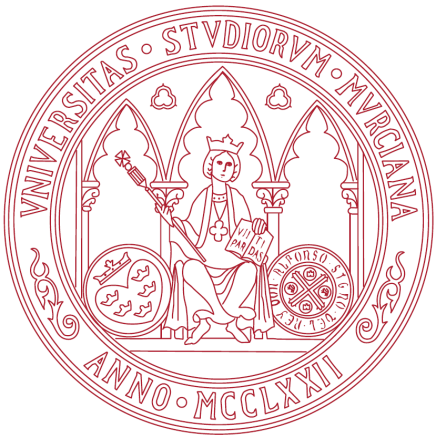
UNIVERSIDAD DE MURCIA

FACULTAD DE INFORMÁTICA

Entorno para la Inteligencia de Negocio
basada en Tecnologías Semánticas

D. Ángel Esteban Gil

2015



UNIVERSIDAD DE MURCIA

FACULTAD DE INFORMÁTICA

Entorno para la inteligencia de negocio basada en tecnologías semánticas

D. Ángel Esteban Gil

2015

Entorno para la inteligencia de negocio basada en tecnologías semánticas

Tesis doctoral presentada por Ángel Esteban Gil dentro del Programa de Doctorado: Informática y Matemáticas Aplicadas en Ciencias e Ingeniería

Dirigida por los Doctores
Jesualdo Tomás Fernández Breis
Francisco García Sánchez

Departamento de Informática y Sistemas
Facultad de Informática
Universidad de Murcia

2015

Agradecimientos

En primer lugar quiero dar las gracias a mis directores, Jesualdo Tomás Fernández Breis y Francisco García Sánchez, por darme la oportunidad de comenzar en el mundo de la investigación y por toda su ayuda en el desarrollo de esta tesis.

Por acogerme en su grupo de investigación, y dedicarme el tiempo y la atención necesaria para enriquecer este trabajo, gracias a Martin Boeker, del *Center for Medical Biometry and Medical Informatics* en la *University of Freiburg*.

Quiero dar las gracias a todos las personas que han colaborado en la validación de este trabajo:

- Gracias al Programa de Prevención de Cáncer de Colon y Recto de la Región de Murcia por su colaboración y la cesión de datos reales. Concretamente a los doctores: José Cruzado, Francisco Pérez-Riquelme y Fernando Carballo.

- Gracias al Departamento de Recursos Humanos del Hospital Clínico Universitario Virgen de la Arrixaca por implantar y usar el modelo de evaluación de activos de conocimiento y el módulo de planificación.

- Gracias al Grupo de Investigación en Enfermedades Respiratorias Infantiles coordinado por el doctor Luis García-Marcos. Sin su colaboración los procesos de desarrollo y validación del cuaderno de recogida de datos semántico hubieran sido mucho más complejos.

- Gracias al Grupo de Investigación en Regionalización Cerebral y Genes de Desarrollo coordinado por el doctor Salvador Martínez Pérez, por permitirme colaborar en uno de sus proyectos europeos facilitando el pilotaje y la validación de la solución de clasificación y explotación de contenidos multimedia.

Muy especialmente, quiero agradecer a mis compañeros del Departamento de Informática y Sistemas de la Fundación para la Formación e Investigación

Sanitarias de la Región de Murcia por compartir este camino conmigo. También quisiera dar un agradecimiento especial a Juan Pedro Serna Mármol, mi antiguo jefe, ya que recibí todo su apoyo y comprensión para compaginar mi trabajo de responsable del departamento con la investigación.

A todos los compañeros que han pasado por el grupo de Tecnologías para el Modelado, Procesamiento y Gestión del Conocimiento, por la ayuda que me han prestado y por los buenos recuerdos que me llevo de todos ellos. Especialmente me gustaría agradecer a Mari Carmen su orientación, ya que me ha facilitado la realización de este trabajo.

Quiero dedicar este trabajo a toda mi familia, padres, hermanas y abuelas. Su apoyo ha sido fundamental durante toda mi vida académica.

Por último, gracias María, eres la persona que mejor me comprende y que más confianza deposita en mi.

Índice

Índice de figuras

Índice de Tablas

Bloque I

Introducción, Estado del Arte y Objetivos

Capítulo 1

Introducción

La Inteligencia de Negocio (IN) se define como un conjunto de metodologías y herramientas que permiten transformar los datos en información y la información en conocimiento [1]. Ese conocimiento se convertirá en un activo más que permita analizar el estado de cualquier tipo de organización, dando soporte a la toma de decisiones estratégicas. Desde el punto de vista de las tecnologías de la información, la inteligencia de negocio consiste en depurar e integrar datos de múltiples sistemas de información de una entidad, transformarlos en una única fuente estructurada, y explotar esa fuente de información a través de diversas herramientas (buscadores avanzados, generación de informes, cuadros de mando o predictores).

La arquitectura de este tipo de soluciones suele tener tres ámbitos de actuación [2]: la integración de datos, la entrega de información, y el análisis y la planificación de la organización. En esos tres ámbitos se describen trece características esenciales que cualquier plataforma debería incluir [3]. La evolución de la Inteligencia de Negocio alcanza tres fases. La Inteligencia de Negocio 1.0 está centrada en analizar datos estructurados y cubre ocho de las trece características esenciales [3]. La Inteligencia de Negocio 2.0 pretende proporcionar herramientas para añadir la información no estructurada a este tipo de sistemas. Está orientada principalmente a integrar datos generados en aplicaciones colaborativas basadas en Web 2.0 y en redes sociales [4]. La Inteligencia de Negocio 3.0 está orientada a incluir los nuevos dispositivos (teléfonos móviles y tabletas) a la Inteligencia de Negocio. Aún no está claro cuáles serán los servicios que aportará esta nueva evolución de la IN pero son tecnologías que probablemente cambien el mercado [5].

Actualmente, las mayoría de productos de IN tienen sus núcleos centrados en la funcionalidad de entrega de información. Sin embargo el análisis y sobre todo la mejora de las capacidades de integración se han convertido en una

línea de interés que crece en las grandes compañías que desarrollan productos de IN [2], ya que el 80 % de la inversión en la implantación de estas soluciones se realiza en la fase de integración [6].

El impacto que ha tenido la IN en las organizaciones se suele medir en los resultados económicos de la empresa [7]. También es muy útil para poder medir la calidad de la información que gestionan las organizaciones [8]. Como ya se ha comentado, la IN ayuda a analizar lo que está pasando y a la toma de decisiones, además de convertir los datos en información, y la información en conocimiento. Este conocimiento está orientado al análisis de la actividad productiva y económica de la empresa, pero no tiene en cuenta otros activos de conocimiento como el capital humano, el clima laboral, la evaluación del desempeño, la cultura corporativa, los valores empresariales o las relaciones con otras empresas, entre otros. A este tipo de activos se les denomina "activos de conocimiento" [9] y no suelen estar incluidos en las soluciones de IN, lo que supone una carencia de información en el proceso de toma de decisiones.

Por otro lado, la Web Semántica describe una metodología para dotar de significado a múltiples tipos de contenido [10]. Gracias a que esa información pueda ser procesada por agentes software, permite que se puedan usar herramientas de razonamiento automático, que inferirán nuevo conocimiento y serán capaces de comprobar la consistencia lógica de los datos. La Web Semántica se ha idenficado como un marco tecnológico que puede integrarse en las soluciones de IN, a nivel de integración de la información [11; 12; 13], a nivel de evaluación de la calidad de la información [14], a nivel de explotación de la información [15; 16; 17] y, por supuesto, a nivel de compartición y comparación de soluciones de IN heterogéneas e implantadas en diferentes compañías [18; 19].

La mayoría de propuestas de integración de tecnologías semánticas con productos de IN se basan en usar estas tecnologías como apoyo a algunos procesos de las plataformas existentes en el mercado. Por ejemplo, a la anotación semántica de datos no estructurados [11], integración semántica de la información [12] o el uso de ontologías que sirvan como una representación de los metadatatos del almacén de datos [11]. Es decir, se trata fundamentalmente de extensiones semánticas a las soluciones de las que se dispone en el mercado. A día de hoy no existe una solución integral que tome como base la Web Semántica para generar herramientas que permitan cubrir el ciclo de vida de una plataforma de IN, desde su diseño, hasta su puesta en marcha y su mejora evolutiva.

En esta tesis se propone una solución totalmente funcional que ofrezca metodologías y herramientas que permitan cubrir los servicios de la mayoría de soluciones de IN del mercado, usando tecnologías de la Web Semántica.

El hecho de partir de un modelo basado en ontologías en la fase de integración de la información, permitirá que ese mismo modelo se convierta en una representación formal del conocimiento generado, que podrá ser entregado, analizado y explotado usando diferentes tecnologías semánticas. Como resultado final, se obtiene una serie de productos software que permiten la generación y compartición del conocimiento aprovechando las ventajas que poseen las tecnologías semánticas. Además, se presentan dos modelos adicionales: (1) añadir información estructurada al almacén de datos semántico, lo que permitirá gestionar y explotar la información en entornos de gran heterogeneidad, y (2) evaluar el conocimiento y analizar el impacto de las diferentes decisiones que se toman en la organización.

1.1 Organización del documento

Este documento se divide en cuatro bloques principales, cada uno de ellos compuesto por una serie de capítulos. En el bloque I se ofrece una visión general del estado del arte de aquellos aspectos relevantes para este trabajo y se definen los objetivos de la tesis. En el bloque II se describe la solución semántica para la Inteligencia de Negocio y se presentan diversos entornos de validación de este enfoque. El bloque III ofrece las conclusiones finales de la tesis, una discusión de aspectos relevantes de la investigación realizada, una descripción de las posibles vías futuras de trabajo y el listado de publicaciones científicas y contribuciones a congresos relacionadas con este trabajo. Por último, el bloque IV incluye un resumen en inglés de toda la tesis.

Los tres primeros capítulos pertenecen al Bloque I. El capítulo 1 proporciona una breve introducción a la tesis y su organización. El capítulo 2 ofrece una visión del estado del arte de los diferentes métodos y tecnologías relacionadas con la tesis. En la primera parte del capítulo se revisa la Inteligencia de Negocio, sus diferentes fases y retos, finalizando con diferentes ejemplos de aplicación. Se continúa con una explicación de las diferentes tecnologías de la Web Semántica que se usan en la solución propuesta, comentando trabajos que combinan la Inteligencia de Negocio con Web Semántica. Posteriormente se define la Web 2.0 y su relación con la Web Semántica. Se dedica un apartado especial a los diferentes métodos de integración de la información, ya que es uno de los procesos clave de la IN, y se contextualiza con diferentes aproximaciones realizadas con tecnologías semánticas. A continuación se detalla uno de los servicios que se suelen encontrar en todas las soluciones de IN como es el Procesamiento Analítico en Línea (OLAP). Por último, se finaliza el capítulo hablando de herramientas y métodos para evaluación del conocimiento de las compañías. El capítulo 3, último del Bloque I, describe

los objetivos de la tesis y la metodología de investigación seguida.

El bloque II está formado por dos capítulos. El capítulo 4 presenta las diferentes metodologías y herramientas que se han desarrollado e integrado entre sí para obtener un marco de trabajo para la Inteligencia de Negocio Semántica. Además se comentan las diferentes plataformas que se han desarrollado usando este marco. En el capítulo 5, último del Bloque II, se describen diversos escenarios en los que se han aplicado y validado los resultados de este trabajo. Se incluyen dominios como el económico-financiero, el sector clínico, varios campos de la investigación biomédica y evaluación del desempeño del personal de un hospital.

El bloque III está constituido por dos capítulos. En el capítulo 6 se hace una discusión sobre los aspectos más relevantes del trabajo realizado y se describen las posibles vías futuras. En el capítulo 7 se comentan las conclusiones finales y se listan las publicaciones científicas y contribuciones a congresos relacionados con esta tesis.

El bloque IV incluye un único capítulo con un resumen de la tesis en inglés.

Capítulo 2

Estado del Arte

En este capítulo se revisan exhaustivamente los métodos y tecnologías relacionadas con este trabajo. En la sección 2.1, se revisa la Inteligencia de Negocio y se repasa su evolución. En la sección 2.2 se introduce la Web Semántica y sus tecnologías. El estado de la Web 2.0 será descrita y analizada en el contexto de la Web Semántica y de la Inteligencia de Negocio en la sección 2.3. Se definen las metodologías y herramientas para la integración de la información en la sección 2.4, haciendo especial hincapié en tecnologías de transformación semánticas. En la sección 2.5, se describe el procesamiento analítico en línea y, por último, en la sección 2.6 se analizan las metodologías para medir y evaluar los activos de conocimiento.

2.1 Inteligencia de negocio

El término Inteligencia de Negocio (IN) se acuñó en 1958 [1] y se define como un conjunto de metodologías y herramientas que permite transformar los datos en información y la información en conocimiento. Ese conocimiento permitirá asesorar en la toma de decisiones estratégicas de cualquier organización. Desde el punto de vista de las tecnologías de la información y las comunicaciones, la IN consiste en depurar e integrar datos de los sistemas de información de una organización para transformarlos en una fuente de información estructurada y con un modelo de datos más formal. Posteriormente, esta información será explotada por diversas herramientas (buscadores avanzados, generación de informes, cuadros de mando, etc) que ayudarán al proceso de toma de decisiones.

La principal diferencia entre los sistemas de IN y los sistemas de información tradicionales reside en cómo se almacena la información. Mientras que en los sistemas de información el modelo de datos está optimizado para regis-

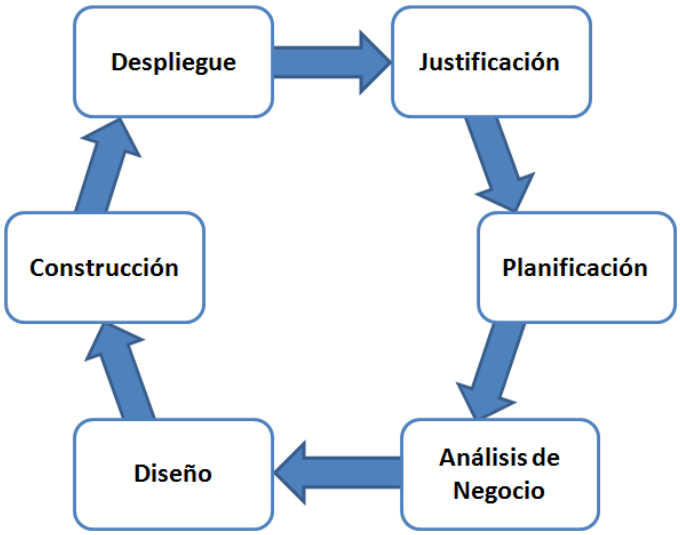


Figura 2.1: Ciclo de vida de una solución de IN

trar datos y gestionar procesos, en los sistemas de inteligencia de negocio la información se estructura en un formato en el que sea mucho más fácil hacer preguntas y recibir respuestas. El objetivo principal de una solución de IN es facilitar qué ocurre en la organización actualmente, cuál es el motivo de lo que pasa, qué podría pasar (basándose en datos históricos) y qué decisiones se deberían tomar.

En la figura 2.1 se puede ver un diagrama con el ciclo de vida de una solución típica de IN [20]. En la fase de justificación se evalúan las necesidades de negocio que requieren ser cubiertas. En la fase de planificación se desarrollan las estrategias para identificar de qué sistemas de información se va a recuperar la información y cómo ésta va a ser integrada. En la fase de análisis de negocio se identifican los requisitos que deben ser cubiertos por los servicios de análisis y explotación de la solución. En la fase de diseño se conceptualiza tanto el proceso de integración como el de explotación de la información. En el proceso de construcción se realizan los desarrollos específicos necesarios y se parametriza la solución de IN para adaptarla a los requisitos concretos de negocio de la organización. Por último, en la fase de despliegue se realiza la extracción, transformación y carga desde los orígenes de información al almacén de datos. A partir de ese momento, los usuarios pueden medir y analizar la actividad de la organización. Este ciclo de vida se repite cada vez que se establecen nuevas necesidades de negocio.

En [3] se definen las trece características esenciales que debe tener cualquier solución de IN. Estas características se clasifican en tres ámbitos de-

pendiendo de su funcionalidad, a saber, integración, entrega de información y análisis. A continuación se describe cada una de ellas:

- **Integración**

    1. **Infraestructura para la IN.** Este servicio es el que integra todas las fuentes de datos de la organización. Además ofrece servicios para la gestión de usuarios, seguridad, imagen corporativa, motores de búsqueda, etc. Es el punto de entrada del usuario a la hora de consultar o analizar la información.
    2. **Gestión de metadatos.** La gestión de metadatos es un servicio fundamental en estas soluciones. Al igual que la infraestructura para la IN ofrece un marco corporativo de consulta de los datos de la organización, la gestión de metadatos ofrece el lenguaje que se usa para analizar esos datos. Este servicio debe permitir la definición de jerarquías, nombrar indicadores, etc.
    3. **Herramientas de desarrollo.** Las plataformas de IN deben proporcionar mecanismos de interoperabilidad con otras aplicaciones, y la posibilidad de crear nuevos servicios. Estas soluciones suelen permitir que se creen nuevos componentes usando asistentes gráficos, sin la necesidad de el asesoramiento de un profesional TIC.
    4. **Colaboración.** Otro servicio importante de los sistemas de IN es la posibilidad de que se establezcan foros de debate entre los usuarios a partir de los datos, que se puedan hacer anotaciones, añadir métricas, etc.

