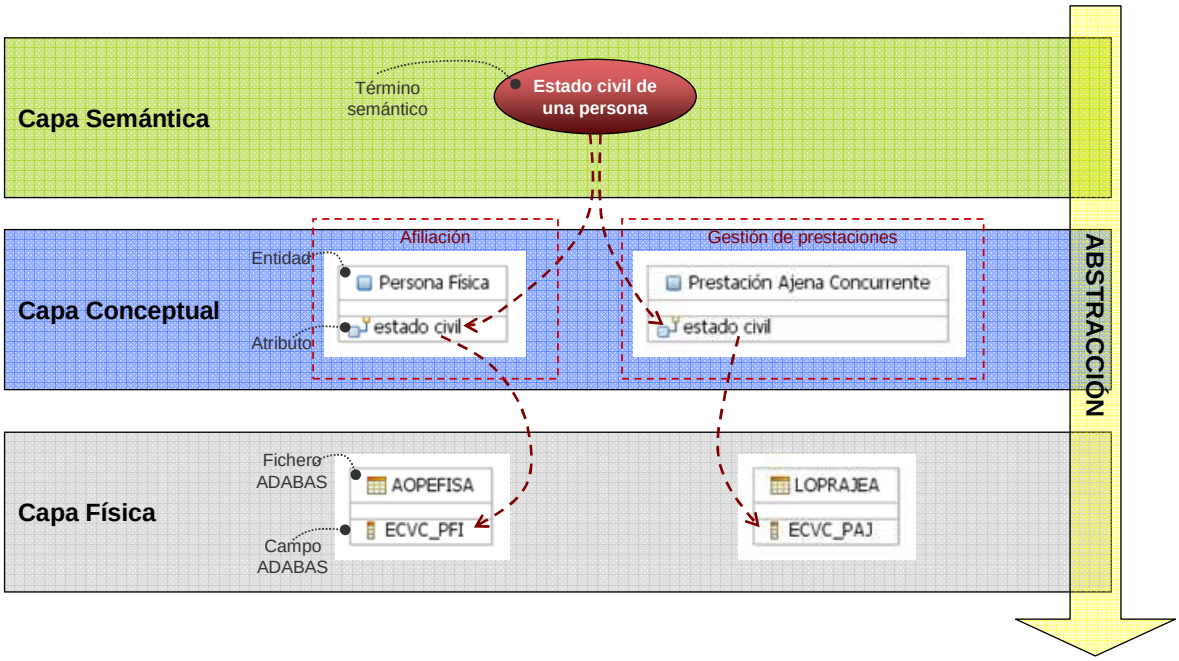

Figura 4. Ejemplo de trazabilidad de metadatos en el Diccionario de Datos

Partiendo de la parte superior de la figura, se observa como en la capa semántica se aloja el concepto de "Estado civil de una persona". Este concepto se encuentra enlazado a dos atributos lógicos de la capa conceptual, que se corresponden con sendas entidades de dos modelos de datos de dos aplicaciones (Afiliación y Gestión de Prestaciones). Si se sigue bajando, de los atributos conceptuales se llega al campo ADABAS del correspondiente fichero ADABAS en la capa física.

Herramientas

Herramienta de modelado de Datos

El entorno de modelado de datos utilizado es *Rational Data Architect* de IBM. Se trata de un entorno basado en Eclipse que ofrece herramientas de modelado de datos y facilidades de administración de bases de datos. En su instalación puede integrarse con *Rational Software Architect* (RSA), también de IBM, que es el entorno de desarrollo Java EE con modelado UML utilizado en la GISS.

RDA permite la elaboración de diferentes modelos relacionados con los datos. La estrategia seguida en el proyecto es extender esta plataforma para ofrecer soporte directo para que dichos modelos puedan ser cargados y leídos del Diccionario de Datos desde el mismo entorno de desarrollo:

- **Modelos lógicos de Datos**. Son modelos E-R extendidos. Se corresponden con los modelos cargados y leídos de la capa conceptual del Diccionario de Datos.
- **Modelos de dominio de Datos**. Definen de forma específica dominios de datos. Un dominio de datos es un tipo de dato concreto, junto con su documentación y restricciones. Por ejemplo, el domino Fecha en formato AAAAMMDD.
- **Modelos físicos de Datos**. Representan el diseño físico de los datos en la Base de Datos.

