tecnimap Murcia 2004

23

EII, UN NUEVO PARADIGMA PARA LA INTEGRACIÓN DE INFORMACIÓN DISPERSA Y HETEROGÉNEA EN LA ADMINISTRACIÓN. EXPERIENCIA: VIXÍA, UN SERVICIO DE VIGILANCIA TECNOLÓGICA

David Sánchez Sánchez
Director de Operaciones
Denodo Technologies

Alberto Pan Bermúdez
Investigador dentro del programa Ramón y Cajal
Universidad de La Coruña

Angel Viña Castiñeiras
Catedrático del Departamento de Electrónica y Sistemas
Universidad de La Coruña

1. INTRODUCCIÓN

Hoy en día, y para soportar todo tipo de procesos, las administraciones local, regional, estatal o europea, disponen de multitud de sistemas de información desarrollados a lo largo del tiempo y que hacen uso de distintas fuentes de información. Habitualmente, estas fuentes de información presentan esquemas de consulta muy variados, están desarrolladas sobre tecnologías diferentes (bases de datos relacionales, ficheros planos...) y de acuerdo a modelos de datos heterogéneos.

En Europa, con la actual situación de descentralización administrativa, se hace cada vez más necesaria la colaboración entre las distintas administraciones. Surge la necesidad de integrar los distintos sistemas de información existentes, con objeto de compartir datos entre ministerios, consejerías o concejalías sin perder las inversiones ya realizadas en ellos. Adicionalmente es preciso ofrecer servicios de consulta de información administrativa útil para el ciudadano, independientemente de la dispersión de esta información en diferentes organismos.

Otra importante tendencia a considerar es la expansión de Internet, ejemplo paradigmático de dispersión de información y heterogeneidad de formatos. Sus millones de páginas web, el inmenso mundo de información de Internet, son en muchas ocasiones un complemento fundamental para la información residente en los sistemas.

Claramente es necesario hacer espacio a tecnologías que faciliten el acceso eficiente y un óptimo aprovechamiento de la totalidad de la información disponible; es decir, que permitan su extracción en tiempo real, su filtrado flexible y su homogeneización en un modelo de datos unificado, listo para ser procesado por las nuevas aplicaciones.

Esta tendencia ineludible se apoya en un dato absolutamente revelador: en vez de desplegar nuevas tecnologías o extensiones de funcionalidad, un porcentaje muy importante de las inversiones en TI (hasta un 40% de su presupuesto total de TI) se está consumiendo actualmente en la integración de los sistemas y aplicaciones existentes.

A lo largo de la presente comunicación, pretendemos definir más claramente el problema, plantear las alternativas tecnológicas disponibles, y argumentar a favor de un nuevo paradigma llamado a jugar un papel decisivo en las arquitecturas software de la administración del futuro. Finalmente, presentaremos una experiencia de aplicación real de estas tecnologías en la Administración.

2. LAS FUENTES DE INFORMACIÓN

Para definir con exactitud el problema, es importante caracterizar adecuadamente las distintas fuentes de información. Una clasificación básica, atendiendo fundamentalmente a la flexibilidad del modelo de datos, debe diferenciar fuentes no estructuradas, fuentes estructuradas y fuentes semiestructuradas.

2.1. Fuentes No Estructuradas

Son aquellas fuentes que no presentan ningún tipo de esquema para la información que contienen (e.g. documentos en texto libre de cualquier tipo, Word, PDF, la gran mayoría de las páginas web estáticas, etc.). La naturaleza no estructurada de los datos contenidos en estos documentos complica enormemente su tratamiento automatizado.

Existen herramientas capaces de obtener información de este tipo de fuentes empleando técnicas de indexación y conceptualización basadas en estadística, inteligencia artificial o aprendizaje inducido. Pese a su complejidad, sólo permiten la realización de búsquedas por palabra clave o por concepto que, ordenadas según algún indicador de relevancia, son directamente utilizadas para presentar sus resultados (de forma más o menos completa y relacionada) al usuario final.