- **Entrega de información**

    1. **Generación de informes.** Uno de los servicios que se encuentra en todas las soluciones de IN consiste en la generación de informes. La generación de informes suele estar caracterizada por (1) realizar una búsqueda de los datos sobre la que se elabora el informe, (2) representar esos datos de una forma que sea ilustrativa al usuario y (3) generar el informe cada cierto tiempo, dependiendo de la frecuencia con la que se deba revisar esa información. Un tipo especial de informe son las alertas. Este informe incluye información sobre aquellas actividades que no se están realizando adecuadamente, o que superan los plazos requeridos para su puesta en marcha. Este tipo de informe es muy importante para que los sistemas informen de aquellos aspectos que no están operando correctamente en la organización.

2. **Cuadros de mando.** Este servicio es una representación gráfica de los datos del almacén. Esta representación suele mostrar gráficamente el estado de alguna de las líneas de negocio de la organización.
3. **Generación de consultas *ad hoc*.** Esta herramienta permite que los usuarios puedan hacer sus propias preguntas al almacén de datos. Estas consultas son las que permiten definir informes o cuadros de mando personalizados. En la mayoría de soluciones, se usa la gestión de metadatos para añadir una capa semántica que ayude a los usuarios a navegar por las fuentes de información.
4. **Integración con Microsoft Office.** En muchas ocasiones, las plataformas de IN son usadas como un middleware al que se accede a través de un cliente externo. En la mayoría de soluciones del mercado, ese cliente es Microsoft Excel, llegando la integración en muchos casos a permitir que desde Excel se puedan modificar los datos.
5. **Buscadores avanzados.** Este servicio permite que se puedan hacer búsquedas, tanto en información estructurada, como en contenido en texto abierto. Este servicio se suele apoyar en la generación de consultas *ad hoc* y también usa una capa semántica que facilite al usuario la búsqueda de contenidos concretos.

- **Análisis**

1. **Procesamiento analítico en línea (OLAP).** Este servicio permite analizar datos de una forma rápida y eficiente. Este tipo de procesamiento permite cambiar la estructura de almacenamiento de la información, optimizada para su gestión, por otro modelo que esté optimizado para su consulta. Las arquitecturas que se pueden encontrar para ese tipo de análisis son muy variadas, tales como sistemas relacionales, multidimensionales, sistemas de almacenamiento en memoria (*in-memory*) y bases de datos NoSQL.
2. **Visualización interactiva.** Este servicio consiste en tener la posibilidad de que la información que se muestra en los diferentes visualizadores sea navegable entre sí. Por ejemplo, si se ha definido un cuadro de mando donde se representa el porcentaje de ventas de los principales productos de una empresa, haciendo clic sobre uno de ellos se podría llegar a la información relacionada con los clientes que han adquirido ese producto.

3. **Modelado predictivo y minería de datos.** Estos servicios usan técnicas matemáticas avanzadas e inteligencia artificial para generar modelos predictivos o nuevo conocimiento a partir del almacén de datos. Los resultados de este tipo de técnicas suelen realimentar la solución para que puedan plasmarse en informes o en cuadros de mando. En [21] se identifican cuáles son los diez algoritmos más usados en minería de datos básandose en comités de expertos, número de citas y encuestas a diferentes comunidades de usuarios. Entre esos algoritmos se encuentran las redes neuronales, algoritmos de clasificación, redes bayesianas y algoritmos genéticos. Estas técnicas cubren los procesos de clasificación de la información, clustering, regresión, predicción y análisis asociativos y en red. Muchos de ellos han sido incorporados a soluciones comerciales y de código abierto disponibles en productos de IN.
4. **Cuadros de mando de indicadores estratégicos.** Este tipo de cuadro de mando relaciona los indicadores reales de la empresa con los resultados que hipotéticamente se pretendían conseguir. Este tipo de servicio debe incluir herramientas para la gestión y evaluación de los activos de conocimiento.

Actualmente, los productos de IN tienen sus núcleos centrados en la funcionalidad de entrega de información, aunque el análisis y la mejora de las capacidades de integración se han convertido en una línea de interés que crece en las grandes compañías de desarrollo de este tipo de soluciones [2].

La importancia de la IN en las organizaciones ha aumentado de forma importante en los últimos años, ya que está reconocida como una herramienta que tiene un impacto positivo real en los beneficios empresariales [7]. Hay un acuerdo general en que hay que aprovechar el enfoque de la IN para mejorar las estrategías de las organizaciones [22]. El rápido progreso que han tenido este tipo de prácticas empresariales ha hecho que el número de soluciones empresariales haya crecido enormemente, dando lugar a altos niveles de heterogeneidad [23].

A continuación se describe la evolución de la Inteligencia de Negocio.

2.1.1 Inteligencia de negocio 1.0

La IN 1.0 está centrada en analizar datos estructurados. Sus principales procesos consisten en la recopilación de datos de diversos orígenes, extracción de éstos y su análisis [24; 25; 7]. Casi todas las soluciones de IN que se encuentran hoy en la industria son IN 1.0, ya que trabajan con datos bastante

estructurados, provenientes de los sistemas operacionales de las compañías, que a su vez suelen estar almacenados en sistemas relacionales de almacenamiento de datos. Las técnicas de análisis más usadas en estos sistemas se popularizaron en los noventa, y provienen de métodos estadísticos desarrollados en los setenta y de técnicas de minería de datos descritas en los ochenta [5].

La gestión y el almacenamiento de los datos son la base de la IN 1.0. El diseño de cómo se agrupan los datos para la extracción, transformación y carga (ETL, *Extraction, Transformation and Load*) es un proceso esencial para convertir e integrar los datos de los diversos orígenes que se encuentran en las compañías. Herramientas como los generadores de consultas a la base de datos, el procesamiento análitico en línea (OLAP), y los generadores de informes son los principales servicios que se encuentran en la IN 1.0 para analizar la información. Estas herramientas son sencillas e intuitivas, y permiten explorar la información de un modo gráfico [26]. Otra característica importante de estas soluciones son las herramientas de gestión del rendimiento empresarial (BPM) usando cuadros de mando para visualizar y comparar diversos indicadores de rendimiento. Por último, las soluciones de IN 1.0 tienen una capa de servicios que se apoyan sobre el resto y que usan análisis estadístico y técnicas de minería de datos para realizar tareas como segmentación de la información, agregación de la información, análisis de regresión, detección de anomalías, y generación de modelos predictivos que ayuden a la toma de decisiones estratégicas. La mayoría de estas herramientas y técnicas han sido incorporadas en las soluciones de IN más comercializadas, ofertadas por las mayores compañías de software del mundo, incluyendo Microsoft, IBM, Oracle, y SAP [2].

Entre las trece características esenciales para una plataforma de IN definidas en la sección 2.1, hay ocho que son recogidas en la IN 1.0: generación de informes, cuadros de mando, consultas bajo demanda, buscadores avanzados, procesamiento analítico en línea, visualización de datos interactiva, modelos predictivos y minería de datos. En el marco de la IN 1.0 hay áreas que se siguen investigando hoy en día, como son los bancos de trabajo en minería de datos, los gestores de bases de datos no relacionales, gestores de bases de datos en memoria, y herramientas de toma de decisiones en tiempo real [5].

En la figura 2.2 se puede ver el esquema típico de una solución de IN 1.0. Como entradas se tienen los diferentes sistemas operacionales de la organización. Esas fuentes de datos son procesadas y transformadas usando técnicas de ETL, y se almacenan en un origen de datos único (*data warehouse*) sobre el que se desarrollan diversos servicios de análisis y consulta de los datos. A continuación se describen cuáles son los principales orígenes de datos que se

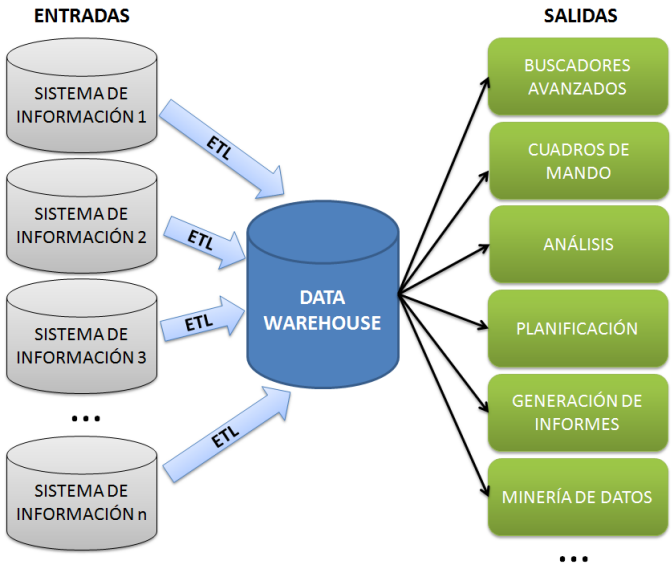


Figura 2.2: Esquema Inteligencia de Negocio 1.0

usan en las soluciones de IN 1.0.

2.1.1.1 Orígenes de datos

Como se ha comentado, la IN basa su funcionalidad en orígenes de datos que han sido optimizados para su explotación. En estos sistemas se integran todos los datos de los diferentes sistemas de información de una organización. Por lo tanto, el tratamiento de la información se realiza fuera de los sistemas de gestión de la empresa, lo que evita que ese tratamiento ralentice el funcionamiento normal de los sistemas informáticos.

En las soluciones de IN suelen existir dos tipos de orígenes de datos, almacén de datos operacional y *Datawarehouse*, que pueden combinarse entre sí. A continuación se describen sus principales características:

1. Almacén de datos operacional. Este tipo de orígenes de datos almacena información homogénea de los diferentes sistemas de información de la organización una vez transformados. Estos sistemas no suelen almacenar datos históricos, es decir, la información suele ser volátil [27]. Son muy útiles para analizar datos en tiempo real [28] o para procesarlos antes de pasar al almacén de datos. Ese procesamiento suele consistir en realizar operaciones con los datos que generan nueva información [29].

2. Almacén de datos o *Datawarehouse*. En el *datawarehouse* se almacena el histórico de todos los datos provenientes del proceso de transformación. Si los datos provienen de almacenes de datos operacionales también suele hacerse un segundo proceso de ETL. En organizaciones con grandes estructuras departamentales o que tienen muchas líneas de negocio, los datawarehouse pueden dividirse en subconjuntos de almacenes (*Data mart*) que permiten explotar la información desde un enfoque de negocio concreto.

Es importante destacar que en ambos casos los datos suelen estar desnormalizados para mejorar el rendimiento de las consultas. Esa desnormalización implica que el proceso de ETL tiene que hacerse correctamente para evitar los errores que pudieran provocarse por esa desnormalización.

Una tendencia habitual en las soluciones de IN consiste en usar orígenes de datos NoSQL [2; 30; 31] que sean mucho más eficientes a la hora de consultar la información.

2.1.2 Inteligencia de negocio 2.0

El desarrollo de Internet durante los primeros años del siglo XXI, permitió establecer el concepto de usuario de la red que consume y que además provee de contenidos. El nivel de interacción de los usuarios con Internet se desarrolló enormemente con el nacimiento de las redes sociales y de portales de comercio electrónico como Amazon o eBay. Este crecimiento de contenidos generados a través de plataformas Web 2.0 ha establecido la necesidad de poder analizar esa información. Por este motivo surge la IN 2.0 que está centrada en analizar contenidos web que carecen de estructura [26].

Una gran cantidad de compañías ofrecen sus productos a través de la Web y organizan su información para que sea atractiva para el cliente, no para ser explotada por herramientas de análisis de negocio. Para analizar el comportamiento de los clientes hacen uso de herramientas que permiten analizar los portales Web, como Google Analytics, que proporciona un rastreo de cuáles han sido las actividades que ha realizado el usuario en el portal y que, además, también revelan cómo se comporta al navegar de un sitio a otro. Las aplicaciones de la Web 2.0, que empezaron a ser populares sobre el 2004, permiten que no sólo las compañías creen contenidos, sino que los usuarios puedan generarlos a través de redes sociales, blogs, foros, sitios para compartir archivos multimedia (imágenes y vídeos), e incluso mundos virtuales y juegos sociales [4; 32]. Las aplicaciones Web 2.0 pueden almacenar eficientemente la retroalimentación y las opiniones de muchos clientes para una gran cantidad

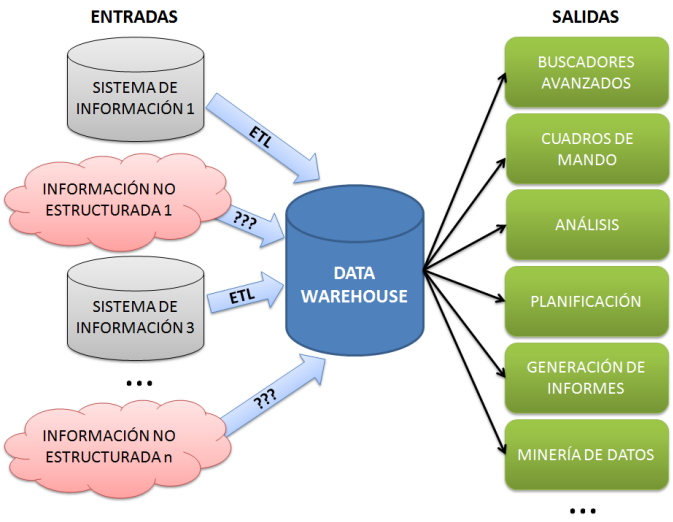


Figura 2.3: Esquema Inteligencia de Negocio 2.0

de tipos de negocio, y las empresas, e incluso los entornos socio-políticos, necesitan analizar esa información.

Muchos investigadores creen que el análisis de la información social es una oportunidad única para que el proceso de compra-venta se convierta en una conversación entre los clientes, en vez de la tradicional conversación vendedor a cliente en un sólo sentido [33]. Las herramientas de la IN 2.0, a diferencia de la IN 1.0, están mucho menos desarrolladas, aunque ya hay soluciones de minería de textos (extracción de información, identificación de temas clave, minería de opiniones o sistemas de pregunta-respuesta), minería web, análisis de redes sociales, y análisis espacio-temporal de orígenes de datos provenientes de soluciones IN 1.0 [26].

Si se vuelve al esquema de la figura 2.2, exceptuando la definición de consultas y las capacidades de búsqueda, no existen capacidades de análisis de textos para contenidos no estructurados dentro de la definición de capacidades esenciales para la IN. En [5] se propone que la Web Semántica y las técnicas de procesamiento del lenguaje natural serán tecnologías a tener en cuenta en la IN 2.0.

La principal diferencia entre la IN 1.0 y la IN 2.0 no se encuentra en los servicios que ofrece, sino en el origen de las fuentes de información (como se puede ver en la figura 2.3). En este caso, además de incorporar datos estructurados, también se intenta analizar otro tipo de fuentes cuya información no está tan estructurada, como redes sociales, foros de discusión, blogs, opiniones de los clientes, etc. En este caso, los procesos de ETL se suelen basar

en técnicas de procesamiento del lenguaje natural y minería de textos.

2.1.3 Inteligencia de negocio 3.0

La IN 3.0 está relacionada con lo que se ha denominado "Internet de las cosas". En 2011, el número de dispositivos como teléfonos móviles y tabletas vendidos superó al número de ordenadores y portátiles [34]. Aunque el número de ordenadores sobrepasó los mil millones en 2008, en [26] se prevee que en 2020 haya diez mil millones de dispositivos móviles conectados a Internet. Estos dispositivos, el ecosistema de aplicaciones disponibles para ellos, y sus capacidades de interconexión (RFID, Bluetooth y etiquetas de radio) han abierto un nuevo mundo al desarrollo de aplicaciones innovadoras. Los nuevos datos de movilidad y localización, incluyendo información recogida de diferentes sensores, se convierten en una nueva fuente de información que debe ser recogida, procesada y analizada.

Aún no existen soluciones de IN 3.0 integradas ni comerciales, pero el informe [5] las incluye como una nueva tecnología que puede cambiar el mercado de soluciones de IN [35].

2.1.4 Ejemplos de aplicación de la Inteligencia de Negocio

A continuación se exponen diversos dominios donde la IN ha tenido un alto impacto. Los ejemplos irán enfocados a: (1) comercio electrónico, (2) gobierno y política, (3) ciencia y tecnología, (4) ciencias de la salud, y (5) seguridad pública.

2.1.4.1 Comercio electrónico

El crecimiento de la IN se ha producido principalmente por la Web y las plataformas de comercio electrónico. La transformación del mercado ha sido liderada por las grandes compañías de comercio electrónico como Amazon o eBay, gracias a sus plataformas altamente escalables e innovadoras y a sus sistemas de recomendación de productos. Gigantes de Internet como Google, Amazon y Facebook continúan liderando el desarrollo de analíticas web, computación en la nube, y redes sociales. El auge de la Web 2.0, con sus consumidores y productores de contenidos ofrece una nueva oportunidad para "escuchar" la voz del mercado a través de clientes, empleados, inversores, etc [32]. A diferencia de los sistemas de información operacionales tradicionales, la información es almacenada en sistemas donde los datos no están estruc-

turados, lo que no significa que de esa información no se puedan extraer las opiniones de los clientes o incluso información sobre el comportamiento de éstos [4].