- **Modelos de correlación de Datos**. Estos modelos permiten especificar el mapeo de modelos lógicos a modelos físicos. Se utilizan en el Diccionario para especificar la trazabilidad de los modelos entre las capas conceptual y física.

RDA ofrece facilidades para trabajar y transformar los modelos que soporta. Por ejemplo, puede hacer ingeniería inversa desde una base de datos existente, elaborar su modelo físico y, a partir de él, generar el modelo lógico correspondiente. Integrado con RSA ofrece, además, transformaciones automáticas entre modelos UML y modelos de datos.

Además, por estar basado en Eclipse se obtiene toda la potencia de la arquitectura de Eclipse para su extensión a través de Plugins. Esta característica se ha aprovechado para implementar diversas extensiones. Por ejemplo, se ha implementado un proceso específico de ingeniería inversa de modelos ADABAS, que permite generar modelos lógicos con una estructura relacional normalizada y con nombres lógicos de negocio a partir la documentación Predict de ADABAS que se encuentra e el Repositorio Corporativo.

Herramienta de carga

En lo relativo a la alimentación de las diferentes capas del Diccionario de Datos, la información de la capa física se carga automáticamente. El Repositorio Corporativo dispone de escáneres que, periódicamente, analizan la información de las Bases de Datos ADABAS y Oracle y actualizan los metadatos de sus estructuras físicas automáticamente.

En cuanto a la capa conceptual, se carga utilizando un sistema de implementación propia. Los procesos de carga pueden lanzarse desde el mismo entorno RDA si se disponen de los permisos adecuados.

La capa semántica se alimenta automáticamente al indexar los modelos cargados en la capa conceptual. Además, se permite la administración centralizada de sus metadatos a través de un interfaz Web para que un experto humano pueda dotar de mayor calidad a la misma.

Herramientas de consulta

En lo relativo a la consulta del Diccionario, se dispone de dos mecanismos:

- Desde el entorno de RDA puede consultarse cualquier modelo cargado en el Diccionario de Datos, así como los dominios y formatos estándares de la organización.
- También puede utilizarse el interfaz Web general que ofrece el Repositorio Corporativo. Este permite consultar la información del Diccionario de Datos, así como generar informes en diferentes formatos.

En la Figura 5 se recoge el esquema de herramientas disponibles.

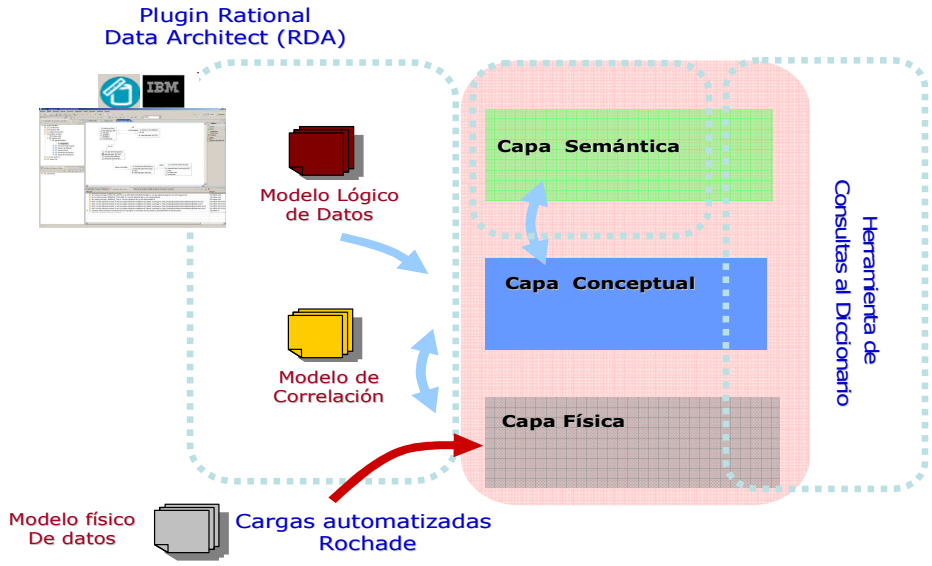

Figura 5. Esquema de herramientas disponibles

Aspectos organizativos

Debido al tamaño de la informática de la GISS, se descartó una estructura de administración de datos centralizada y se apostó por una estructura distribuida. Además, se considera que donde mejor se puede realizar la administración de los modelos de información de las aplicaciones es precisamente allí donde se desarrollan, puesto que la calidad de los modelos de información está íntimamente relacionada con el conocimiento del negocio al que se refieran.