2.2. Fuentes Estructuradas

Son aquellas fuentes que presentan un esquema rígido bien definido, diferenciado de los datos. Un típico ejemplo de fuente estructurada es una base de datos relacional. Presenta un esquema, almacenado en el diccionario de datos, que define la organización interna de la información (en el caso de las bases de datos relacionales, las tablas y sus relaciones) y las restricciones a aplicar sobre los datos (e.g. posibilidad de existencia de valores nulos, rangos de valores, etc.).

Como es sobradamente conocido, el acceso automatizado a este tipo de fuentes es mediante potentes lenguajes de consulta (e.g. SQL). En la actualidad destaca el profundo interés dedicado a los temas relacionados con su integración.

2.3. FUENTES SEMIESTRUCTURADAS

Son aquellas fuentes que no presentan un esquema rígido. Esto quiere decir que el esquema es implícito y está contenido en (o puede deducirse de) los propios datos (*self-describing*). O que aún habiendo un esquema, éste es muy vago y permite cierta flexibilidad (sobre los tipos de datos, las restricciones...).

Las fuentes semiestructuradas están ganando una importancia creciente en la actualidad. Además de ficheros XML, Excel, PDF, los logs emitidos por muchos sistemas o equipos de comunicaciones, o infinidad de documentos con información tabulada, destacan, por el volumen de información y servicios volcados en Internet, las páginas web obtenidas como resultado de algún tipo de consulta o navegación.

Como ejemplo de automatización del acceso a información ofrecida de forma semiestructurada, utilizaremos los resultados de una consulta de búsqueda de libros realizada en un catálogo electrónico. Se trata de una página HTML con información adicional (publicidad, logos, etc.) en la que aparece insertada la lista de libros obtenidos. La información de cada uno de los libros se presenta con un formato similar (como se puede observar en la figura) que es el que define el esquema implícito.

Sin embargo, se pueden apreciar diferencias en el conjunto de datos que aparecen para un item determinado (como por ejemplo que para algunos libros aparece un descuento asociado ó una fotografía de la portada del libro y en otros casos no) o el formato de presentación de resultados (el autor puede aparecer como "Apellidos, Nombre" o como "Nombre Apellidos"). Es posible que cierta información adicional se pueda obtener accediendo a otra página activando el enlace del libro, obligando a una navegación extra. También es posible que sencillamente esa información no esté disponible para algunos items.

Figura 1 Resultado Típico de una Consulta a una Catálogo Web

Una herramienta que automatice las búsquedas en este tipo de fuentes debe realizar esta navegación a través de páginas web, lo cual también incorpora cierta complejidad debido a que:

- Muchas de estas páginas son dinámicas y se obtienen rellenando formularios.
- Muchas de las páginas incorporan control de acceso, de modo que es necesario autenticarse antes de acceder a la información.
- Muchas de estas páginas manejan *cookies* ó identificadores de sesión que se deben mantener durante toda la secuencia de navegación.

Además, la flexibilidad resulta vital, puesto que frecuentemente estas fuentes web cambian el esquema implícito. Es decir, actualizan el *look and feel* de sus páginas, incluyen más información en los items proporcionados, cambian las secuencias de navegación, etc.

Otros problemas añadidos de la consulta sobre fuentes semiestructuradas son:

- Diferencias en el formato o el tipo de datos utilizado para presentar los resultados: incluso en ocasiones no es posible determinar el tipo en base a la información extraída de la fuente. Por ejemplo, la moneda en que se expresa el precio para ciertos productos en tiendas electrónicas.
- Capacidades de consulta limitadas: existe una enorme heterogeneidad y limitación con respecto a las consultas que la fuente permite. En la mayoría de los casos se permite hacer un AND/OR con cierto número de palabras clave y una ordenación por algunos campos, pero las variaciones son muy importantes. Además, en algunos casos, dichas búsquedas son complicadas de homogeneizar. Por ejemplo, en una búsqueda en catálogo, la mayoría de los sites permitirán realizar un filtro por categoría, pero la lista de categorías diferirá de uno a otro.

En definitiva, las limitaciones impuestas por el acceso via web complican su tratamiento automatizado, especialmente si nos vemos obligados a combinar de alguna manera los datos obtenidos de más de una fuente.