Para analizar las opiniones de los clientes se suelen usar técnicas de análisis de textos y análisis de sentimientos (saber a partir del análisis del texto si la opinión es positiva o negativa) [36]. También se han desarrollado diversas técnicas para sistemas de recomendación de productos, como son la generación de reglas de asociación, la segmentación y agregación de las bases de datos, la detección de anomalías, y la minería de grafos [37]. En entornos donde el catálogo de productos y su diversidad es muy grande se han desarrollado soluciones para mejorar las búsquedas y los sistemas de recomendación personalizada [38].

2.1.4.2 Gobierno y política

La llegada de la Web 2.0 también ha cambiado los mecanismos de comunicación de los gobiernos y de los partidos políticos. Este fenómeno ha dado lugar a lo que se conoce como "políticos 2.0", esto es, políticos que participan en foros de debate, en redes sociales, campañas en línea, etc. [26] La ciudadanía reclama cada vez más procesos políticos transparentes y donde haya mucha más participación por parte de los ciudadanos. En este contexto, la Web 2.0 y la IN se han convertido en herramientas esenciales para desarrollar este tipo de plataformas. En este dominio se definen técnicas de minería de opiniones, análisis de redes sociales, y análisis de contenidos multimedia para analizar la participación de los ciudadanos en foros políticos, blogs, etc. [39; 40].

La Web Semántica también se ha considerado como una herramienta importante en este dominio. En [41] se demuestra que el desarrollo de una ontología (OntoCop) para modelar los servicios gubernamentales permite que los servicios públicos puedan organizar y ofrecer una información más cercana al ciudadano.

2.1.4.3 Ciencia y tecnología

Muchas áreas de la ciencia y la tecnología se están beneficiando de diversos dispositivos que permiten generar grandes cantidades de información en cada experimento. Para facilitar la compartición y el análisis de los datos de esos experimentos, la Fundación Nacional para la Ciencia (NFS) ha demandado que cada proyecto o experimento tenga un plan de gestión de los datos. Es en estos casos donde se recomienda la aplicación de soluciones de IN, tanto para almacenar esos grandes volúmenes de datos, como para analizarlos y explotarlos correctamente [42].

2.1.4.4 Ciencias de la salud

El dominio de las ciencias de la salud se ha convertido en el entorno ideal para la puesta en marcha de soluciones de IN. En este dominio, las fuentes de datos suelen ser bastante heterogéneas, desde sofisticados aparatos médicos a comunidades de salud en línea. Además, se encuentran datos de la Historia Clínica Electrónica (HCE), de prescripción farmaceútica, o en los últimos años información genómica como datos de secuenciación, expresión de genes y genotipado [26].

Durante los últimos 15 años, la HCE ha sido implantada en la mayoría de hospitales y clínicas del mundo desarrollado. Esto ha conllevado un aumento del conocimiento clínico y la detección de patrones en el diagnóstico y tratamiento de enfermedades gracias al uso de técnicas de IN aplicadas sobre los datos de la HCE [43; 44; 45].

Además se pueden encontrar muchos sitios como *Daily Strength* y *PatientsLikeMe* que proporcionan un entorno único para la formación de pacientes [46], especialmente en enfermos crónicos. En estos sistemas se encuentran técnicas como la generación de reglas, la agrupación de pacientes en cohortes, la monitorización de las comunidades de pacientes, análisis de texto clínico, efectos adversos de los medicamentos y una gran cantidad de posibles análisis. En este entorno también empiezan a aparecer ontologías clínicas, que ayudan a clasificar el conocimiento.

2.1.4.5 Seguridad pública

Desde los trágicos atentados del 11 de septiembre de 2001, la seguridad pública ha pasado a ser un elemento esencial del modo de vida de la humanidad. Internet se ha convertido en una herramienta poderosa de comunicación, pero también ha permitido que diversas organizaciones criminales la usen como mecanismo de comunicación y orquestación de actividades en contra de la seguridad ciudadana [39]. Las herramientas de IN pueden contribuir de una forma esencial a detectar dónde se podría estar planeando un delito a través del análisis de la información que viaja a través de la red.

2.2 Web Semántica

El término Web Semántica [10] fue acuñado en el año 2001 por Tim Berners-Lee para designar una nueva generación de la Web en la que los contenidos sean algo más que una suma de información y servicios escasamente estructurados. Este nuevo enfoque propone reestructurar y enriquecer los documentos

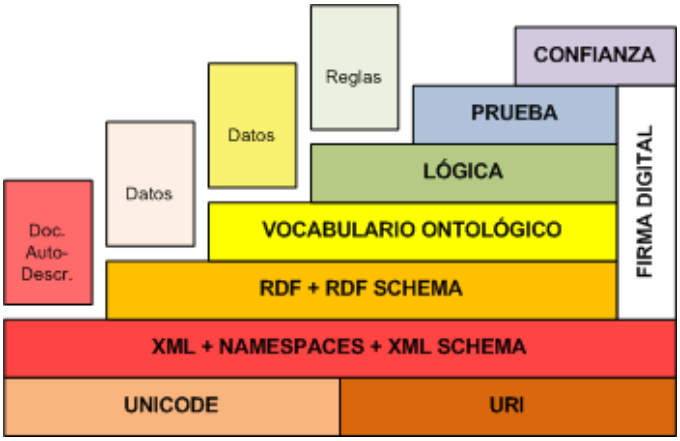


Figura 2.4: Arquitectura de la Web Semántica

y componentes de la Web con información semántica explícita, independiente de la presentación al usuario, y susceptible de ser procesada de forma automática por cualquier servicio software. Se considera que la Web Semántica añadirá estructuras formales y estándar a los contenidos electrónicos, creando un entorno donde los agentes software puedan buscar, explotar e intercambiar la información de manera eficiente [10]. La Web Semántica es una visión: la idea de tener los datos en la Web definidos y enlazados de manera que puedan ser empleados por las máquinas, no sólo con propósito de visualización. La Web Semántica está cambiando la forma de trabajar. Por ejemplo, el desarrollo de un servicio software se podría basar en las búsquedas de los componentes necesarios en la Web y en una especificación formal que describa cómo se orquestarán éstos para que el servicio cumpla con los requisitos especificados.

Tim Berners-Lee propuso una primera arquitectura por capas para implementar la Web Semántica (véase figura 2.4).

Las capas URI (Uniform Resource Identifier)[47] y Unicode permiten identificar unívocamente a cada entidad de conocimiento y asegura su correcta representación en cualquier idioma, respectivamente.

El lenguaje de etiquetas XML [48], los espacios de nombres (NameSpaces) y los XML Schema [49] son ampliamente empleados en la Web. Por ejemplo, el lenguaje XHTML en el que se representan la mayoría de páginas son documentos XML. Aunque este lenguaje añade información sobre la estructura de la información, no indica nada sobre el significado de esta estructura de datos. Esta capa se incluye en la arquitectura de la Web Semántica para que las diversas tecnologías semánticas puedan ser serializadas en XML, pu-

diendo ser transferidas sin necesidad de ninguna infraestructura adicional, ya que para las tecnologías actuales la información semántica sería otro estándar XML.

El lenguaje RDF (Resource Description Framework) [50] permite describir los recursos de la Web Semántica. Un recurso es cualquier cosa que pueda tener otros elementos enlazados. En este lenguaje, la información semántica se almacena en forma de tripletas. Estas tripletas se componen de sujeto, predicado y objeto. El sujeto es el origen de la relación, el predicado expresa la relación en sí y el objeto es el destino de esa relación. Cada uno de los elementos de la relación se identifican unívocamente con una URI. Una estructura de información RDF se define como un grafo dirigido y etiquetado, formado por el conjunto de todas sus tripletas. El sujeto y el objeto serían nodos, y el predicado la arista que los relaciona dirigida desde el sujeto hasta el objeto. RDF Schema (RDFS) [51] es una extensión del lenguaje RDF que permite definir los recursos como clases, organizar a éstas en jerarquías, definir las relaciones entre clases como propiedades y definir sus dominios y rangos.

Estos lenguajes siguen siendo insuficientes para representar el conocimiento de muchos dominios. Por ese motivo se añade una capa de vocabulario ontológico que mejora la expresividad de la capa anterior, proporcionando nuevos conceptos, relaciones y propiedades.

La capa lógica permite definir reglas que ayuden a los razonadores a inferir nuevo conocimiento sobre el que ya disponen. Es decir, son capaces de convertir el conocimiento implícito del modelo semántico en explítico.

Las últimas capas aún no han sido desarrolladas. En general, estas capas pretenden establecer niveles de seguridad que evalúen los recursos y la fiabilidad de los mismos gracias a la ayuda de la firma digital.

A continuación se van a describir con más detalle los principales elementos y tecnologías de la Web Semántica.

2.2.1 Resource Description Framework (RDF)

El objetivo principal de RDF es definir un lenguaje estándar para el intercambio de recursos a través de Internet. Un recurso puede ser cualquier cosa, tanto real como abstracta. En RDF, cada recurso se describe en forma de expresiones sujeto-predicado-objeto, también llamadas tripletas. A continuación se describen cada uno de los componentes de estas tripletas:

- El sujeto de la tripleta permite identificar el recurso (documento, persona, objeto físico, concepto abstracto, etc.). Esa identificación se hace por medio de una URI unívoca.

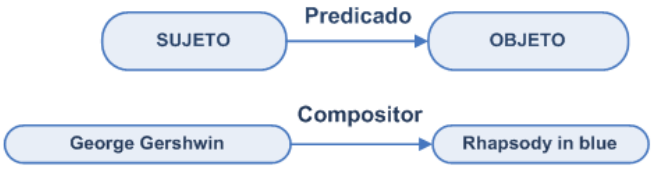


Figura 2.5: Grafo RDF

- El objeto de la tripleta puede ser un literal (cadena de caracteres, un valor numérico, una fecha, etc) o ser otro recurso que vendrá identificado de nuevo por una URI.
- El predicado de la tripleta indica el tipo de relación existente entre el sujeto y el objeto. El predicado también se identifica con una URI. Cuando el objeto es un literal, la relación describe una propiedad concreta del recurso. En el caso de que el objeto sea otro recurso, el predicado representa la relación entre estos dos recursos.

Como se puede apreciar en la figura 2.5, RDF constituye un modelo de datos sencillo para la descripción de recursos y sus relaciones. RDF usa URI para identificar los recursos y sus relaciones. Cuando se va a almacenar un tipo de datos básico se usan literales que opcionalmente pueden ir acompañados de una etiqueta del idioma y que se suelen asociar a la especificación de tipos de datos de XML Schema. En algunos casos, para almacenar alguna información de modelos más complejos, RDF permite que se pueda identificar un recurso sin usar una URI, sino simplemente un identificador local. Este tipo de recursos se llaman nodos en blanco (*blank nodes*).

2.2.2 Ontologías

Dentro del campo de la ingeniería del conocimiento existen tecnologías que no sólo facilitan la representación del conocimiento, sino que también permiten la reutilización y la compartición de componentes del conocimiento [52]. Una de estas tecnologías son las ontologías, que dan cuenta del conocimiento del dominio en términos estáticos. Una ontología puede definirse como una descripción explítica y formal de una conceptualización [53].

Las ontologías se han utilizado para representar conocimiento en distintos tipos de dominios, como los clínicos [54; 55], las memorias organizacionales [56; 57], la gestión del conocimiento [58], la bioinformática [59], el ámbito financiero [60] o el e-Learning [61]. Las ontologías permiten que el conoci-

miento que éstas contienen pueda reutilizarse y compartirse, por lo que su uso conlleva una reducción del esfuerzo necesario para implementar sistemas expertos.

Las ontologías son una tecnología clave para la Web Semántica [62], ya que unen la comprensión simbólica humana con el procesamiento por ordenador. Desde los años 90 han sido un tema puntero de investigación, y han sido estudiadas en varias comunidades científicas como ingeniería del conocimiento, procesamiento del lenguaje natural o representación del conocimiento.

El motivo de su creciente popularidad es principalmente lo que prometen: una comprensión compartida y común en un dominio que puede ser comunicado entre personas y aplicaciones informáticas. El uso de las ontologías ofrece una oportunidad de mejorar las posibilidades de realizar cualquier tarea relacionada con la gestión de la información y el conocimiento.

Para representar las ontologías se ha intentado establecer un lenguaje común. Hay una propuesta de estandarización de lenguaje ontológico llamado OWL [63], especificado por un grupo de trabajo del consorcio W3C (URL), que ayudaría a solucionar los impedimentos actuales para la construcción cooperativa de ontologías entre diferentes plataformas de construcción que usen diferentes modelos. Sin embargo, la mayor limitación actual de los trabajos en ingeniería ontológica es que están centrados en la taxonomía, dejando a un lado otras (no menos importantes) relaciones ontológicas formales para describir contenidos (y por tanto, realizar consultas) en portales de conocimiento.

Otro lenguaje muy extendido para representar ontologías es RDF Schema, que como se ha comentado es una extensión de RDF que proporciona la posibilidad de definir clases, jerarquías, propiedades y restricciones de dominio y rango. Como se describe en la sección 2.2.2.2, OWL proporciona un lenguaje más complejo.

2.2.2.1 RDF Schema (RDFS)

RDFS es una extensión de RDF que permite representar el esquema de los datos usando RDF como lenguaje de descripción. RDFS soporta clases, propiedades, jerarquías de clases y propiedades, y restricciones de rango y dominio para las propiedades. A continuación se decriben brevemente los recursos que pueden ser definidos:

- *rdfs:Class*, para definir clases o conceptos.
- *rdfs:Resource*, para definir recursos.
- *rdf:Property*, para modelar propiedades.

- *rdf:type*, para indicar la relación "instancia de".
- *rdfs:subClassOf*, para modelar jerarquías de clases o conceptos.
- *rdfs:subPropertyOf*, para construir jerarquías de propiedades.
- *rdfs:domain*, para restringir las instancias sujeto de una propiedad como instancias de un conjunto de clases concretas.
- *rdfs:range*, para restringir las instancias objeto de una propiedad como instancias de un conjunto de clases concretas.

RDFS es el lenguaje en el que se han desarrollado muchas ontologías usadas en redes sociales como FOAF [64] o SIOC [65], o para representar el conocimiento de una organización como SKOS [66].

2.2.2.2 Web Ontology Language (OWL)

OWL (Web Ontology Language) es un lenguaje para la creación de ontologías que facilita la interpretación del contenido Web en mayor medida que los estándares XML, RDF, y RDF Schema, proporcionando un vocabulario adicional con enriquecimiento semántico. De esta manera, OWL permite organizar los recursos en clases y crear instancias de esas clases. Permite la definición de propiedades, tanto de tipo primitivo como de tipo objeto, conocidos popularmente como atributos y relaciones, respectivamente. También define jerarquías, restricciones de dominio, rangos, cuantificación existencial y universal, y cardinalidad, entre otras.

En OWL existen tres variantes con diferente expresividad y capacidad a la hora de inferir conocimiento (ver figura 2.6 izquierda). Un grupo de desarrolladores o usuarios puede elegir la forma más apropiada de representación en OWL según las características siguientes [63]:

- OWL Lite: indicado cuando se necesita una clasificación jerárquica y restricciones simples. Esta variante es menos compleja que OWL DL, pero su expresividad es más reducida.
- OWL DL: indicado en casos donde se requiera mayor expresividad, conservando la computacionalidad y resolubilidad, es decir, que cualquier conclusión pueda alcanzarse mediante reglas de inferencia y alcanzables en un tiempo finito respectivamente. En la versión 2 de OWL [67] se introducen tres perfiles para OWL DL con distinta expresividad orientados a aplicaciones del mundo real (ver figura 2.6 derecha):

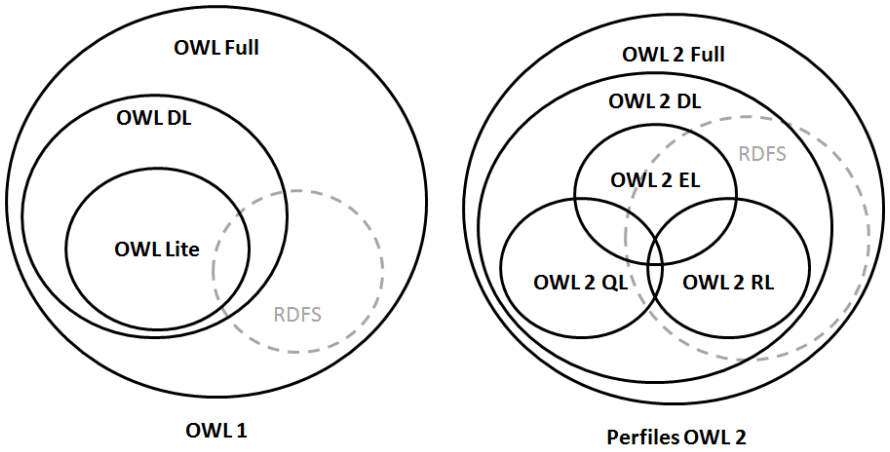


Figura 2.6: Lenguaje OWL

  - OWL 2 EL: perfil cercano a la lógica descriptiva EL++ [68], asegura un tiempo polinomial para todos los problemas estándar de razonamiento. Es adecuado para aplicaciones que requieren ontologías muy grandes y donde se puede sacrificar poder de expresividad con tal de obtener el mejor rendimiento.
  - OWL 2 QL: destinado a aplicaciones que utilizan ontologías relativamente ligeras para organizar gran número de instancias donde es útil acceder a los datos directamente a través de consultas relacionales SQL. Este perfil permite que las consultas se puedan ejecutar en un tiempo computacional razonable.
  - OWL 2 RL: permite la implementación de algoritmos con tiempo de razonamiento polinomial usando tecnologías de bases de datos basadas en reglas que operan directamente sobre RDF. Es particularmente adecuado para aplicaciones que usan ontologías ligeras que almacenan gran número de instancias y donde es útil manejar los datos en RDF.