Participantes y roles

Los participantes del proceso son:

- **Los centros de desarrollo de la GISS** Estos Centros tendrán al menos un Administrador de Datos Delegado, responsable del Diccionario
- **El Área de Diccionarios del Centro de Coordinación**, donde se encuentran los Administradores de Datos Centrales.

Administradores de Datos Delegados

Responsables de administración de datos en cada centro de desarrollo. Gestionan la correcta documentación en el Diccionario de Datos de la información manejada por las aplicaciones desarrolladas en el centro. Sus funciones son:

- Cargar y gestionar la documentación de los datos en el Diccionario de Datos, utilizando para ello las herramientas de administración de que ofrece el Centro de Coordinación.
- Verificar que la información gestionada por las aplicaciones está correctamente documentada en el diccionario de datos.
- Velar por la calidad de los modelos y diseños de datos: verificar que los datos están correctamente modelados; que se utilizan formatos estándar para los datos; que no se duplica información ya existente, etc.

Administradores de Datos Centrales (Área de Diccionarios)

Responsables de administración de datos en la GISS. Supervisan la correcta carga, gestión y explotación del Diccionario de Datos. Sus funciones son:

- Verificar que los datos manejados por las aplicaciones están correctamente documentados en el Diccionario de Datos, como paso previo a su entrada en preexplotación.
- Velar por la calidad de la documentación semántica de los datos.
- Ofrecer asistencia a los administradores de datos delegados en la carga del diccionario de datos y en la utilización de las herramientas de administración de datos.
- Ofrecer asistencia al resto de unidades de la GISS para la explotación de la información almacenada en el Diccionario de Datos.

Procesos

En los siguientes apartados se recoge el flujo de trabajo de la administración de datos en los desarrollos de la GISS contemplando dos escenarios: la carga inicial del diccionario de datos y su mantenimiento.

Proceso de carga inicial

El proceso de carga inicial tiene como objeto cargar en el Diccionario de Datos los modelos de información de las aplicaciones existentes en los sistemas informáticos de la Seguridad Social.

El primer paso consiste en la identificación y catalogación de las aplicaciones informáticas. Para ello se utiliza una aplicación Web denominada "Censo de Aplicaciones", también desarrollada en el Centro de Coordinación de la GISS. A partir de este punto, cada aplicación llevara su propio proceso de carga inicial.

El siguiente paso consiste en elaborar el modelo lógico y de correlación con el físico. Este proceso se agiliza enormemente con las funcionalidades de ingeniería inversa que ofrece RDA.

En la Figura 6 se recoge un resumen de las tareas que deben realizarse para el proceso de carga inicial.

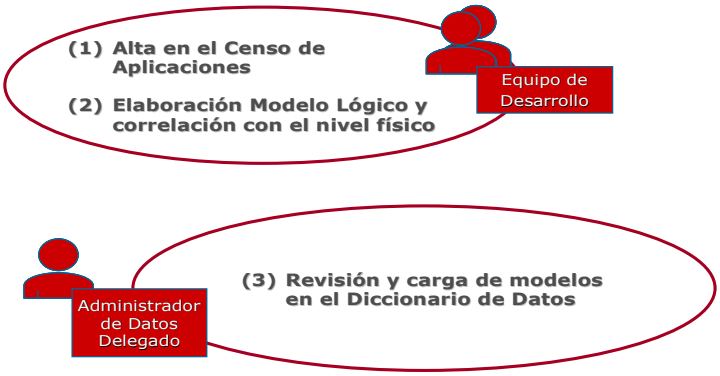

Figura 6. Proceso de carga inicial del Diccionario de Datos

Proceso de mantenimiento

La característica esencial que se exige a los desarrollos es que, como paso previo a la entrada en producción de las aplicaciones, éstas tengan sus datos correctamente documentados en el Diccionario de Datos.

En concreto, se considera esencial que las unidades de administración de datos delegadas aprueben la creación de nuevas estructuras físicas. Esto es necesario porque de lo contrario, el enorme número de desarrollos que acomete la GISS produce inevitablemente una proliferación de duplicidades, derivadas de la decisión de crear nuevas estructuras físicas antes de reutilizar las existentes.