3. ESTRATEGIAS PARA LA INTEGRACIÓN DE INFORMACIÓN HETEROGÉNEA

Parece evidente que es importante disponer de tecnologías que gestionen la complejidad de tratamiento e integración de la información heterogénea. El desafío que nos proponemos consiste, por tanto, en **ser capaces de acceder y aprovechar en nuestras aplicaciones, de forma automatizada, la enorme cantidad de información que tenemos disponible en fuentes semiestructuradas.**

Las diferentes soluciones que podemos plantear son las siguientes:

- Homogeneización de las fuentes semiestructuradas mediante acuerdos con los proveedores de la información.
- Mediante Tecnologías de Data Warehouse.
- Desarrollo de un **Mediador** que permita su tratamiento automatizado y disminuya en lo posible la necesidad de modificación y colaboración de las fuentes:
- A medida.
- Mediante **Soluciones EII (Enterprise Information Integration)** o Tecnologías de Base de Datos Virtual.

Todas las alternativas tienen su validez en ciertos entornos, por lo que es muy importante conocer y discriminar su aplicabilidad en cada situación real. Como veremos, sólo **la última opción es aplicable para desarrollar aplicaciones de prestaciones avanzadas que precisen de acceso a datos actualizados semiestructurados y estructurados de múltiples fuentes con un importante grado de variabilidad.**

3.1. Homogeneización

Este primer planteamiento exige un alto grado de colaboración con las entidades que gestionan los datos. Es necesario que los responsables de dichas fuentes aborden la adaptación de todas ellas, unificando todas las heterogeneidades que pueden presentar. En caso extremo, incluso la unificación física de todas las fuentes en una sola. Entre estas heterogeneidades, destacan, como hemos visto, el acceso a la información y los modelos de datos heterogéneos.

Pese a que este enfoque puede resultar apropiado y sencillo de implementar para ciertos entornos, especialmente si el número de fuentes de información es pequeño y poco cambiante, presenta dificultades muy claras que es importante considerar:

- Enfoque muy intrusivo: es necesaria la adaptación de la fuente que en muchos entornos es implanteable. Incluso si hay acuerdo para hacerlo, la implementación de los cambios necesarios en todas las fuentes a las que accedemos puede implicar un coste inabordable o innecesario.
- Dependencia: la aplicación que pretendemos desarrollar tiene una fuerte dependencia de la fuente de información. La pérdida de colaboración de la fuente puede tener efectos muy negativos que invalidan este enfoque, especialmente en el caso de aplicaciones críticas.

- Pese a que haya voluntad de colaboración, problemas de coordinación con múltiples fuentes muy cambiantes puede provocar desde fallos constantes en el acceso para un número importante de ellas, (o incluso que la información esté continuamente desactualizada), con la consiguiente degradación de la calidad del servicio.

Respecto al impacto de la homogeneización, el caso extremo es probablemente la integración de todas las bases de datos en una Base de Datos Universal. Este enfoque precisa de una compleja reingeniería que debe, al menos:

- Desarrollar, una vez diseñado el modelo unificado de datos, la nueva base de datos con toda la información a agregar. Han de considerarse mecanismos de transición que permitan incorporar la información que residía en las bases de datos anteriores.
- Rehacer los sistemas de información soportados por esas bases de datos, incurriendo además en nuevos costes de formación para los usuarios tradicionales que deben aprender a utilizar las nuevas interfaces de operación.

Enlazando con el apartado siguiente, si no se desea abordar la homogeneización del acceso, otra alternativa a considerar es el envío periódico de ficheros con la información (en algún formato predefinido, por ejemplo en XML) que pasaría a cargarse automáticamente en una base de datos local. Evidentemente, este nuevo planteamiento exige tecnologías que homogenicen los datos para generar los ficheros XML.

Concluyendo, pese a que no sea un enfoque descartable para casos muy particulares, es necesario asumir la enorme y creciente heterogeneidad con la que vamos a tener que tratar y plantearse soluciones avanzadas que aíslen y permitan evolucionar a los sistemas sin preocuparse del acceso a la información que precisan.