- OWL Full: indicado en casos de máxima expresividad y libertad sintáctica de RDF, pero no ofrece garantías computacionales. OWL Full permite a una ontología aumentar el significado de los vocabularios predefinidos en RDF u OWL. Como consecuencia resulta imposible diseñar un sistema que sea capaz de realizar razonamiento sobre un lenguaje de estas características. OWL Full puede ser visto como una extensión de RDF, mientras OWL Lite y OWL DL puede ser visto como extensiones de vistas restringidas de RDF.

Las ontologías empleadas en la tesis se encuentran definidas en el lenguaje OWL, más concretamente en OWL-DL.

2.2.2.3 Ingeniería Ontológica

La ingeniería ontológica es la disciplina que investiga los principios, métodos y herramientas para el diseño, desarrollo, mantenimiento y reutilización de ontologías. Por este motivo se han desarrollado varias metodologías de construcción de ontologías que pueden clasificarse en:

- Creación manual de ontologías. En esta propuesta de construcción, el desarrollo de la ontología se hace de forma manual y se va construyendo desde cero. En la literatura se pueden encontrar diferentes metodologías que siguen este enfoque. En [69], en una primera fase se extraen de forma manual aquellas entidades de negocio que se encuentran en distintas fuentes. En una segunda fase, se usa software de procesamiento del lenguaje natural validado manualmente para extraer más entidades y en una tercera fase únicamente se usa el software para completar la ontología. En [70], la creación de la ontología se divide en: (1) identificar el propósito de ésta, (2) capturar los conceptos, relaciones entre estos y terminología y (3) codificar la ontología. En general, las metodologías que se encuentran en este enfoque [71; 72; 73] extraen las entidades principales, reutilizan modelos ontológicos existentes sobre esas entidades, extendiéndolos o acotándolos dependendiendo del problema concreto a resolver. En muchos casos se usan herramientas de procesamiento del lenguaje natural para extraer conceptos y relaciones de las fuentes de conocimiento originales.

- Creación colaborativa de ontologías. En los casos anteriores, las metodologías están pensadas para casos donde el equipo de diseño y desarrollo de la ontología se encuentra en un entorno centralizado. En los casos en los que el equipo de desarrollo se encuentra geográficamente disperso se han definido una serie de metodologías para la construcción de ontologías de forma colaborativa. En [74] se definen cinco pasos principales: (1) construcción; (2) adaptación local; (3) análisis; (4) revisión; (5) actualización local. En este esquema, existe una ontología global y varias comunes por cada equipo de trabajo. Cada cierto tiempo se analizan esas ontologías locales y se deciden las modificaciones en la ontología global.

- Reutilización de ontologías. En esta metodología se construyen ontologías automáticamente a partir de la reutilización de otras. Las mejo-

res prácticas en ingeniería ontológica recomiendan reutilizar ontologías existentes, y cuando esto no es posible, crear ontologías modulares [75]. La implementación de estas recomendaciones permite generar una red de ontologías relacionadas entre sí a partir de la reutilización de sus conceptos. La *OBO Foundry* [76] ha desarrollado una serie de principios de modularidad, ortogonalidad y reusabilidad para construcción de ontologías.

- Aprendizaje de ontologías (*"Ontology Learning"*). En esta metodología se propone la construcción automatizada de ontologías a partir del análisis de grandes volúmenes de información. Este enfoque se apoya en las disciplinas de Procesamiento del Lenguaje Natural, la Inteligencia Artificial y el Aprendizaje Computacional [77]. En [78] se definen cinco pasos principales para la construcción automatizada de ontologías: (1) identificación de los términos y sus sinónimos, (2) descubrimiento de conceptos a partir de los términos, (3) extracción de relaciones taxonónimas entre los conceptos, (4) generación de relaciones semánticas no taxonómicas entre conceptos, y (5) generación de reglas y axiomas entre los elementos ontológicos.

2.2.2.4 Técnicas de razonamiento

El modelado conceptual (ontológico) aborda la cuestión de la descripción declarativa y abstracta de la información de un dominio de aplicación, su vocabulario relevante y cómo restringir el uso de los datos. Los correspondientes lenguajes abstractos de representación (por ejemplo OWL) soportan la comprensión de tales descripciones, su rápido desarrollo, mantenimiento y reutilización. A raíz de dichos lenguajes han surgido diversas herramientas que permiten la realización de razonamientos y comprobación formal de las estructuras de conocimiento como Pellet [79], Fact++ [80] y Hermit [81].

2.2.2.5 Anotación semántica

La anotación semántica es un proceso que consiste en añadir información semántica a contenidos de cualquier tipo. Es decir, consiste en enlazar contenidos con entidades semánticas que los describan [82]. Los procesos de anotación se dividen en [83]:

- Anotación manual. En este proceso será el usuario el que anote manualmente las entidades de negocio sobre el modelo ontológico. Este proceso necesita de expertos del dominio capaces de generar anotaciones precisas.

- Anotación automática. Se basa en la búsqueda de términos en textos. No requiere la interacción de los usuarios, pero si de unas reglas de anotación, de cuya calidad dependerá la precisión de la misma.

- Anotación semi-automática. Este caso se basa en el proceso anterior. La principal diferencia es que en vez de anotar automáticamente en base a las reglas de anotación es el usuario el que valida las anotaciones.

2.2.2.6 Etiquetado semántico

La interacción entre usuarios y la inmensa mayoría de plataformas sociales (blogs, wikis, etc.) se lleva a cabo a través de campos en texto abierto, y el único modo de estructurar los datos y organizar los contenidos se realiza por medio de palabras clave. Estas palabras clave sirven para indexar los contenidos, facilitando la búsqueda y la navegación entre diferentes contenidos, aunque es el uso de una misma palabra clave por más de un usuario lo que permite generar esquemas de clasificación cooperativos [84]. Además, las palabras clave no suelen basarse en modelos formales y carecen de modelos semánticos bien definidos.

Para solventar esta limitación, en [85] se propone MOAT (Meaning of a Tag), un novedoso mecanismo para representar folksonomías asociando las etiquetas con su significado, por ejemplo con recursos semánticos identificados por una URI.

2.2.2.7 Similitud semántica

Gracias a los procesos de anotación semántica se pueden comparar entidades que de otra manera no serían comparables [86]. Por ejemplo, si se quisiese comparar una imagen de una persona con un modelo de datos que refleja los rasgos de esa persona. La similitud semántica va a obtener la probabilidad de que esa imagen (anotada semánticamente) sea de la persona que se ha descrito en el modelo de datos. Un método de similitud semántica es una función que devuelve cómo de parecidos semánticamente son dos entidades. Existen dos propuestas principales para el cálculo de la similitud semántica:

- Métodos basados en aristas [87]: se basan en contar el número de aristas existentes entre los dos términos comparados. La técnica más común es la de la distancia. Este método sólo es válido en ontologías donde los nodos y las aristas se distribuyen de forma uniforme y donde las aristas al mismo nivel en la ontología son equivalentes semánticamente.

- Métodos basados en nodos [88]: en este caso se usan los nodos y sus propiedades. Se basan en comparar las propiedades relacionadas entre las entidades de las que se quiere evaluar su similitud, ya sea directamente, con sus ancestros o con sus descendientes.

Estos métodos son útiles para comparar elementos de la misma ontología. Existen otros para medir la similitud entre ontologías [89]:

- Los métodos basados en características miden la similitud basándose en términos similares de la taxonomía.
- Los método híbridos son una combinación de las propuestas anteriores [90]. La similitud se calcula en base a sinónimos, términos vecinos y propiedades de los términos.

Una consideración importante al diseñar funciones de similitud es que asignen similitudes mayores a aquellos términos que estén más cerca (en términos de distancia del camino en el grafo) y sean más específicos (más profundos en la jerarquía), que a términos que estén igualmente cercanos pero más altos jerárquicamente.

2.2.3 Lenguajes de consulta: SPARQL

SPARQL (SPARQL Protocol And RDF Query Language) [91] define un lenguaje de consulta estándar para modelos de datos RDF(S). Actualmente está recomendado por el W3C como lenguaje de consulta para la Web Semántica. La funcionalidad de SPARQL se encuentra definida en tres especificaciones diferentes:

- SPARQL Query Language: en esta especificación se explica la sintaxis para la creación de sentencias y su concordancia.
- SPARQL Protocol for RDF: utiliza WSDL 2.0 para definir protocolos HTTP y SOAP para consultar remotamente bases de datos RDF.
- SPARQL Query Results XML Format: describe el formato XML de cómo se devuelven los resultados de una consulta.

El lenguaje SPARQL posee tres componentes importantes: URI, literales y variables procedentes del lenguaje RDF.

- URI: sirven para especificar unívocamente los recursos y propiedades de la ontología. Se definen entre "<" y ">".

- Literales: se describen como una cadena de caracteres encerradas entre comillas.
- Variables: estas variables son globales y además deben de ser prefijadas por "?" no formando parte del nombre de la variable.

SPARQL ha sido diseñado para ser escalable y permite hacer consultas sobre orígenes de datos distribuidos, independientemente del formato.

Otra característica de SPARQL es que puede ser serializado en RDF haciendo uso de SPIN [92]. SPIN permite definir reglas y restricciones en modelos de datos semánticos basados en RDF.

2.2.4 Linked Data

El término *Linked Data* representa un conjunto de buenas prácticas para publicar y enlazar datos estructurados en la Web. Estas prácticas fueron propuestas por Berners-Lee en [93] y se conocen como principios *Linked Data*:

- Utilizar URI para identificar los recursos.
- Utilizar HTTP URI para que se puedan localizar los identificadores de los recursos.
- Proporcionar información útil asociada a la URI de los recursos.
- Enriquecer la descripción de los recursos enlazando con otros recursos.

Linked Data usa RDF para representar la información y enlaces Web para enlazar las diferentes fuentes. Con este enfoque se pretende alcanzar un espacio único global de intercambio de datos llamado "Web de Datos" que contiene datos tan variados como personas, empresas, libros, publicaciones científicas, música, televisión, cine, genes, ensayos clínicos, etc [94]. Esto ha llevado a la aparición de nuevas aplicaciones que pueden explotar esos datos permitiendo obtener respuestas más elaboradas según van apareciendo nuevas fuentes de datos en la Web [95].

El ejemplo más notable de uso de los principios de *Linked Data* es el proyecto *Linked Open Data* (LOD) de W3C [96], que tiene como objetivo impulsar la Web de Datos identificando conjuntos de datos existentes publicados bajo licencias abiertas, convertirlos a RDF siguiendo los principios de *Linked Data* y publicarlos en la Web. Es un proyecto abierto a cualquiera que quiera publicar datos según *Linked Data*, lo que ha favorecido su éxito. El proyecto comenzó en 2007 con 13 conjuntos de datos y en abril de 2014

se contabilizaron un total de 1014 conjuntos. En [97] se puede ver el diagrama que representa los conjuntos de datos publicados y su interrelación en la actualidad.

Berners-Lee propuso un esquema de despliegue de *Linked Open Data* con cinco niveles puntuados con estrellas [93]:

- ⋆ Los datos están disponibles en la web (en cualquier formato) con una licencia abierta. Se clasifican como Open Data.
- ⋆⋆ Los datos están disponibles de forma estructurada por lo que pueden ser procesados por un ordenador.
- ⋆⋆⋆ Cumple el nivel 2(⋆⋆) y además los datos están en un formato no propietario.
- ⋆⋆⋆⋆ Cumple todos los niveles anteriores pero además usa estándares abiertos de la W3C (RDF y SPARQL) para identificar los recursos de manera que estos puedan ser enlazados.
- ⋆⋆⋆⋆⋆ Cumple todos los puntos anteriores y además los datos están relacionados con otros conjuntos de datos que les proporcionan más contexto.

Es decir, con las tecnologías de la Web Semántica se consiguen los niveles más altos de este esquema de estrellas. Utilizando RDF y haciendo un uso apropiado de URI se alcanza el nivel de cuatro estrellas y enlazando los conjuntos de datos con recursos externos se llega al nivel cinco.

2.2.5 Métodos de transformación semántica

A continuación se describen los principales métodos para llevar la información proveniente de múltiples orígenes de datos heterogéneos a modelos semánticos formales y bien estructurados. Las metodologías de transformación de información a representaciones semánticas siguen enfoques semejantes. Parten de unos datos fuente, que son transformados siguiendo un modelo semántico. El proceso de extracción y transformación de la información se basa en la definición de correspondencias entre el modelo que siguen los datos origen y el modelo semántico de salida.

2.2.5.1 Definición del modelo semántico

Un modelo semántico debe definir formalmente el dominio que pretende representar. Para la definición de este modelo se distinguen dos casos: (1) el

modelo de salida se diseña a partir de los datos fuente, y (2) el modelo de salida se define usando técnicas de ingeniería ontológica al dominio donde se aplicará, sin tener en cuenta los datos origen.

Tim Berners-Lee propuso una primera metodología [98] para relacionar la Web Semántica y los modelos de datos relacionales, a través de una correspondencia directa:

- Una fila de una tabla es un nodo RDF.
- Cada campo de la tabla es una propiedad RDF.
- Cada valor de una fila para un campo concreto es un valor de un determinado tipo para el modelo semántico.

En [99] se amplía este modelo para generar un esquema RDFS a partir de una base de datos relacional:

- Cada tabla R se mapea con una clase RDFS C.
- Por cada entrada en la tabla R se crea un nodo RDF I cuyo tipo es C, es decir, se crea una instancia de la clase C.
- Por cada campo *field* en la tabla R se crea una propiedad RDF P.
- Para cada entrada en la tabla R, el valor del atributo *field* se asocia al valor de la propiedad P para el nodo I.

La W3C creó un grupo de trabajo, RDB2RDF Working Group [100], para la estandarización de la definición de correspondencias entre las bases de datos relacionales y los esquemas semánticos como RDF u OWL. Uno de los resultados fue Direct Mapping [101], una metodología para asociar datos relacionales a RDF.

2.2.5.2 Definición de correspondencias

El primer paso para poder transformar datos en información semántica consiste en definir correspondencias entre ambos modelos. Esta definición puede ser un proceso manual, dirigido por el usuario, o puede ser el resultado de un proceso de equivalencia de esquemas.

A continuación se describen los dos lenguajes de definición de correspondencias más extendidos:

- **R2RML: RDB to RDF Mapping Language.** El lenguaje R2RML [102], creado por el consorcio RDB2RDF, sirve para expresar correspondencias entre bases de datos relacionales y conjuntos de datos RDF. Las correspondencias en este lenguaje están diseñadas para: (1) construir repositorios RDF a partir de las bases de datos relacionales, (2) acceder a la base de datos relacional a través de un punto de acceso virtual SPARQL y (3) crear una interfaz *Linked Data*.

- **D2RQ Mapping Language.** Este lenguaje permite definir la relación entre un esquema de base de datos relacional y un vocabulario RDFS u OWL de un modo declarativo [103]. Una correspondencia D2RQ está formada por un "*ClassMap*", que asocia una clase de la ontología con una URI construida con valores de la base de datos. Cada "*ClassMap*" contiene un conjunto de "*PropertyBridge*", que especifican las propiedades y relaciones de la clase.

2.2.5.3 Transformación de datos

La transformación de datos es el proceso en el que la información del modelo de entrada es convertida en un modelo de salida semántico, normalmente basado en RDF, RDFS u OWL. Estos procesos se clasifican en [99]: (1) totalmente automáticos, (2) semi-automáticos y (3) manuales. Normalmente, en los procesos automáticos el modelo de salida también se genera automáticamente, ya que sería muy complejo realizar este proceso de forma automática cuando el modelo de salida ha sido producido manualmente o por otra herramienta. Atendiendo al tipo de acceso a la información semántica, estas herramientas de transformación se clasifican en: (1) las que hacen una transformación completa de la fuente de información origen al modelo semántico destino y (2) las que crean vistas virtuales que pueden ser consultadas a través de lenguajes de consulta como SPARQL.

A continuación se describen las principales herramientas de transformación semántica:

- **D2RQ.** Es una plataforma que permite consultar datos albergados en bases de datos relacionales en SPARQL gracias a la generación de grafos RDF virtuales [104]. Es un proceso totalmente automático.

- **Triplify.** Es una herramienta que permite publicar en Linked Data información proporcionada por consultas SQL [105]. En muchas ocasiones se usan vistas SQL para hacer la transformación más flexible. Además tiene la posibilidad de anotar las columnas con vocabularios existentes. Es un proceso semi-automático.

- **Linked Data Views de Virtuoso.** OpenLink Virtuoso [106] es un sistema gestor de bases de datos que permite trabajar con múltiples modelos de persistencia (relacional, XML, objeto-relacional, virtuales y RDF). Entre sus múltiples servicios, la herramienta Linked Data Views [107] permite consultar fuentes de datos relacionales con SPARQL a partir de la definición de correspondencias. Es un proceso totalmente automático y permite integrar la información de diferentes fuentes siempre que estén en alguno de los motores de persistencia de Virtuoso.

- **XS2OWL.** Este modelo de transformación permite la representación de esquemas XML en sintaxis OWL [108]. Gracias a esa generación del modelo de salida semántico permite la consulta de bases de datos XML que sigan ese esquema a través de SPARQL, que es traducido a XQuery (lenguaje de consulta para XML) dirigido por la ontología. Es un modelo automático.