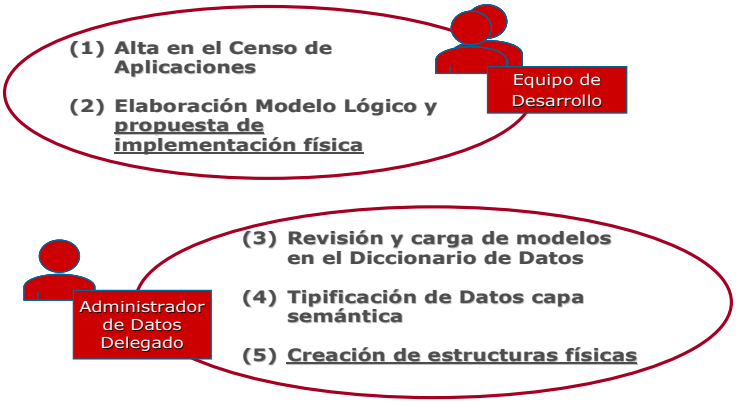

Figura 7. Proceso de mantenimiento en el Diccionario de Datos

Arquitectura técnica

La Figura 8 recoge los principales componentes que conforman la arquitectura del Diccionario de Datos, que son descritos en los siguientes apartados.

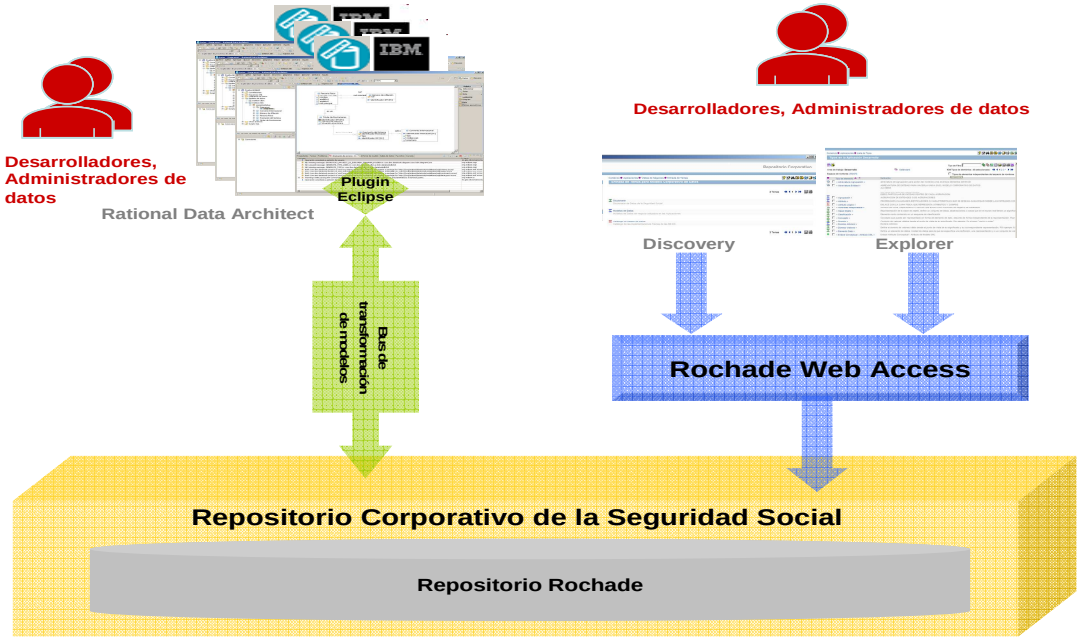

Figura 8. Principales componentes del Diccionario de Datos

Repositorio Rochade de ASG

El Repositorio Corporativo se implementa sobre ASG Rochade. Se trata de una implementación de un repositorio de metadatos para grandes organizaciones. Permite cargar un número arbitrario de tipos de metadatos facilitando el enlace o trazabilidad entre los mismos. Diversos fabricantes ofrecen soluciones para alimentar el repositorio a partir de diferentes orígenes. A estas soluciones se les denomina buses o escáneres y se ha hecho uso de varios de ellos para cargar metainformación de multitud de activos de la GISS: código fuente Java, estructuras PREDICT de ADABAS, esquemas Oracle, librerías NATURAL, etc. En la Figura 1 se recoge un esquema general del Repositorio Corporativo.

Por otra parte, Rochade ofrece dos aplicaciones Web que permiten consultar y gestionar la información del repositorio, una de carácter general denominada *Explorer*, y otra que permite la construcción de interfaces Web parametrizadas, denominada *Discovery*. Sobre esta última se ha elaborado un interfaz de consulta específico para el Diccionario de Datos.