3.2. Data Warehouse

Las soluciones de Data Warehouse se basan en crear un gran repositorio central de información (Warehouse o Almacén) sobre el que se ejecutan las aplicaciones sin necesidad de que éstas interactúen directamente con las fuentes de información. La complejidad queda por tanto asociada al diseño eficiente de la estructura del Data Warehouse y de los procesos de carga de la información en su repositorio central.

Estos sistemas suelen proporcionar dos tipos de herramientas:

- Herramientas de extracción, transformación y carga. Son comúnmente llamadas herramientas ETL (Extraction, Transformation and Load).
- Herramientas de consulta OLAP (On-Line Analitic Processing), que facilitan la realización de consultas multi-dimensionales sobre los datos.

La ventaja principal de los Data Warehouses es la eficiencia. Como los datos residen localmente, pueden efectuarse en tiempos aceptables procesos muy pesados, tales como los que llevan a cabo aplicaciones de Data Mining o de análisis estadístico. Este tipo de tecnologías se han utilizado con éxito para soluciones (e.g. dentro del ámbito del Business Intelligence) destinadas a encontrar estadísticamente patrones de comportamiento interesantes en grandes volúmenes de datos, en muchos casos provenientes de históricos. Sin embargo muestran un importante número de limitaciones que los invalidan para otras aplicaciones:

- Un Data Warehouse no permite acceder directamente a los datos de las fuentes sino que trabaja con los datos precargados en el repositorio. Esto es especialmente crítico si la aplicación a desarrollar requiere forzosamente del uso de datos actualizados.
- No pueden tratar fuentes de datos semiestructurados, como las fuentes web. Tradicionalmente, los sistemas de Data Warehouse no han podido tratar fuentes de datos con sus capacidades de consulta limitadas y cuyo esquema de datos es impreciso. Sólo pueden tratarlas mediante consultas imprecisas (por palabra clave o similares, es decir, son tratadas como si fueran fuentes no estructuradas), o desarrollando programas ad-hoc.
- Exige la construcción de un gran repositorio central que manejará enormes volúmenes de datos. Esto tiene un impacto muy alto sobre los tiempos de implantación y los costes asociados de desarrollo y mantenimiento. En algunos casos, el volumen de datos puede ser incluso demasiado elevado para que el enfoque sea factible. De hecho, la mayoría de las aplicaciones de *Data Warehousing* no integran todos los datos de las fuentes, sino sólo aquellos subconjuntos considerados relevantes para sus propósitos.
- En el caso de integración de datos de fuentes autónomas (externas) el enfoque de extracción y carga masiva de datos puede no ser eficiente o viable.
  El enfoque de Data Warehousing requiere volcados periódicos de la información de las fuentes. Esto suele precisar de cierto grado de colaboración por parte de la fuente, tanto para el envío de avisos en caso de cambios como para la realización de descargas masivas de datos. En general esta colaboración va a ser imposible o demasiado costosa.
  Aún en el caso de que las entidades que gestionan las fuentes acepten colaborar, si son plenamente autónomas, cada organización deseará mantener el control sobre cómo su información es consultada y no estará dispuesta a transferir sus datos periódicamente a un almacén que no está bajo su control.

Por todas estas razones, los sistemas de Data Warehouse resultan poco adecuados para muchas de las aplicaciones de integración de la administración actual. Cuando cualquiera de los inconvenientes resulta un obstáculo clave, se hace imprescindible utilizar otras estrategias.

3.3. Mediadores Adhoc

Un primer paso en la dirección apropiada es eliminar la dependencia realizando un software que gestione la heterogeneidad y permita integrar los datos sin necesidad de alterar las fuentes.

Este esfuerzo puede realizarse con un desarrollo ad hoc que integre un conjunto bien definido de fuentes diversas. Este enfoque, aunque no es intrusivo presenta una restricción muy importante que afecta a su escalabilidad:

- Las exigencias de mantenimiento en caso de múltiples fuentes cambiantes pueden llevar a disminuir la calidad del servicio por fallos en el acceso o información desactualizada. Realmente la independencia de la fuente es relativa, porque cada alteración de la fuente exige alterar la aplicación.