- **RDB2OWL.** Es una propuesta para transformar información almacenada en bases de datos relacionales a una representación basada en RDF o una ontología OWL preexistente [109]. Las correspondencias entre el modelo de entrada y el de salida se almacenan en una base de datos relacional. Estas correspondencias permiten la generación automática de un script SQL que transforma los datos relacionales en tripletas RDF o en instancias OWL. El modelo de generación de correspondencias es manual, por lo que para ontologías grandes puede ser tedioso.

- **Karma.** Es una herramienta que permite asociar al modelo origen ontologías existentes para generar una representación semántica de la fuente de los datos [110]. Gracias a esa generación de un modelo semántico de origen, se pueden generar automáticamente correspondencias con el modelo destino. En este modelo se define un proceso semi-automático, siempre y cuando se disponga de un modelo origen anotado semánticamente.

- **Populous**. Populous [111] es una herramienta que sirve como asistente a la creación de ontologías. Esta herramienta hace uso de patrones para guiar el proceso de recogida del conocimiento y cómo éste va a incorporarse al modelo ontológico. Una vez creada la ontología, Populous es capaz de importar datos provenientes de hojas de cálculo y de archivos tabulares. Las columnas de estos archivos deben estar asociadas a las variables de la ontología.

- **Sistema OGO.** En [112] se propone una herramienta para la integración de varios repositorios relacionales en un único almacén semántico basado en una ontología. La transformación se lleva a cabo a través de la definición de correspondencias entre los distintos esquemas relacionales y la ontología global que modela el dominio.

- **SWIT.** Esta herramienta [113] permite transformar un modelo de entrada (base de datos relacional o XML) en un modelo de salida semántico que puede estar en RDF o en OWL. La transformación se realiza a partir de la definición de correspondencias entre los esquemas de entrada y de salida. Las correspondencias se definen de forma declarativa, permitiendo que se puedan reutilizar y que, una vez definidas, el método de transformación sea automático. La transformación se realiza en tres fases: (1) definición de las reglas de mapeo entre los campos de la base de datos y la ontología, (2) generación de la información en OWL, e (3) importación de los datos OWL en un repositorio semántico.

SWIT alcanza el nivel 5 estrellas propuesto por Berners-Lee para la publicación de datos en formato abierto. El resto cumplen el nivel 4 estrellas, a excepción de D2RQ y XS2OWL, que no llegan realmente a transformar los datos a un estándar abierto.

2.2.5.4 Discusión

Como se ha comentado, existen diferentes herramientas para la transformación de datos provenientes de diversos sistemas (bases de datos relacionales, archivos XML, etc.) en información semántica. En esta sección se analizan cuáles son los modelos más convenientes en cada caso, usando la flexibilidad como la principal medida de las diferentes herramientas.

Muchas de las herramientas que se han descrito sólo realizan una transformación sintáctica de los datos, guiada por el modelo lógico de los sistemas de origen. La principal ventaja de estas herramientas es que el proceso de transformación de las fuentes de datos a información semántica es totalmente automático, lo que simplifica notablemente este proceso. Sin embargo, este tipo de herramientas no son adecuadas en entornos muy heterogéneos en los que hayan múltiples orígenes de datos, ya que no permitirán el uso de modelos semánticos formales que se acerquen a la representación más cercana al usuario del dominio objeto de explotación.

En otras herramientas la transformación es guiada por la ontología gracias a una definición de correspondencias entre los modelos de origen y el modelo semántico. En estas herramientas se pueden definir ontologías inde-

pendientemente del modelo lógico de los orígenes de datos, lo que proporciona una gran usabilidad. El principal problema de estas soluciones es que, dependiendo del volumen del dominio, esos mapeos entre modelos pueden ser bastante costosos. Entre las diferentes herramientas disponibles para la transformación de información destacan Populous y SWIT. Ambos hacen uso de patrones que pueden reutilizarse para definir correspondencias, reduciendo así la complejidad de esta tarea. Populous acepta como fuentes de entrada ficheros tabulares, y SWIT permite ficheros XML y conexión a diversas bases de datos relacionales.

2.2.6 Inteligencia de Negocio y Web Semántica

La convergencia entre la Inteligencia de Negocio y la Web Semántica se ha visto plasmada en diferentes trabajos científicos. La mayoría de los estudios que se han realizado han utilizado tecnologías semánticas como apoyo a diferentes fases de la IN, principalmente vinculados a: (1) los procesos de extracción, transformación y carga (ETL) [12]; (2) al modelado formal del dominio de negocio [11]; y (3) para validación de los datos con razonadores [13].

En [114] se hace un estudio de las posibles sinergias que se pueden encontrar entre estas dos áreas de investigación, analizando los puntos fuertes y débiles del uso de tecnologías semánticas en el ámbito de la IN. En este trabajo, parten de las tres capas típicas que se pueden encontrar en la mayoría de soluciones de IN: la integración de datos, la entrega de información y la capa de análisis. En este trabajo mencionan los repositorios RDF como posible fuente de datos, y la anotación semántica como fuente de generación de esos datos. También comentan cómo se pueden usar ontologías para hacer un proceso de extracción, transformación y carga guiado por éstas. Por último, resalta la potencia de los razonadores semánticos para agregar información o generar nuevo conocimiento.

En el ámbito de los procesos ETL, en [15] se describe una propuesta para usar representaciones basadas en grafos como modelos conceptuales. Las transformaciones son definidas sobre estos modelos que pueden ser ontologías del dominio.

En trabajos como [14] se analizan las limitaciones de las soluciones de integración de datos sobre diferentes dominios, como pasa por ejemplo en biomedicina, donde se tienen modelos de diagnósico, terapeúticos, epidemiológicos, genéticos, etc. Además se propone un almacén de datos semántico (*Semantic Data Warehouse*), que pueda guardar anotaciones de diversas ontologías interrelacionadas.

A continuación se describe uno de los proyectos que más ha trabajado en el uso de tecnologías de la Web Semántica para soluciones de IN.

2.2.6.1 Proyecto EU MUSING

MUSING es el acrónimo de "*MUlti-industry, Semantic-based next generation business INtelliGence*". Su objetivo principal es integrar las tecnologías de la Web Semántica y del Procesamiento del Lenguaje Natural con metodologías basadas en reglas y enfoques estadísticos para mejorar las capacidades de adquisición de conocimiento y de razonamiento en aplicaciones de IN [115]. El proyecto se desarrolló entre 2006 y 2010, con una financiación superior a los catorce millones de euros [115].

Las principales tecnologías que se plantearon en este proyecto son:

- Gestión del conocimiento basada en Web Semántica.
- Ingeniería ontológica.
- Razonamiento.
- Estadísticas bayesiana.
- Minería de datos.
- Herramienta de integración de datos para anotación automática y semiautomática.
- Sistemas de evaluación avanzados.
- Aplicación a estándares y soluciones que se aplican en diversos sectores de la industria.

Los principales resultados que han obtenido están orientados a la parte de integración de datos. En [11; 116; 117] presentan varios modelos para la extracción de la información basada en ontologías. En [11] usan técnicas de procesamiento del lenguaje natural para la integración de la información en soluciones de IN. En [116] desarrollan cómo se pueden modelar y explotar semánticamente el estándar XBRL (Lenguaje extensible de informe de negocios) [118]. También han integrado análisis estadísticos basados en el teorema de Bayes para analizar los riesgos de los sistemas operacionales usando ontologías del dominio.

2.3 Web 2.0

La Web 2.0 y las redes sociales son muy importantes para la investigación en sistemas colaborativos. El término Web 2.0 está asociado a una serie de tecnologías y patrones de diseño que permiten desarrollar aplicaciones Web que facilitan la compartición de información, interoperabilidad y la colaboración [32]. Se pueden ver las redes sociales como una plataforma que ofrece múltiples servicios con tecnologías Web 2.0. Siendo éstas fundamentales para el éxito de las aplicaciones Web 2.0 [84]. Las redes sociales son muy útiles en la IN, ya que pueden verse como una colección de individuos (normalmente personas) y organizaciones enlazadas por un conjunto de relaciones [119]. Las entidades de las redes sociales son conocidas como nodos, y las relaciones entre éstas como aristas. Este tipo de tecnologías facilita que la identificación de la importancia que tiene un servicio, para un determinado usuario, venga calculada por el número de usuarios que están relacionados con él [84]. Las redes sociales han sido explotadas con éxito en diferentes contextos [120].

Como ya se ha comentado, las plataformas Web 2.0 son un servicio esencial para la IN según [2], ya que permiten generar foros de discusión, compartición y anotación de la información, y evaluación de la información generada.

2.3.1 Web Social Semántica

Antes del surgimiento de la Web Semántica, las redes sociales se han representado como un conjunto de nodos, propiedades de éstos y relaciones entre ellos [121]. Actualmente, la Web Semántica se ha convertido en una tecnología muy presente en las redes sociales. Prueba de ello son los estándares semánticos FOAF o SIOC.

1. FOAF [64] (*Friend of A Friend*) es una ontología codificada en RDF Schema que modela que una persona conoce a otra. Además, permite definir el perfil social de cada individuo.

2. SIOC [65] (*Semantically-Interlinked Online Communities*) también es una ontología codificada en RDF Schema. Es una extensión de FOAF que permite modelar comunidades de usuarios y cómo éstas se relacionan entre sí.

En los últimos años se ha visto que la Web Semántica y la Web 2.0 son complementos naturales uno del otro. La combinación de estas tecnologías ha dado lugar a lo que se denomina Web Social Semántica [122].

Numerosos estudios de investigación han combinado ambas tecnologías. En [84], los autores representan una red social como información que se comparte a través de relaciones semánticas. Estas relaciones permitirán la adquisición de nuevo conocimiento generando relaciones nuevas entre los datos cuando sea posible. Los autores de [123] presentan una propuesta más ambiciosa. Ellos anticipan la llegada de unos nuevos dispositivos sociales que serán ordenadores que ayuden a los usuarios en las tareas de administración de las redes en las que participan dejándoles a ellos únicamente la parte creativa. Para ello, los autores consideran que es importante que los contenidos de la Web se organicen como un grafo que interconecte a personas e ideas.

En [85], los autores proponen una nueva ontología, MOAT *(Meaning of a Tag)* que es un mecanismo semántico de representación de folksonomías asociando palabras clave con su significado. Por último OPO (*Online Presence Ontology*) [124], es una ontología que modela la presencia de un usuario en una determinada comunidad en línea.

En [125], el autor distingue entre "inteligencia acumulada" e "inteligencia colectiva". La Web 2.0 permite acumular inteligencia, pero si se combina con Web Semántica se puede alcanzar una inteligencia colectiva, esto es, un nivel más alto de "entendimiento" que permita dar respuesta a cuestiones no resueltas explícitamente por las contribuciones individuales de los usuarios.

A continuación se describen algunos ejemplos de aplicaciones sociales y semánticas:

- La Semantic MediaWiki es un claro ejemplo de aplicación que combina ambas tecnologías. Es un intento de formalizar en RDF la MediaWiki, facilitando que los contenidos puedan ser explotados por agentes software [126].

- SMOB [127] es una plataforma para microblogging usando tecnologías semánticas. SMOB hace uso de FOAF, para representar a los proveedores y consumidores de contenidos, y SIOC para modelar la naturaleza de las comunidades de una red de microblogging. También hace uso de MOAT y de OPO.

- SemSLATES [128] es una metodología que usa FOAF, SIOC y MOAT para modelar la gestión del conocimiento en un ámbito empresarial. Esta metodología incluye el uso de otras ontologías del dominio que permitan clasificar el conocimiento existente en los sistemas operacionales de las organizaciones.

2.4 Integración de la información

La cantidad de datos que las organizaciones almacenan en formato electrónico requiere cada vez más mecanismos para explotar esa información. Como ya se ha comentado, la heterogeneidad de la información con la que trabajan los sistemas operacionales hace que esa explotación requiera acceder a un repositorio unificado de los datos, lo que lleva a la creación de metodologías de integración [129; 130]. Se define integración de datos como el proceso de combinar datos provenientes de fuentes diferentes. Ese proceso de combinación genera un modelo global y unificado de los datos que puede ser explotado por un usuario del sistema [131]. El proceso de integración de la información supone más del 80 % del trabajo necesario para implantar una solución de IN [6]. Existen dos tipos de integración de la información:

- Integración horizontal. Este tipo de integración se realiza entre información complementaria que se encuentra en modelos de datos distintos. Por ejemplo, se pueden integrar datos de gestión de fondos de proyectos científicos con bases de datos de producción científica.

- Integración vertical. En este caso, la integración se realiza sobre datos de una misma línea de negocio, pero que se generan por organizaciones diferentes con sistemas operacionales diferentes. Por ejemplo, si una multinacional quiere comparar datos de ventas de varias sedes en distintos países.

Las propuestas de integración más usadas en sistemas operacionales se clasifican en términos del modelo de datos que usan: texto, datos estructurados o registros enlazados. En el primer caso, los sistemas de integración proporcionan sistemas de codificación basados en palabras clave, terminologías, etc. Si el sistema operacional trabaja con datos estructurados, las soluciones se dividen en almacenes de datos centralizados o los sistemas basados en vistas, que permiten obtener una vista de los datos en el momento en que son requeridos. La última propuesta considera los datos como elementos enlazables y navegables, lo que facilita que se pueda navegar desde un elemento a otro. Este tipo de integración es útil para enlazar la información pero no para su explotación conjunta.

A continuación se describen los modelos de integración de la información más empleados.

2.4.1 Integración basada en almacén de datos

Este modelo de integración guarda la información de los sistemas operacionales en un único almacén de datos que ha sido optimizado para su explotación [129]. Esto permite que los procesos de búsqueda y análisis de los datos se hagan sin afectar al rendimiento de los sistemas operacionales. Este tipo de sistemas requiere la definición de correspondencias entre las fuentes de datos y el esquema global. Este proceso se ha denominado ETL (Extracción, Transformación y Carga) [132] y existen múltiples soluciones en el mercado [5]. Este proceso consiste en:

1. Extracción: obtención de los datos necesarios de los diferentes sistemas de información, que pueden ser internos o externos a la organización.

2. Transformación: filtrado, limpieza, depuración y homogeneneización de los datos.

3. Carga: actualización de los datos y de los posibles metadatos en el almacén de datos.

Integrar toda la información en un único almacén y en estructuras que estén optimizadas para su explotación (aunque la información esté desnormalizada) mejora la eficiencia de las consultas. Además, el preprocesado que se hace antes de integrar la información permite filtrar, validar y modificar los datos obtenidos de las diferentes fuentes [132]. Por el contrario, estos sistemas tienen problemas de mantenimiento ya que suelen tener volúmenes de datos demasiado grandes, y resulta costoso mantenerlos actualizados, ya que tienen que actualizarse periódicamente en busca de modificaciones y nueva información en las fuentes operacionales.

2.4.2 Integración basada en mediadores

Los sistemas basados en mediadores, también conocidos como sistemas basados en vistas, se caracterizan por mantener la información en los sistemas operacionales [133]. Uno de los elementos de este modelo es un esquema global de la información que unifique los datos de los diferentes sistemas fuente. Otro de los elementos es un motor de consultas que traduce las consultas del modelo de integración a los diferentes modelos de datos de los sistemas operacionales.

Al contrario que en el caso de los almacenes de datos, la información no se importa, sino que se transforma al hacer la petición al modelo de datos

origen. Para realizar esa transformación es necesario definir las correspondencias entre las fuentes y el esquema global. A la hora de definir estas correspondencias se han propuesto dos esquemas [134]:

- GAV (global-as-view): Este esquema necesita que la correspondencia se exprese en términos globales respecto al modelo central de explotación.
- LAV (local-as-view): Este enfoque implica que las correspondencias se definen de forma independiente en cada uno de los sistemas operacionales, y luego éstas se corresponden con vistas del esquema global.

Cada una de las propuestas tiene sus ventajas y desventajas. LAV favorece la extensibilidad del sistema, ya que en el caso de que hubiera un nuevo sistema operacional, sólo habría que definir las correspondencias locales de este sistema con el almacén global. Sin embargo, en un enfoque GAV añadir una nueva fuente podría requerir redefinir todo el modelo de datos central con el esfuerzo que ello conlleva. Centrándose en el procesamiento de consultas, el modelo GAV es mucho más eficiente, ya que estas consultas no tienen que traducirse a cada uno de los esquemas de cada sistema operacional, como sí habría que hacerlo con un enfoque LAV.

2.4.3 Integración basada en enlaces

Los sistemas de integración basados en enlaces intentan aprovechar el hecho de que la mayoría de recursos están disponibles en formato Web y que cualquier navegador puede acceder a ellos para resolver consultas [135]. Este tipo de integración permite generar grafos donde los datos de los sistemas operacionales se enlazan entre sí. Este tipo de integración ayuda a los usuarios a obtener la información vinculada entre los diferentes recursos de negocio, pero no permite explotarla de una forma integrada.

2.4.4 Combinación de información

Los sistemas que usan este modelo de integración combinan fuentes de información y funcionalidades para generar contenidos más completos o generar nuevos servicios por agregación [136]. Dos de los tipos de fuentes de datos más usados en este enfoque son los servicios REST [137] y la sindicación de contenidos [138], que pueden combinarse fácilmente para generar nuevos recursos de negocio.