Bus de modelos EMF

En el piloto inicial del Diccionario de Datos los modelos eran especificados mediante Hojas Excel. Este sistema era poco productivo porque la creación de modelos se convertía en una tarea muy laboriosa. Por ello, se decidió que el sistema de carga debía dar soporte directo a los modelos elaborados con herramientas de modelado profesionales.

Se decidió apostar por la tecnología EMF (*Eclipse Modeling Framework*). Se trata de un lenguaje de especificación de metamodelos. Los metamodelos permiten definir formalmente la estructura formal de un lenguaje. En la actualidad EMF se trata de la implementación *de facto* de la especificación MOF (Meta-Object Facility) del OMG. Por ejemplo, el proyecto UML2 que utiliza RSA es una implementación del lenguaje UML utilizando EMF. Del mismo modo, LDM es la especificación de un lenguaje E-R para el entorno RDA basado en EMF. Hay que señalar que al utilizar EMF se dispone de toda una infraestructura para procesar modelos de forma programática mediante la plataforma Java.

Para la carga y lectura de modelos desde RDA se diseñó un sistema de transformación de modelos. Dicho sistema se basa en el refinamiento sucesivo de modelos utilizando un motor de transformación de modelos (ATL). Un motor de transformaciones permite, dados un metamodelo de entrada y un metamodelo de salida, especificar cómo se mapean los elementos de ambos metamodelos. Con esta especificación, el motor es capaz de transformar instancias del metamodelo de entrada en instancias del metamodelo de salida.

Lo que se hizo fue especificar formalmente el metamodelo de los modelos E-R soportados por el Diccionario de Datos. Esta decisión estuvo influenciada por el hecho de que el lenguaje LDM no es un estándar como UML, y no se deseaba estar cautivo con la notación de IBM. Así mismo, se creó el metamodelo EMF de la estructura interna de Rochade, así como un proceso que permitiese leer y escribir estos modelos en Rochade. Con estos componentes se construyó un bus donde, para cargar, los modelos de IBM son transformados en el metamodelo propio del Diccionario de Datos, luego son transformados al metamodelo de Rochade y finalmente se cargan. Para la lectura de modelos el sistema es el mismo, si bien el orden de las transformaciones es inverso. El esquema de carga puede verse en la Figura 8.

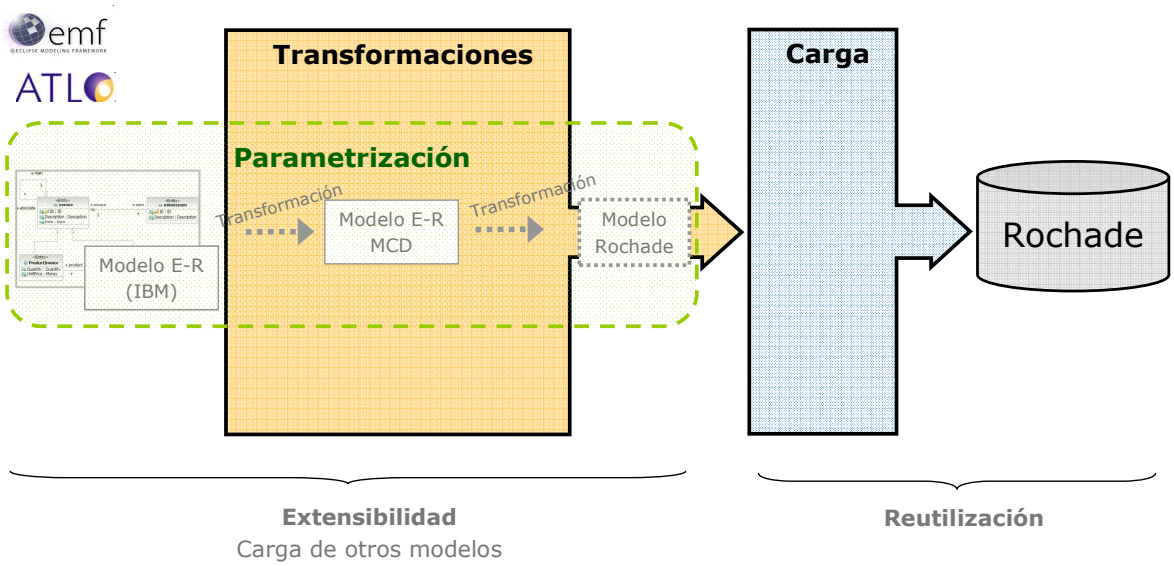

Figura 9. Sistema de transformación de modelos

El sistema diseñado es configurable y el mismo bus de carga se ha utilizado para cargar diferentes modelos (como los modelos de dominio, o los modelos de correlación de RDA).