La solución final debe por tanto prestar atención a este aspecto. **La tecnología que soporte las aplicaciones basadas en datos semiestructurados debe ser suficientemente escalable como para permitir desarrollar aplicaciones que ofrezcan servicios de calidad.**

4. SOLUCIONES EII O BASES DE DATOS VIRTUALES

4.1. EII

Como objetivo fundamental, las tecnologías EII permiten operar con información dispersa y heterogénea como si de una base de datos local se tratase.

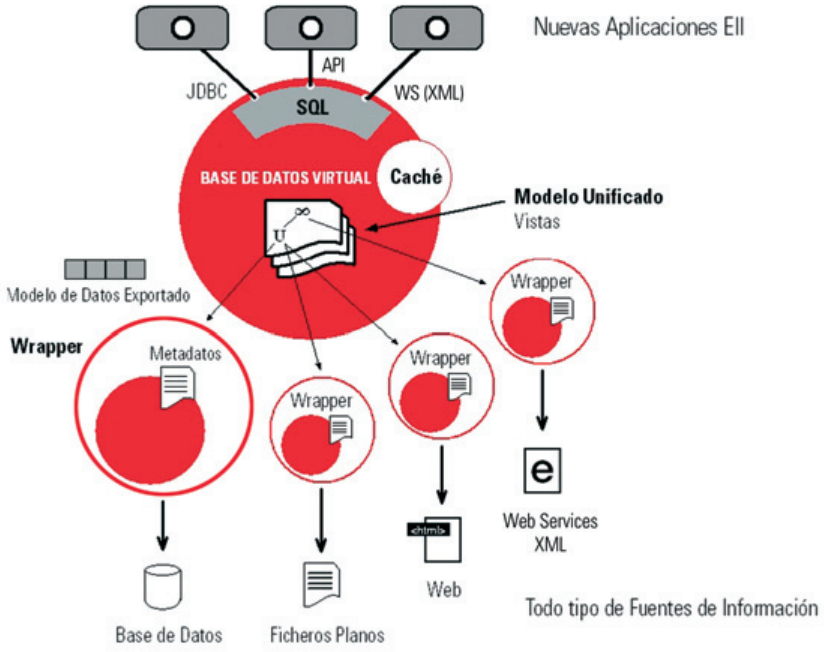

Figura 2 Arquitectura de una Solución EII

Sin necesidad de entrar en excesivo detalle técnico puede visualizarse su importante valor añadido. Se observa en la figura 2 como cada fuente de información es accedida a través de un robot de extracción, denominado **Wrapper.** Es importante destacar que la generación de estos wrappers debe ser rápida y poco costosa porque no exige en ningún momento el desarrollo de código; basta la simple configuración **(metadatos)** de un componente tecnológico.

Los wrappers posibilitan el acceso a los datos formateados según un **Modelo de Datos Exportado.** El módulo principal, la **Base de Datos Virtual,** permite definir un Modelo Unificado compuesto de **Vistas** que combinan como se desee los modelos de datos de los wrappers. Estas vistas unificadas son la que serán accedidas por las aplicaciones.

Las **Nuevas Aplicaciones EII** pueden acceder a la información desde un único punto y utilizando diferentes interfaces de acceso (driver xDBC, API Java, Web Service -XML-...). Y lo hacen realizando consultas sobre el modelo unificado como lo harían sobre una base de datos normal, utilizando un lenguaje de consultas (e.g. SQL). La Base de Datos Virtual debe descomponer esas consultas y lanzarlas en paralelo a los wrappers involucrados. Estos van a resolver la ejecución de cada subconsulta y la extracción de la información deseada, mientras la Base de Datos Virtual llevará a cabo la integración final en tiempo real de toda la información extraída.

Es importante notar que los datos residen en la fuente y sólo se accede en cada momento a los estrictamente necesarios para resolver las consultas solicitadas. Esto posibilita sistemas más ligeros y evita la duplicación de la información o la creación de grandes repositorios centralizados. Además, por consideraciones de rendimiento o necesidades propias de la aplicación, es también posible la activación de una potente Caché de consultas (almacena temporalmente los resultados de consultas anteriores para poder reutilizarlas automáticamente en próximas ocasiones) o la carga de los datos integrados en una base de datos para su uso posterior.