Este enfoque no puede considerarse como una verdadera integración, ya que al igual que en el enfoque anterior, no permite explotar los datos de forma centralizada.

2.4.5 Integración mediante arquitecturas orientadas a servicios

Estos sistemas usan servicios Web como SOAP [139] o REST para extraer la información de los sistemas operacionales heterogéneos. Los sistemas que integran esta información la obtienen de esos servicios Web y la transforman en información homogénea, gracias a un modelo o esquema común de datos.

Los servicios Web, y en concreto las arquitecturas orientadas a servicios (SOA) son herramientas muy extendidas para integrar información entre los sistemas operacionales de las organizaciones [5].

2.4.6 Integración mediante arquitecturas dirigidas por modelos

Este enfoque puede verse como una extensión del modelo basado en mediadores. En este modelo, las correspondencias entre los sistemas operacionales y el sistema de IN se definen en base a un arquitectura dirigida por modelos [140], lo que implica que usan tipos de datos y un vocabulario común. Ésto evita que se tengan que definir adaptadores entre los diferentes sistemas. Este enfoque sólo es posible entre sistemas que ya están muy integrados entre sí, ya que sería inviable entre fuentes muy heterogéneas.

2.4.7 Integración de aplicaciones

Este tipo de sistemas se basa en el desarrollo de aplicaciones propias cuyo objetivo es integrar la información de los diferentes sistemas operacionales de la organización. Su reutilización es muy difícil debido a que están diseñados para el problema de integración que pretenden resolver.

2.4.8 Integración por flujos de trabajo

Este tipo de integración puede combinarse con el resto. Se basa en extraer la información en un momento concreto del flujo de trabajo de uno o varios sistemas operacionales. En este enfoque no se recupera la información de fuentes de datos o servicios Web, sino que se envía secuencialmente.

Ejemplos de este tipo de integración se pueden encontrar encontrar en integraciones basadas en mensajería. Por ejemplo, en el ámbito de la historia clínica electrónica, estándares de integración como HL7 envían una mensajería concreta con la información de un determinado evento que se ha producido en la aplicación [141].

Este tipo de modelo de integración tiene dos grandes ventajas:

- No requiere de un modelo común de integración.
- Al hacer una integración por fases, es posible saber en cada momento en qué estado se encuentra la integración.

Este modelo facilita la integración de aplicaciones y la explotación en tiempo real de los datos, aunque puede ser muy costoso de implementar dependiendo del número de procesos definidos en cada uno de los sistemas operacionales.

2.4.9 Integración semántica de la información

Las tecnologías de la Web Semántica ofrecen un entorno ideal para la integración y transformación de la información [142]. Entre estas tecnologías, las ontologías son un elemento fundamental ya que permite definir modelos formales para dar soporte a la integración de datos y eliminar la heterogeneidad de la información que se encuentra en los diferentes sistemas operacionales. Por un lado, las ontologías sirven como un vocabulario común clave para el proceso de estandarización. Por otro lado, se pueden usar como esquema global y para definir las correspondencias con los orígenes de datos o con otras ontologías que sirvan de esquemas locales. Se pueden encontrar tres enfoques entre las arquitecturas de integración que utilizan ontologías como elemento integrador [143]:

- Enfoque de ontología global. Usa una única ontología para modelar el dominio usando un único vocabulario común. En esta arquitectura se definen correspondencias directas entre los sistemas operacionales y la ontología. La principal desventaja de este enfoque está en su capacidad de tolerancia de cambios en los sistemas operacionales, haciendo que cada cambio afecte tanto a la ontología como a las correspondencias.
- Enfoque de ontología múltiple. Define una ontología para cada sistema operacional. Esto implica que se hagan correspondencias entre los sistemas operacionales y sus ontologías, y correspondencias entre las propias ontologías. Esta arquitectura es muy flexible a cambios, pero las correspondencias entre ontologías pueden llegar a ser muy complejas dependiendo de la heterogeneidad de la información.
- Enfoque híbrido. Define una ontología para cada uno de los sistemas operacionales basándose en una ontología común que proporciona un

vocabulario compartido. Esta ontología común facilita la generación de correspondencias entre las ontologías de los sistemas operacionales. La principal desventaja de este modelo es que las ontologías generadas, al estar definidas en base a la ontología común, no pueden ser reutilizadas fácilmente.

Los enfoques de ontología única son más propios para sistemas basados en almacén de datos o mediadores con enfoque GAV, mientras que el enfoque múltiple y el híbrido son más adecuados para mediadores basados en un enfoque LAV.

2.4.9.1 Ejemplos de integración semántica

En el ámbito de las tecnologías de la Web Semántica para integración de la información, uno de los usos más comunes es la aplicación de Linked Open Data (LOD). Bio2RDF [144] es un ejemplo de conjunto de datos biológico que ofrece una red federada de puntos de acceso a datos en RDF que han sido integrados de diversas fuentes.

También se han usado para la construcción de repositorios semánticos y sistemas basados en mediadores usando las ontologías como metadatos del modelo [145]. YeastHub [146] es un almacén de datos RDF que permite la integración de diferentes tipos de datos genómicos de la levadura.

El sistema OGO [147] es un ejemplo de repositorio integrado de información biomédica de genes y proteínas, relacionadas con su comportamiento biológico.

SWIT (Semantic Web Integration Tool) [113] es un motor de transformación semántica capaz de generar RDF y OWL desde bases de datos relacionales y repositorios XML. Una característica importante de SWIT es que previene la generación de información incosistente gracias a la integración con razonadores DL que evitan que los datos incorrectos se transformen al modelo semántico.

En Ontocop [41] se integra información de bases de datos relacionales que almacenan una discusión de ideas políticas con una ontología que modela el conocimiento en un dominio específico.

2.5 Procesamiento analítico en línea

El Procesamiento Analítico en Línea, más conocido como OLAP (*On-Line Analytical Processing*), es una de las herramientas más utilizadas en IN [5].

La función principal de OLAP es agilizar la consulta de grandes cantidades de datos, a partir de diferentes técnicas [148]:

- Uso de bases de datos relacionales en donde las tablas se diseñan principalmente para agilizar su consulta. En la mayoría de casos usan tablas desnormalizadas y detallan más la información para evitar datos agregados. Los sistemas que siguen este modelo reciben el nombre de ROLAP (Relational OLAP).
- Uso de bases de datos multidimensionales, cuyo funcionamiento consiste en almacenar resúmenes de los datos, con valores precalculados. Suelen usar algoritmos de compresión para disminuir su espacio en disco debido a toda la información precalculada que generan. Los sistemas que siguen este modelo reciben el nombre de MOLAP (Multidimensional OLAP). Estos sistemas también reciben el nombre de Cubos OLAP.
- HOLAP (Hybrid OLAP) consiste en almacenar parte de los datos en ROLAP y el resto en MOLAP, dependiendo de las características de cada uno de ellos.

Una de las principales desventajas de este tipo de tecnologías consiste en que necesitan alimentarse de los sistemas fuente periódicamente, y al estar optimizados para consultas de información, no lo están igualmente para las operaciones de manipulación, por lo que ese proceso de alimentación de los datos puede ser muy costoso. Para solucionar este problema se han empezado a comercializar soluciones en tiempo real llamadas RTOLAP (Real Time OLAP) [149].

2.5.1 OLAP y Web Semántica

El interés de combinar tecnologías de la Web Semántica con OLAP ha sido una línea de investigación en la que se han desarrollado varias propuestas. En [13] se propone el mapeo de diferentes orígenes de datos a ontologías RDF u OWL. En este caso, el usuario define el esquema MOLAP donde almacenará los datos y usa consultas SPARQL para rellenar el almacén de datos. En [15] se continúa este trabajo con la propuesta de la generación de esquemas OLAP a partir del modelo semántico automáticamente. En [150] se propone usar técnicas de la Web Semántica para recomendación de consultas basado en los logs de consultas que se realizan en un motor OLAP. En [17] se propone la combinación de OLAP para un almacén de datos estructurados con un almacén de datos no estructurados (documentos, imágenes, etc.) que estén anotados semánticamente. Por último, en [14] se propone el uso de consultas

SPARQL que permitan identificar las dimensiones y los valores calculados de un modelo MOLAP.

Como se ha comentado, estas soluciones están enfocadas a combinar OLAP con tecnologías semánticas o a usar éstas para facilitar el diseño y puesta en marcha de un almacén de datos OLAP. En [18] se define el vocabulario QB (The RDF Data Cube Vocabulary) propuesto por el W3C. Este estándar está pensado para disponer de un vocabulario común para que las organizaciones puedan compartir sus modelos de cubos OLAP a través de un modelo semántico común, es decir, no está pensado para el análisis de esos cubos. Para superar esta deficiencia, en [19] se propone un nuevo vocabulario QB4OLAP que permite realizar consultas concretas a los datos de un cubo OLAP usando SPARQL.

El denominador común de estas soluciones es que ya se dispone de una solución OLAP en la compañía, pero no existe una solución OLAP puramente semántica. Probablemente se deba a las propiedades especiales que definen la Web Semántica, en las que se pueden analizar los datos desde diferentes perspectivas sin necesidad de hacer una transformación del modelo de datos. Es decir, la Web Semántica va a permitir trabajar a varios niveles de detalle, generando indicadores o yendo al detalle de los datos para conocer cómo se calculan esos indicadores. Además, con el uso de razonadores o de lenguajes de definición de reglas, se puede generar nuevo conocimiento automáticamente. Es decir, se podrían generar valores calculados en tiempo real desde el propio modelo semántico.

2.6 Modelos de evaluación de "activos de conocimiento"

El impacto que ha tenido la IN en las organizaciones se suele medir en los resultados económicos de la empresa [7]. También es muy útil para poder medir la calidad de la información que gestionan las organizaciones [8]. Como ya se ha comentado, la IN ayuda a analizar lo que está pasando y a la toma de decisiones, además de convertir los datos en información, y la información en conocimiento. Este conocimiento está orientado al análisis de la actividad productiva y económica de la empresa, pero no tiene en cuenta otros activos de conocimiento como el capital humano, el clima laboral, la evaluación del desempeño, la cultura corporativa, los valores empresariales o las relaciones con otras empresas, entre otros. A este tipo de activos se les denomina "activos de conocimiento" [9]. Como ya se ha comentado, en el modelo de IN definido en [5] se establece la importancia de que los usuarios de la solución

de IN puedan evaluar las diferentes representaciones del conocimiento de la organización. También se ha comentado que esta característica esencial no suele estar incorporada en la mayoría de soluciones comerciales [26].

En esta sección se describen varias metodologías de evaluación del conocimiento que pueden integrarse en soluciones de IN. Concretamente se comentan las metodologías de evaluación "360 grados" y de "activos de conocimiento".

2.6.1 Evaluación "360 grados"

Este modelo de evaluación se ha usado para la evaluación de competencias en el ámbito de la gestión de recursos humanos [151]. Este método consiste en definir cuáles son las competencias que una persona debe desempeñar en su puesto de trabajo, establecer unas métricas de puntuación y, por último, realizar evaluaciones periódicas de dichas competencias. Recibe el nombre de "360 grados" porque las evaluaciones se hacen por diferentes personas que intervienen en ese puesto de trabajo como el jefe directo, compañeros de departamento, compañeros de otros departamentos que solicitan algún tipo de servicio y en algunos casos los propios clientes de la compañía. Además, cada persona puede autoevaluarse para tener una visión de cómo se ve él y cómo le valoran el resto de personas con las que colabora.

Este modelo de evaluación suele dividirse en tres fases [152]:

1. En una primera fase se realizan entrevistas y cuestionarios para definir cuáles serían las competencias necesarias que se deben cubrir en cada uno de los puestos de trabajo definidos en una compañía. En esta fase también se define cada uno de los comportamientos asociados a esa competencia. Los comportamientos son los elementos que posteriormente serán evaluados para saber el grado de desempeño de cada competencia. Por ejemplo, una competencia podría ser: "Trabajo en equipo", y para poder evaluarla se podrían definir los comportamientos: (1) "Fomenta el intercambio de información y experiencia en la búsqueda de resultados", (2) "Genera buen clima de trabajo y espíritu de cooperación", y (3) "Muestra interés y predisposición para trabajar en equipo".

2. En una segunda fase se elegirían las personas que van a evaluar a cada individuo y se realizaría la evaluación. El propio individuo objeto de la evaluación también se autoevalúa. Es importante que cada uno de los comportamientos y la evaluación que pueden recibir para cada uno de

ellos esté bien descrita, incluyendo unas pequeñas instrucciones para evitar el componente subjetivo.

3. Por último, se le dan los resultados a cada una de las personas, comentando sus puntos fuertes y recomendando dónde puede mejorar. Si se ha hecho más de una evaluación también se le pueden presentar datos evolutivos sobre su desempeño en la compañía.

En diversos artículos [152; 153] se demuestra el impacto positivo que tiene este tipo de evaluación en el desarrollo profesional de las personas. En [154] se comenta la importancia que tiene la evaluación del desempeño en el ámbito de la gestión del conocimiento. En este trabajo se usa esta metodología para identificar líderes expertos en algún tipo de conocimiento de la compañía.

2.6.2 Evaluación de "activos de conocimiento"

La evaluación de los activos de conocimiento suele ser un proceso complicado debido a su intangibilidad [9]. En esta sección se hace una revisión de las principales metodologías de evaluación de activos de conocimiento, comentando sus ventajas e inconvenientes.

2.6.2.1 Navegante de Skandia

Skandia fue la primera gran compañía que midió sus activos de conocimiento durante la decada de los noventa [155]. Su modelo de evaluación se conoce como Navegante (Navigator) y se enfoca en cinco áreas: finanzas, clientes, procesos, renovación y capital humano. Este modelo se basa en la integración del capital humano con el resto de factores que, sumados juntos, permiten identificar los activos de conocimiento y evaluarlos [156]. El capital humano se define como la combinación de conocimientos, competencias y habilidades que cada empleado de la compañía posee. El resto de áreas, que se denominará capital estructural, son el resto de áreas cuyos elementos tangibles son usados por el capital humano para mejorar su productividad. Ejemplos de esto podrían ser los edificios, los equipos, el software, las patentes o los procesos.

La principal ventaja de este modelo es que permite modelar los activos de conocimiento de una compañía en una representación basada en los activos tangibles y los intangibles, que básicamente son el capital humano. Este modelo ayuda a tener una vista rápida de la estructura de la compañía y es uno de los primeros modelos que permite representar los procesos y la renovación.

La principal desventaja de este modelo es que está preparado para obtener una fotografía de cómo está el conocimiento de la organización en un determinado momento, es decir, no permite realizar una evaluación de la evolución de los activos de conocimiento. Otro de los problemas de este modelo es que asume que los trabajadores usan correctamente los activos tangibles de la empresa (ordenadores, software, etc.), cuando eso no tiene por qué ser así [157].

2.6.2.2 Índice de Capital Intelectual (IC-Index)

Este modelo consiste en generar diferentes indicadores e intentar correlacionarlos para ver cómo los activos de conocimiento influyen en los cambios en el mercado de la compañía [158]. Estos indicadores y sus correlaciones ayudan a identificar las prioridades y las relaciones existentes entre los diferentes resultados de la evaluación.

Este modelo permite que los gestores puedan medir el impacto que ha tenido una determinada decisión estratégica en la empresa, gracias a la medición de esos indicadores. En [159] se concluye que la mejora en el IC-Index de una compañía con respecto a otra quiere decir con toda certeza que esa compañía lo está haciendo mejor. Sin embargo, puede ser un índice engañoso para compañías que empiezan desde cero, ya que su IC-Index mejorará considerablemente, lo que no implica que se estén tomando buenas decisiones.

Como principal desventaja, es un índice que depende mucho de la definición de los indicadores. Si esa definición no ha sido correcta, el índice devolverá valores que no se adecúan a la evaluación real de los activos de conocimiento de la compañía.

2.6.2.3 Technology Broker

Esta propuesta consiste en tres modelos de medida del conocimiento para ayudar a calcular el valor en dólares de los activos de conocimiento [160]. En este modelo se definen los activos de conocimiento como una combinación de: activos de mercado, activos centrados en el personal, activos de propiedad intelectual e infraestructuras. Los activos de mercado se definen como los clientes, los canales de distribución, contratos, acuerdos con licencias o franquicias, y líneas de producto. Los activos centrados en el personal almacenan indicadores sobre experiencia, creatividad y capacidad de resolución de problemas, liderazgo y evaluación de las competencias de cada empleado. Los activos de propiedad intelectual contienen los mecanimos legales para proteger el conocimiento generado por la compañía. Incluyen el *know-how*, *copyright*, patentes, derechos de autor y secretos comerciales. Por último,

las infraestructuras almacenan información sobre tecnologías, metodologías y procesos con los que la compañía funciona, incluyendo cultura corporativa, metodologías de evaluación de riesgos, bases de datos, estructuras financieras, información sobre el mercado de clientes y los planes de comunicación.

La forma de evaluar todos estos aspectos es a través de cuestionarios que rellenan diferentes agentes de la organización. Cada uno de estos cuestionarios está relacionado con uno o varios de los activos que han sido descritos previamente. Una vez que se han rellenado los cuestionarios, se ofrecen tres métodos para calcular el valor en dólares de cada activo:

- Evaluar el coste que tendría reemplazar un determinado activo.
- Comparar el valor de sus activos con el precio de mercado.
- Evaluar los ingresos con la capacidad de producción de cada activo.