Eclipse PDE

La herramienta de modelado Rational Data Architect está basada en Eclipse. Esto significa que puede extenderse utilizando las facilidades de extensión de la plataforma, mediante componentes ejecutables denominados Plugins. Eclipse constituye un contenedor de ejecución de aplicaciones Java realmente avanzado, donde su extensibilidad se manifiesta en su propia arquitectura interna. Se basa en una implementación de un contenedor OSGI (Equinox) sobre la que funciona un núcleo que, básicamente, permite desplegar y ejecutar plugins.

El principal reto que ha tenido el equipo de desarrollo para implementar plugins de Eclipse ha sido el desconocimiento inicial de la plataforma. No obstante, la alta calidad de la arquitectura de Eclipse hace que sea un entorno de desarrollo muy productivo una vez que se comprenden los fundamentos de la misma. El plugin del Diccionario de Datos contribuye una serie de asistentes, vistas, menús y acciones con las que ofrece las diversas funcionalidades comentadas.

Metodología de Desarrollo

El equipo de desarrollo ha adoptado SCRUM como metodología ágil de desarrollo. Así mismo, se han adoptado multitud de buenas prácticas de desarrollo y gestión de configuración, que incluyen la implementación exhaustiva de pruebas automáticas, buenas prácticas de diseño y codificación, refactorización de código y la utilización de un servidor de integración continua.

Este sistema ha arrojado muy buenos resultados permitiendo obtener una versión funcional de las herramientas desde etapas tempranas de desarrollo, pese al componente de incertidumbre técnica existente, y permitiendo la continua evolución y ampliación del software desarrollado.

Conclusiones

El proyecto del Diccionario de Datos surgió como solución a la necesidad de mantener la visión global sobre los datos que la GISS maneja en sus desarrollos. Se trata de un proyecto de gran alcance que ha presentado grandes retos tanto a nivel organizativo como técnico.

A nivel organizativo, el reto ha sido conseguir que los Centros de Desarrollo se involucren activamente con el proyecto, especialmente en lo relativo al proceso de carga inicial del Diccionario. Para que el Diccionario resulte útil tiene que alimentarse y esto significa la creación los modelos de datos de la multitud de aplicaciones existentes. Esta tarea exige recursos de los que no siempre disponen los equipos de desarrollo, cuyo día a día puede establecer otras prioridades.

En la actualidad se encuentran documentadas en el Diccionario de Datos más de 40 aplicaciones que abarcan a los 3 Centros de Desarrollo de la GISS y al propio Centro de Coordinación. Cabe destacar la reciente carga del corazón del sistema de Prestaciones del Centro de Desarrollo del Instituto Nacional de la Seguridad Social (INSS). A modo indicativo, el número de entidades conceptuales cargadas supera los 2.000 y el número de atributos los 14.000.

A nivel técnico, el proyecto se ha movido sobre cierta incertidumbre técnica debido a la novedad del desarrollo planteado. El hecho de que el almacén físico sea ASG Rochade dota de características muy particulares a la arquitectura del software desarrollado. Además, el procesamiento y transformación de modelos, y el desarrollo de plugins para Eclipse son áreas donde no existe mucho conocimiento en la industria, si bien ha sido solventado por la calidad del equipo de desarrollo que ha participado en el proyecto.

Finalmente, las futuras líneas de desarrollo del proyecto pasan por hacerlo útil para los equipos de desarrollo, desarrollando herramientas que no solo mejoren la calidad de sus desarrollos, sino también su productividad. El proyecto se encuentra actualmente en las últimas fases de integración con GADA, un generador de código acceso a datos para la arquitectura Pros@ de la GISS. Además, se está realizando una labor de estudio y catalogación de los dominios existentes en la organización, y se está trabajando para que los desarrolladores puedan utilizar directamente dicho conocimiento a la hora de elaborar sus modelos.

La experiencia adquirida por la Gerencia Informática de la Seguridad Social en el desarrollo y gestión de su Diccionario de Datos puede ser exportada y servir de base para afrontar la realización de un Diccionario de Datos más ambicioso y global, que aporte el conocimiento sobre la información de carácter común a todos los ámbitos de la Administración Pública.