4.2. Ventajas EII

A modo de comparativa, hay que considerar como principales ventajas competitivas:

- Frente a la necesidad de colaboración de las fuentes, permite mantener la **autonomía** e independencia del nuevo sistema respecto de las fuentes.
- Frente a sistemas que permiten consultas imprecisas por palabra clave, permiten ejecutar **consultas potentes SQL,** incluyendo bases de datos relacionales, sistemas de data warehouse, ficheros planos en un formato predefinido (e.g. XML, logs...), hojas de cálculo, fuentes web, etc.
- Frente a enfoques que implican una costosa reingeniería o adaptaciones de los sistemas/fuentes heredados, permite realizar **soluciones de integración no intrusivas.**
- Frente a sistemas que requieren de costosos mantenimientos, ofrecen la posibilidad de **mantener el sistema de forma declarativa** y por tanto, rápida y poco costosa.
- Frente a soluciones que trabajan con datos precargados, ofrece también la flexibilidad de **trabajar con los datos actualizados** directamente extraídos de la fuente.
- Frente a soluciones muy pesadas y grandes repositorios de información, permite **sistemas ligeros** que saben acceder a los datos necesarios sin necesidad de almacenarlos.
- Frente a soluciones de extracción web que trabajan con ficheros o páginas estáticas, permiten acceder a la **web dinámica automatizando cualquier tipo de navegación.**
- Frente a soluciones de integración parciales (exclusivamente corporativas o en Internet), permiten incluso desarrollar sistemas que complementen la información corporativa con el subconjunto de información de Internet considerado relevante.

4.3. EIIs vs. EAIs

Dejando de lado otras tecnologías relacionadas con integración de información que, directa o indirectamente, compiten o complementan las soluciones EII (agregadores de contenidos, bases de datos federadas, herramientas de indexación y conceptualización de documentos…), especial interés merece apuntar brevemente algo acerca de los EAIs (Enterprise Application Integration).

Los EAIs, o sistemas de integración de aplicaciones, proporcionan un enfoque de integración orientado a proceso (también esencialmente declarativo) basado en mensajes XML. Pese a su éxito comercial, estos sistemas están limitados a tratar en cada paso información de una sola fuente, y no pueden combinar entre sí la información proveniente de varias fuentes o aplicaciones para proporcionar una visión a un nivel de abstracción superior. En ese sentido **son tecnologías complementarias a los EIIs:** estos pueden proveerles de información integrada que son consumidos en cualquier etapa de los procesos incorporados a los EAIs.

5. Integración de información en la Administración Pública

A la vista de lo anteriormente discutido, la utilización de EIIs en el sector público debe ser contemplada con interés.

Tradicionalmente, y en esos entornos el enfoque más común para abordar la integración de fuentes de información dispersas consiste en crear nuevos repositorios centralizados de datos que aglutinen toda la información necesaria. Los altos costes que conlleva, la adecuación de todas las fuentes a un sistema común de almacenamiento de datos, la migración de la ingente cantidad de información existente, el desarrollo de nuevos sistemas de información o la formación de los empleados públicos (con el perjuicio que se ocasiona al ciudadano hasta la adaptación al nuevo sistema), son obstáculos habituales que este nuevo enfoque puede ayudar a salvar; preservamos así las inversiones ya realizadas por las distintas administraciones en sus sistemas.

Independientemente de la integración de información interna, **las soluciones EII pueden ayudar a generar vistas de información que posibiliten la integración con otras administraciones** (bien directamente, bien a través de EAIs o web services…). Es decir, facilitan la rápida generación de las vistas necesarias sobre sus sistemas propietarios. Casos típicos se producen cuando la administración central precisa tomar una decisión respecto a un asunto en el que parte de las competencias han sido transferidas a las CC. AA. (y, por tanto, disponen de parte de los datos necesarios), o viceversa.

Desde el punto de vista del **ciudadano** también existen beneficios, ya que con las soluciones EII se pueden **crear con rapidez servicios útiles que pongan a su alcance información integrada:** datos censales, tributarios, de la seguridad social, evolución de licencias administrativas solicitadas, empleo público, ayudas, becas o subvenciones...