Este método es un instrumentro de gran utilidad para identificar activos intangibles en las organizaciones [161]. Su principal desventaja es que los métodos de cálculo del valor del activo basado en dólares conllevan distintas dificultades. En algunos casos, es complicado encontrar precios de mercado para un determinado activo, sobre todo si éste es intangible. El enfoque basado en ingresos añade un componente de subjetividad, ya que en muchas compañías hay procesos transversales que suelen tener una implicación indirecta en la producción y que podrían salir con una evaluación baja cuando son imprescindibles para el funcionamiento de la compañía. Por último, el enfoque basado en el coste que tendría reemplazar un determinado activo también puede ser un valor difícil de calcular en aquellos casos en los que es un activo intangible.

2.6.2.4 Monitor de activos intangibles

En [162] se propone un entorno de trabajo conceptual basado en tres familias de activos intangibles: estructuras externas (marcas, clientes y relaciones con los proveedores), estructuras internas (la organización: gestión, estructura legal, manuales de trabajo, actitudes, software); y competencias individuales (formación y experiencia, básicamente).

Para evaluar estos activos se identifican tres indicadores (crecimiento y renovación, eficiencia y estabilidad) para cada uno de los tres activos intanbigles. Por ejemplo, para medir la competencia profesional se definirían los siguientes indicadores:

- Crecimiento y renovación: número de años en la profesión, nivel académico, formación, etc.

- Eficiencia: proporción de profesionales de la compañía, valor añadido por cada profesional, el efecto palanca de cada profesional, etc.
- Estabilidad: media de edad, antigüedad, salario, etc.

Como se puede observar, una característica que diferencia a este modelo es que no se basa en indicadores económicos, lo que algunos autores consideran una ventaja [161]. De este modelo se han desarrollado cursos de formación como TANGO [163] que tiene un simulador para aprender a definir y evaluar los indicadores que permiten monitorizar activos intangibles.

2.6.2.5 MVA y EVA

EVA son las siglas de Valor Económico Añadido (Economic Value Added) y es un sistema de medida que se basa en variables de carácter financiero y estratégico como planificación financiera, establecimiento de metas, presupuestos, comunicación con accionistas, compensación de incentivos, etc [159]. EVA ha sido descrito como un lenguaje y punto de referencia para discutir sobre la creación de valor económico añadido. En [164] se define EVA como el resultado neto de todas las actividades directivas.

EVA surge para ofrecer mejoras al método de cálculo de Valor de Mercado Añadido (MVA, Market Value Added). MVA representa la progresión de la inversión inicial en un negocio y su valor real en el presente. En [159] se considera que MVA puede representar una evaluación del valor neto presente de una compañía y su capacidad para invertir en nuevos proyectos. En este caso EVA es la métrica que ayuda a evaluar el impacto que han tenido dichos proyectos nuevos, ayudando a MVA a calcular la progresión hasta el valor neto actual de la compañía hoy en día.

EVA es una medida que permite evaluar el conocimiento desde un punto de vista financiero, identificando como activos intangibles a aquellos que generan algún tipo de valor añadido.

Su principal inconveniente es que está adaptado a empresas con gran capital en los que hay accionistas, lo que dificulta que se pueda usar a nivel global. Además, en algunos estudios (por ejemplo en [164]) se ha demostrado que no es un buen predictor para medir el valor de las acciones y su posible variación.

2.6.2.6 Patentes de Citas Ponderadas

Este modelo permite evaluar las patentes como activos de conocimiento en las compañías [155]. Se basa en un metodología que tiene seis procesos:

1. Definición del rol del conocimiento en el negocio.
2. Evaluación de las estrategias de la competencia y de los activos de conocimiento.
3. Clasificación de las compañías por sus activos de conocimiento.
4. Evaluación del valor de los activos para mantenerlos, desarrollarlos, venderlos o abandonarlos.
5. Investigación en áreas donde se puedan abrir puertas.
6. Integración del nuevo conocimiento identificado y repetición del proceso sucesivamente.

Este modelo ha sido empleado únicamente para la evaluación de las patentes. Uno de los indicadores que siempre se emplean es el número de veces que se usa dicha patente o se comercializa a un tercero. La ponderación de esas "citas" se realiza con la inversión realizada y con los ingresos recibidos, por ese motivo recibe este nombre.

Capítulo 3

Objetivos

3.1 Motivación

La IN requiere de la integración de diversas fuentes de datos que se suelen encontrar en las organizaciones y en sus entornos. En muchas ocasiones, el conocimiento generado por estas organizaciones se encuentra distribuido en sistemas heterogéneos o está en formatos en los que es prácticamente imposible su clasificación (imágenes, vídeos, etc.)

Para poder integrar la información de las organizaciones se han definido procesos denominados ETL (*Extract-Transform-Load*) que permiten extraer información de las bases de datos, estandarizarlos, eliminar lo que no es válido o no es necesario y cargarlos en otra base de datos o *data warehouse* donde poder analizarlos. Este proceso es útil cuando se analiza una organización concreta, pero se complica cuando se comparan organizaciones o parte de ellas entre sí.

Como se ha comentado, el impacto que ha tenido la IN en las organizaciones se suele medir en los resultados económicos de la empresa [7]. La IN ayuda a analizar lo que está pasando y a la toma de decisiones, además de convertir los datos en información, y la información en conocimiento. Este conocimiento está orientado al análisis de la actividad productiva y económica de la empresa, pero no tiene en cuenta otros ámbitos como el capital humano, el clima laboral, la evaluación del desempeño, la cultura corporativa, los valores empresariales o las relaciones con otras empresas, entre otros. Estos conceptos intangibles que aparecen en las empresas reciben el nombre de "activos de conocimiento" [9]. Como ya se ha comentado, en el modelo de IN definido en [5] se establece la importancia de que los usuarios de la solución de IN puedan evaluar las diferentes representaciones del conocimiento de la organización. También se ha comentado que esta característica esencial

no suele estar incorporada en la mayoría de soluciones comerciales [26].

Entre las propuestas para dar solución a estos problemas se encuentran las tecnologías semánticas, que han sido muy usadas tanto para representación como para explotación de la información de las compañías [12; 11]. Ontologías como SKOS [66] han sido ampliamente usadas en diversos trabajos sobre gestión del conocimiento [165]. En muchos ámbitos, el uso de OWL como lenguaje de representación del conocimiento ha demostrado su viabilidad para la explotación y generación de conocimiento nuevo, gracias al uso de motores de razonamiento [13]. Además, iniciativas como *Linked Open Data* proponen la publicación de datos en la Web en un formato procesable por otros ordenadores, y que fácilmente pueden enlazarse con otras fuentes de conocimiento para poder analizarlas y compararlas. El proyecto EU MUSING [115] ha sido uno de los pioneros en usar tecnologías de la Web Semántica para desarrollar una solución de IN, aunque sus principales resultados han estado enfocados a la integración de la información no estructurada usando tecnologías de procesamiento del lenguaje natural.

En esta tesis se proponen metodologías y herramientas para generar una solución de Inteligencia de Negocio basada en tecnologías semánticas que recoja las características esenciales definidas para este tipo de solución [5]. Además aportará grandes ventajas en el ámbito de la IN 2.0, ya que integra información de redes sociales. Por último, y como principales novedades, la solución propuesta será capaz de generar cuestionarios semánticos que aporten nuevo conocimiento e integrará una herramienta para medir y evaluar los activos de conocimiento. Entre los servicios de generación de conocimiento que ofrece esta solución se destacan: (1) evaluar los activos de conocimiento de cualquier organización, (2) realizar análisis personalizados de cada una de las actividades, (3) recomendar las medidas que se deben llevar a cabo y (4) analizar el impacto que esas medidas han tenido en la organización a nivel cuantitativo y cualitativo.

3.2 Objetivos

El objetivo principal de esta tesis es la investigación y el desarrollo de herramientas basadas en tecnologías semánticas que se integren para formar una solución de IN que aproveche las características de éstas para la explotación del conocimiento de una organización. Además, esa explotación podrá ser compartida y comparada con otras organizaciones. Para conseguir este objetivo se han definido las siguientes tareas:

- Diseño e implementación de un modelo de integración semántica de

cualquier tipo de contenido, estructurado y no estructurado.

- Diseño e implementación de un modelo de entrega de información basado en tecnologías semánticas.
- Diseño e implementación de un modelo de evaluación de activos de conocimiento.
- Diseño e implementación de un modelo de explotación y análisis de la información.
- Diseño e implementación de una plataforma que, apoyándose en el resto de servicios, permita integrar y explotar los datos de cualquier tipo de organización.
- Aplicación y validación de los resultados obtenidos en entornos simulados y reales de los dominios económico-financiero, evaluación del desempeño, clínico y de apoyo a la investigación biomédica.

3.3 Hipótesis

La hipótesis principal de esta tesis es que se pueden usar tecnologías de la Web Semántica para construir una solución integral de Inteligencia de Negocio que (1) integre la información de fuentes de datos heterogéneas, (2) permita su explotación y (3) use características de inferencia para generar nuevo conocimiento que facilite la toma de decisiones. Esta hipótesis se divide en las siguientes sub-hipótesis:

- **Las tecnologías de la Web Semántica permiten integrar la información proveniente de fuentes heterogéneas en un almacén de datos semántico. La integración permitirá asociar datos estructurados y no estructurados, y permitirá identificar criterios de evaluación de los activos de conocimiento implicados.** Comprobar esta hipótesis requiere contestar a las siguientes preguntas:
  1. ¿Cuáles son las metodologías existentes para integrar diferentes orígenes de datos en una fuente común?
  2. ¿Cómo se pueden integrar y explotar contenidos que no estén estructurados?
  3. ¿Cómo se pueden clasificar los activos de conocimiento para que puedan ser medidos y evaluados?

4. ¿Qué ventajas tienen las tecnologías de la Web Semántica para integrar diferentes tipos de datos que a su vez son heterogéneos entre sí?

- **Es posible usar un modelo semántico para añadir nueva información a recursos de negocio.** Comprobar esta hipótesis requiere contestar a las siguientes preguntas:

  1. ¿Es posible incorporar nuevo conocimiento, no recogido en los orígenes de datos, a las soluciones de Inteligencia de Negocio?
  2. ¿Es posible usar la Web Semántica para gestionar, integrar y explotar esa información en una solución de Inteligencia de Negocio?

- **Es posible desarrollar un buscador semántico guiado por la ontología que permita navegar por todo el conocimiento del almacén de datos.** Comprobar esta hipótesis requiere contestar a la siguiente pregunta:

  1. ¿Cómo las tecnologías semánticas pueden ayudar a desarrollar un buscador guiado por la/s ontología/s del dominio que se usan como modelo de integración de datos?

- **Las representaciones semánticas permiten generar modelos más fácilmente explotables por los usuarios de una solución de Inteligencia de Negocio.** Comprobar esta hipótesis requiere contestar a la siguiente pregunta:

  1. ¿Qué ofrece la Web Semántica para generar modelos de datos equivalentes a OLAP?

- **Los buscadores semánticos y las redes sociales de evaluación de activos empresariales evalúan el conocimiento generado por una solución de Inteligencia de Negocio.** Comprobar esta hipótesis requiere contestar a las siguientes preguntas:

  1. ¿Qué función puede tener la integración de herramientas sociales en una solución de Inteligencia de Negocio?
  2. ¿Cómo se puede usar la Web Semántica para implantar metodologías de evaluación del conocimiento de forma genérica?

- **Los servicios semánticos permiten al usuario generar informes, hacer búsquedas semánticas avanzadas, generación de cuadros de mando semánticos personalizados, análisis del impacto de**

la toma de decisiones, sistemas de planificación y recomendadores. Comprobar esta hipótesis requiere contestar a las siguientes preguntas:

1. ¿Cómo se puede usar el buscador semántico para generar servicios más avanzados?
2. ¿Cómo se pueden integrar algoritmos de recomendación y predictivos en entornos semánticos?
3. ¿Cómo se pueden integrar los activos de conocimiento con la evaluación de los mismos para analizar el impacto de una determinada decisión estratégica?

3.4 Metodología

La metodología que se ha seguido parte del estudio del estado del arte y sigue con la definición metodológica de la solución, su implementación y su validación en diferentes dominios.

- Estudio del estado del arte:
  - Inteligencia de negocio: estudio del estado actual de la Inteligencia de Negocio y de cuáles son sus principales retos, además de estudiar algunos ejemplos de aplicación en diferentes entornos. También se han descrito cuáles son los principales trabajos que se han desarrollado combinando Inteligencia de Negocio y Web Semántica.
  - Web Semántica: estudio de las diferentes tecnologías de la Web Semántica como RDF, OWL, y técnicas de ingeniería ontológica. También se hace un análisis de las propuestas existentes para obtención de representaciones semánticas a partir de diversos tipos de orígenes de datos no semánticos (XML, bases de datos relacionales, etc.).
  - Web 2.0: análisis del estado actual de la Web 2.0 y su importancia en el ámbito de las empresas. Estudio de las propuestas que han integrado las tecnologías de la Web 2.0 y la Web Semántica.
  - Integración de la información: estudio de las propuestas existentes para la integración de repositorios heterogéneos, así como la aplicación de web semántica a este ámbito.

  - Procesamiento analítico en línea (OLAP): análisis de los diferentes modelos OLAP existentes en el mercado. Estudio de las propuestas de integración de tecnologías semánticas en este tipo de herramientas.
  - Evaluación de activos de conocimiento: estudio de diferentes metodologías para evaluar activos de conocimiento, haciendo especial hincapié en el conocimiento que es intangible, y que en muchos casos no se tiene en cuenta en las soluciones de IN.

- Definición metodológica de la solución propuesta:
  - Estudio del modelo de integración de la información para la construcción de un almacén de datos semántico. Esta fase incluirá el uso de herramientas de transformación semántica para infomación estructurada, y diferentes métodos de integración de información no estructurada, diferenciando contenido en texto abierto y elementos multimedia.
  - Desarrollo de un proceso de integración de nueva información semántica. Este modelo permitirá incluir nueva información para completar el modelo.
  - Definición de una herramienta que permita que los usuarios puedan definir consultas SPARQL desde una interfaz gráfica, sin la necesidad de tener conocimientos informáticos avanzados.
  - Desarrollo de un modelo de evaluación de activos de conocimiento, que permita identificar y evaluar los activos intangibles de la organización.
  - Desarrollo de un proceso de definición de modelos semánticos que permita, a los usuarios de la plataforma, trabajar con la información que realmente interesa al análisis personalizado que ellos requieren.
  - Estudio y selección de diferentes soluciones que exploten la información generada por las fases anteriores. Esta nueva información podrá realimentar el almacén semántico.

- Implementación de las diferentes herramientas propuestas usando tecnologías semánticas. Las herramientas desarrolladas se integran entre sí para formar una solución de Inteligencia de Negocio en la que se usa la Web Semántica en todas las fases de su ciclo de vida.

- Validación de la solución propuesta en diversos entornos entre los que se encuentran dominios económico-financieros, de evaluación del desempeño, clínicos y para investigación biomédica. En cada uno de los casos se usan las herramientas y servicios de la solución que son necesarios.

Bloque II

Metodologías y herramientas para la Inteligencia de Negocio Semántica

Capítulo 4

Entorno para la Inteligencia de Negocio Semántica

En este capítulo se presentan una serie de métodos y herramientas modulares que pueden integrarse entre sí para dar lugar a soluciones completas de Inteligencia de Negocio Semántica. Siguiendo el enfoque descrito en [5], este modelo tendrá tres ámbitos de actuación: integración, entrega de información y análisis.

En las figuras 4.1 y 4.2 se pueden ver las diferencias entre el modelo tradicional y el modelo semántico propuesto. Como entradas al sistema se encuentran cualquier tipo de fuente de datos, ya sea estructurada, en lenguaje natural o contenido mutimedia (imágenes, vídeos, etc.). Además, la solución propuesta permite extender esos modelos de información usando "cuestionarios semánticos". Estos cuestionarios permiten que los usuarios puedan extender el modelo de datos de cualquier recurso de negocio sin la necesidad de tener conocimientos TIC avanzados. Esta información adicional se representa en forma de ontología y los datos se almacenan en RDF.

Para la fase de integración de la información existen tres posibles servicios:

1. El primero consiste en un método de transformación de la información. En la sección 2.2.5.3 se describieron las diferentes herramientas de integración, y dependiendo de cada problema se propone el uso de una o la combinación de varias de ellas.

2. El segundo servicio es un anotador semántico que permite clasificar el contenido no estructurado o el multimedia.

3. El tercero es un servicio especial de anotación, en el que cada recurso se clasifica en base a unos criterios de evaluación. Gracias a esta anotación,

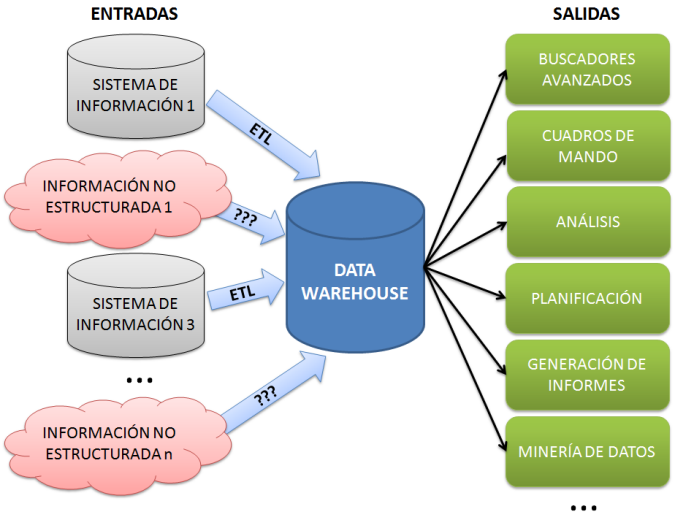


Figura 4.1: Esquema Inteligencia de Negocio 2.0

no sólo se clasifican semánticamente los recursos sino que se puede evaluarlos.