Sin ánimo de ser exhaustivos, otras aplicaciones de interés para la administración son:

1. Vigilancia Tecnológica & Business Intelligence.
   Sistema de búsqueda, detección y análisis del entorno, encaminado a difundir y transmitir de una forma continua y precisa, informaciones y conocimiento sobre los principales movimientos de diferentes áreas tecnológicas (Vigilancia Tecnológica) o sectores económicos (Business Intelligence). Estas herramientas son vitales para la toma de decisiones estratégicas y la mejora de la capacidad competitiva e innovadora de las empresas.
2. Servicios on-line o sistemas de reporting en tiempo real.
   Sistemas de reporting o nuevos servicios on-line que acceden a datos en tiempo real de los sistemas de información (e.g. sistemas de información completa de tramitaciones en curso del ciudadano).
3. Repositorios Virtuales de Datos / Normalización de Catálogos.
   Sistemas de consulta centralizados sobre catálogos/repositorios distribuidos (e.g. catálogos de bibliotecas, repositorios virtuales en medios de comunicación…)
4. Repositorios de Datos Base para Servicios Móviles o de Voz.
   Infraestructura necesaria para permitir la extracción de la información de las fuentes de interés y la homogeneización de los resultados de las consultas para construir servicios móviles o portales de voz.
5. Automatizaciones de operaciones web…

5.1. Experiencia: VIXÍA, un Observatorio Tecnológico en la Administración

La Figura 3 muestra la arquitectura básica del proyecto VIXÍA.

El proyecto VIXÍA, liderado por el Centro de Innovación y Servicios (CIS) de la Consellería de Industria e Comercio de la Xunta de Galicia y con la colaboración de Denodo Technologies (proveedora de la tecnología EII) y Ferroatlántica I+D, es un sistema de vigilancia tecnológica basado en una arquitectura EII.

El Instituto Galego de Promoción Económica, IGAPE, subvenciona también parte del proyecto, un valioso ejemplo de cómo desde la administración se puede fomentar la mejora de la competitividad de las pymes. El sistema se pondrá en operación en modo piloto para un grupo de 30 empresas de sectores altamente competitivos y considerados estratégicos dentro de la economía gallega: sector textil, sector químico, sector metal-mecánico y sector eléctrico.

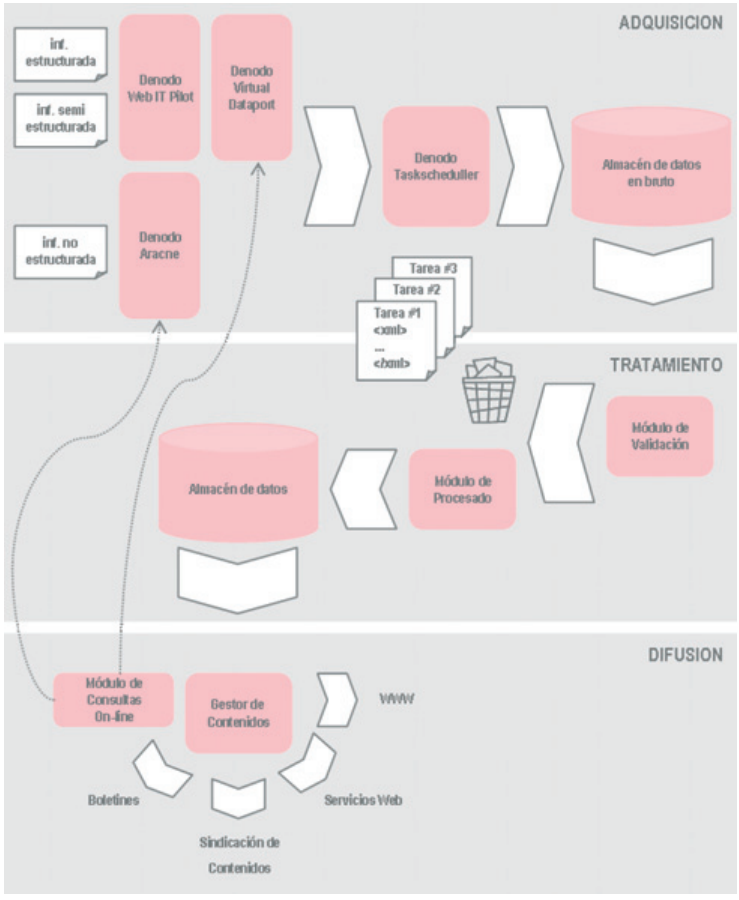

Figura 3. Arquitectura de Vixía

Ciñéndonos a aspectos técnicos, las labores de extracción e integración de información se organizan en tareas, gestionadas por el Task Scheduler. Denodo Web ITPilot es el componente utilizado para extraer y estructurar de forma automática los contenidos de las fuentes web semiestructuradas, mientras que Denodo Aracne es utilizado para indexar y clasificar en categorías los

documentos no estructurados. Todas las necesidades de integración de información, solucionando sus heterogeneidades estructurales y semánticas, son tratadas a través de Denodo Virtual Dataport (el corazón de la solución EII de Denodo).

Dependiendo de su uso, las tareas de extracción e integración de información pueden ejecutarse, o bien en tiempo real a través del módulo de consultas on-line, o bien en modo batch con periodicidad fija. La información extraída por las tareas de extracción, puede ser validada para asociarla a sus receptores adecuados (e.g. asignar a cada usuario sólo los ítems de información que le interesan) y post-procesada (por ejemplo, para extraer el contenido útil de un documento ignorando contenidos accesorios como publicidad, menús de navegación, etc.). Si se desea, la información puede también ser almacenada en un repositorio local que puede ser consultado posteriormente.

Por último, la arquitectura ofrece una interfaz web para el acceso y la gestión de los contenidos del sistema.

El proyecto VIXÍA utiliza tecnologías EII para automatizar el proceso de explorar, extraer e integrar información proveniente de más de 80 fuentes. El grueso de las fuentes utilizadas son sitios web disponibles en Internet (bien gratuitamente o bien vía suscripción), si bien también se accede a información proveniente de bases de datos y proveedores de contenidos on-line.

Entre la información que se recolecta, filtra y permite acceder de forma personalizada para cada empresa de acuerdo a sus requisitos particulares, se encuentran: noticias de la empresa y el sector, legislación relevante, líneas de subvención, eventos y congresos, proyectos de i+d, publicaciones, patentes, información de mercado o de empresas del sector, actividades de formación etc.

El proyecto VIXÍA también proporciona a las empresas usuarias funcionalidades para el tratamiento de su información interna, así como la posibilidad de realizar peticiones de información "a la carta", que son atendidas por un equipo de expertos del CIS ayudados por la herramienta.

REFERENCIAS

[1] G. Wiederhold. Mediators in the architecture of future information systems. Computer, 25(3), March 1992.

[2] S. Abiteboul. Querying semi-structured data. In Proceedings of the International Conference on Database Theory (ICDT), Delphi (Greece) 1997.

[3] A. Pan, M. Álvarez, J. Raposo, P. Montoto, V. Orjales, A. Molano and A. Viña. Mediator systems in E-Commerce Applictions. Proceedings of the 3th IEEE International Conference on Electronic Commerce and Web Information Systems (WECWIS). Newport Beach, California (USA) 2002.

[4] A. Pan, M. Álvarez, J. Raposo, P. Montoto, V. Orjales, J. Hidalgo, L. Ardao, A. Molano and A. Viña. The Denodo Data Integration Platform. Proceedings of the 28th International Conference on Very Large DataBase Systems (VLDB). Hong-Kong 2002.

[5] Alberto Pan, Juan Raposo, Manuel Álvarez, Justo Hidalgo and Ángel Viña. Semi-automatic wrapper generation for comercial web sources. Proceedings of the IFIP WG8.1 Conference on Engineering Information Systems in the Internet Context. Kanazawa (Japan) 2002.

[6] Alberto Pan, Manuel Álvarez, Juan Raposo, Paula Montoto, Anastasio Molano and Ángel Viña. A Model for Advanced Query Capability Description in Mediator Systems. Proceedings of the 4th International Conference on Enterprise Information Systems (ICEIS). Ciudad Real (Spain) 2002.