En el centro de la arquitectura aparece el almacén de datos semántico. Este almacén se compone de dos orígenes de datos: (1) un gestor documental donde se almacenan las ontologías en OWL con la representación formal del dominio de la organización, y (2) un repositorio RDF donde se almacenan las instancias y las propias ontologías OWL. En este caso se debe destacar que las ontologías del almacén de datos acompañan todo el ciclo de vida del software, es decir, esas ontologías guían tanto la integración como la explotación de la información, proporcionando un vocabulario común a los usuarios y servicios software.

Sobre el almacén de datos semántico se construye el primer servicio, llamado "Buscador guiado por la ontología" (Ontology-Driven Searcher u ODS), que permite definir consultas SPARQL a través de una interfaz gráfica que usa los conceptos, relaciones y propiedades de las ontologías OWL. En la arquitectura propuesta, la información semántica puede tener varias entradas, pero la entrega se hace a través de este servicio unitario.

Sobre ODS se construyen diferentes servicios de análisis, recomendación y planificación, además de ofrecer cuadros de mando o buscadores avanzados. Entre esos servicios de salida y el almacén semántico hay dos servicios intermedios que producen entradas y salidas de información:

- Perfiles semánticos. Se define "perfil semántico" como un resumen de

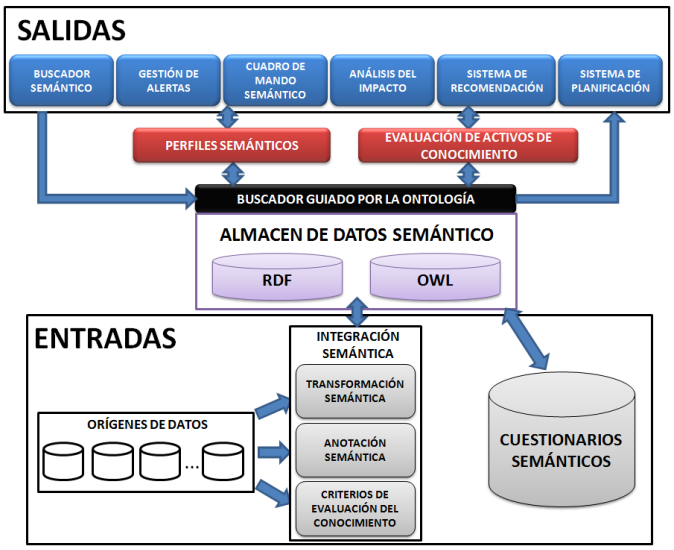


Figura 4.2: Esquema Inteligencia de Negocio Semántica

las características más significativas de un recurso, es decir, cuáles son las propiedades y relaciones con otros conceptos que se van a usar en un análisis concreto.

- Evaluación de activos conocimiento. Como se ha comentado previamente, las soluciones de IN están orientadas a generar conocimiento, pero no hay herramientas integradas que evalúen los activos intangibles de la organización y que, además, añadan los resultados de esa evaluación al almacén semántico. Este servicio es una herramienta Web 2.0 que hace uso de diferentes métodos para medir y evaluar el conocimiento de la organización. A partir de esa evaluación permite analizar el impacto de las decisiones estratégicas ejecutadas.

Usando este marco de trabajo para la Inteligencia de Negocio Semántica, se han desarrollado cinco plataformas completamente funcionales, que combinan diferentes métodos y herramientas:

1. Red social semántica.
2. Plataforma semántica para la planificación.
3. Plataforma semántica para Inteligencia de Negocio con bases de datos epidemiológicas.
4. Cuaderno de recogida de datos semánticos.

5. Plataforma para la clasificación semántica de recursos basada en contenidos multimedia.

A continuación se exponen detalladamente cada una de las metodologías y componentes software que integran esta solución.

4.1 Modelo de Integración de la Información

En esta sección se van a describir los diferentes mecanismos que permiten integrar la información proveniente de fuentes de datos heterogéneas en un modelo semántico común. Debido a los múltiples casos de uso que se pueden encontrar a la hora de abordar esta integración, en este trabajo se provee de diferentes herramientas que podrán usarse conjuntamente o por separado, dependiendo del problema que se pretenda resolver.

En la imagen 4.3 se puede ver la arquitectura de transformación de diferentes fuentes al modelo semántico. La transformación semántica es usada para transformar los datos provenientes de fuentes de datos estructuradas al modelo semántico. Se pueden usar algunas de las herramientas que se describieron en la sección 2.2.5. Para generar información semántica a partir de información en lenguaje natural o de contenidos multimedia se pueden usar diferentes mecanismos de anotación semántica. Por último, además de los elementos anteriores, también existe la posibilidad de anotar los recursos sobre un catálogo de criterios de evaluación (ver sección 4.1.4) que permita clasificar y evaluar los activos de conocimiento de la organización.

En la sección 2.4 se definen las metodologías existentes para la integración de la información. Como ya se ha comentado, la Web Semántica ha sido identificada como una tecnología ideal para la integración y transformación de la información [142]. Debido a la flexibidad del almacén de datos, la plataforma podrá usar los tres esquemas de integración semántica que ya han sido comentados, aunque con algunas limitaciones, como se verá a continuación.

4.1.1 Almacén de datos semántico

El objetivo principal del modelo de integración consiste en almacenar los datos provenientes de los orígenes de información en el almacen de datos semántico. Este almacén semántico se compone de dos elementos:

- **Repositorio OWL.** Permite almacenar en un gestor documental las ontologías que sirven como modelo del dominio. Se usa un gestor documental para disponer de un control de versiones, y así poder identificar

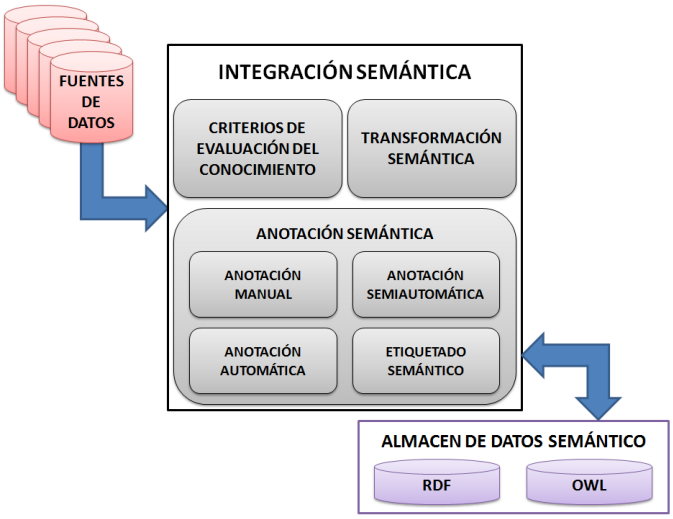


Figura 4.3: Modelo de Integración Semántica

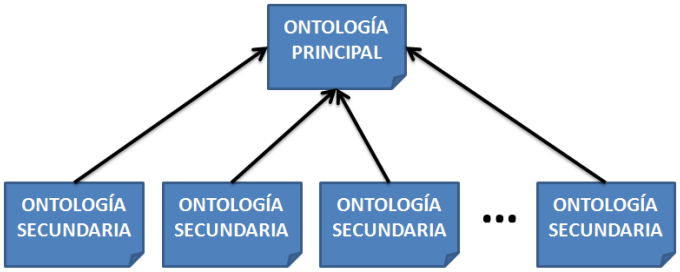


Figura 4.4: Modelo de Integración Semántica

en qué modelo concreto se modelaba la información semántica en un momento determinado del ciclo de vida de la aplicación. Todas las ontologías del repositorio se cargan en memoria para agilizar los procesos de consulta. Estas ontologías podrán generarse manualmente, venir de ontologías ya definidas o generarse usando herramientas que generan modelos ontológicos a partir de otros esquemas automáticamente.

- **Repositorio RDF.** Almacena los datos provenientes de los sistemas origen. La única característica que debe cumplir este almacén es que permita realizar consultas SPARQL de manipulación y recuperación de la información.

Una de las características principales del repositorio OWL es que necesita disponer de una ontología primaria que importa y relaciona el resto de

ontologías del dominio (véase figura 4.4). Esto quiere decir que cada vez que se quiera añadir una nueva ontología a la herramienta de IN, ésta deberá ser relacionada con, como mínimo, uno de los conceptos de esa ontología primaria. Otra característica importante de esta arquitectura es que las mismas ontologías que guían el proceso de transformación son usadas en el proceso de explotación, disponiendo de un vocabulario común en todo el ciclo de vida de la solución de IN.

El repositorio RDF puede ser cualquier plataforma que permita almacenar RDF. También debe permitir usar SPARQL como lenguaje de consulta y modificación de la información. La información deberá estar almacenada siguiendo los principios de *Linked Data* definidos en la sección 2.2.4. Esto otorga a las herramientas gran flexibilidad al poder adaptarse a múltiples entornos y modelos. Por ejemplo, se podrían usar mediadores semánticos que permitan ejecutar consultas SPARQL sobre modelos relacionales, y considerar a éstos como el repositorio RDF. En este caso no se podrían usar ni los cuestionarios semánticos ni el modelo de evaluación del conocimiento ya que generan nueva información semántica que debe ser almacenada en RDF. También se podrían usar algunas de las herramientas de transformación que permiten llevar datos de las fuentes originales a las fuentes destino.

Esto significa que el modelo propuesto separa el modelo semántico (ontologías) de los datos semánticos (repositorio RDF). Esta separación permite emplear un esquema más eficiente de análisis y explotación de la información, aunque limita el uso de razonadores. Para solventar este problema, la herramienta dispone de un exportador de toda o parte de la información del repositorio RDF a OWL.

4.1.1.1 Buscador guiado por la ontología

El almacén de datos semántico tiene un módulo que permite que los servicios que se construyen sobre él puedan consultar la información. Este módulo se denomina "Buscador guiado por la ontología" (ODS - *Ontology-driven searcher*). Esta herramienta ofrece una interfaz gráfica para construir consultas SPARQL guiada por la ontología. Además ofrece un mecanismo que permite persistir las consultas generadas para su parametrización y su posterior reutilización.

La construcción de las consultas parte de un concepto principal de la ontología. Por ejemplo, si se deseara buscar facturas ODS partiría de un concepto inicial "Factura". En función de ese concepto ODS permite que el usuario vaya declarando cláusulas con los diferentes filtros que desea aplicar. Esas cláusulas pueden volver a anidarse entre sí cuando el usuario usa alguna *owl:ObjectProperty* mientras que está filtrando. Por ejemplo, si se desea bus-

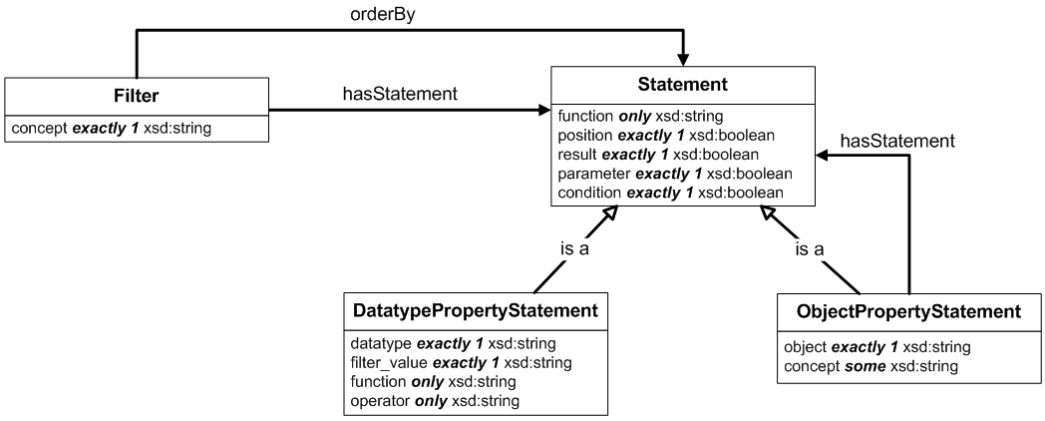


Figura 4.5: Modelo semántico para serialización en ODS

car facturas cuyo proveedor sea español, se tendría una cláusula anidada que indicase que el valor de la propiedad nación del concepto "Proveedor" fuese España. Cada una de esas cláusulas puede ser marcada como parámetro de agrupación, es decir, puede ser usada para agregar los resultados de la búsqueda. Siguiendo con el ejemplo anterior, en vez de buscar por proveedores españoles se podría agrupar por esa cláusula y los resultados serían el número de facturas por nacionalidad del proveedor. Cada una de las cláusulas puede marcarse como una variable resultado, esto es, el usuario desea que la variable destino de la cláusula le sea devuelta como resultado de la consulta.

Para la serialización de estas consultas se ha definido un modelo en OWL que permite clasificarlas para posteriormente ser reutilizadas. En la figura 4.5 se puede ver un esquema de la ontología[1]. A continuación se van a describir cada uno de sus conceptos:

- *Filter*. Este concepto modela la consulta en sí. Tiene un concepto principal desde el que se parte para generar las cláusulas de la consulta. Además almacena un listado de cláusulas y los parámetros por los que se quieren ordenar los resultados.

- *Statement*. Las cláusulas son representadas en este concepto. Hay dos subtipos de cláusulas, que son disjuntas entre sí, para filtrar por el tipo de relación: (1) *owl:objectProperty* y (2) *owl:datatypeProperty*. En este concepto genérico se representa si se quiere marcar la variable de la cláusula como que es un resultado de la consulta, si se quiere aplicar alguna función agregada sobre la variable devuelta(*count*, *avg*, etc.) o

[1]http://sele.inf.um.es/ontologies/SPARQL_Serialization.owl

si se va a indicar que es un parámetro que podrá ser reusado posteriormente. En este concepto también se puede indicar que los resultados de la consulta estarán agregados por la variable usando la propiedad "*group_by*".

- *ObjectPropertyStatement*. Este concepto representa una cláusula en la que se usa una *owl:objectProperty* para filtrar o para añadir una relación con otro concepto. Esta cláusula permite que se aniden los filtros a partir de las relaciones con otros conceptos.

- *DatatypePropertyStatement*. Este concepto representa una cláusula en la que se usa una *owl:datatypeProperty* para filtrar. En este caso no se pueden anidar más cláusulas. El usuario puede indicar el valor del filtro y el operador a aplicar (es igual que, contiene, menor que, mayor que, etc.). También puede aplicar funciones sobre la variable a la hora de filtrar. Por ejemplo, si se tiene un campo de fecha en el modelo semántico y se quiere filtrar por el campo año de esa fecha se puede usar la función "*year*".

Esta herramienta permite que cualquier persona pueda construir sus consultas SPARQL sin necesidad de tener conocimientos informáticos avanzados, ayudado por las relaciones y propiedades de cada uno de los conceptos ontológicos del modelo semántico.

4.1.2 Metodología de transformación

En esta sección se define una metodología genérica de transformación semántica que hará uso de las herramientas que fueron descritas en la sección 2.2.5.3, dependiendo del problema concreto a resolver, aprovechando de forma óptima las fortalezas de cada una de ellas e intentando limitar sus debilidades.

Como se comenta en la sección 2.2.5.4, cada una de las herramientas descritas tiene sus ventajas e inconvenientes. Desde un punto de vista teórico, la mejor opción es usar una herramienta que haga una transformación semántica de la información pero, como ya se ha comentado, en algunos dominios muy extensos la definición de correspondencias puede suponer una tarea muy compleja. Por ese motivo no se limita el uso de cualquiera de las herramientas citadas. Por ejemplo, se podría usar una transformación sintáctica de algunos datos de las fuentes originales que en su modelo son muy semejantes al modelo semántico definido, situación común en muchos casos. Por ejemplo, una tabla de un modelo relacional que almacene personas del sistema será claramente mapeada a la clase Persona del modelo semántico.

La metodología de transformación propuesta incluye las siguientes fases:

1. Búsqueda y desarrollo de las ontologías que modelan el dominio de la organización. En esta fase es posible usar algunas de las herramientas de transformación que generan automáticamente un modelo semántico a partir de los orígenes de datos de la organización.

2. Modelado de la ontología principal sobre la que se representa el sistema de IN semántico.

3. Elección de la herramienta o las herramientas de transformación semántica, optimizando en la medida de lo posible que la transformación que se realice sea guiada por las ontologías que modelan el conocimiento.

4. Realizar la definición de correspondencias entre los diversos orígenes de datos y el modelo final. Esta fase no será necesaria si se usan los modelos de transformación sintáctica.

5. Extracción de la información. En este caso, se podría usar alguna de las herramientas que permiten realizar consultas SPARQL directamente sobre los orígenes de datos como D2RQ. Este tipo de herramientas permitiría hacer una transformación entre datos que ya están originalmente en RDF.

6. Generación de la información en OWL o en RDF. Se intentará priorizar la generación en OWL para poder aplicar técnicas de razonamiento a los datos que garanticen su consistencia.

7. Carga de los datos al repositorio RDF usando SPARQL.

4.1.3 Mecanismos de anotación semántica

Además de la información almacenada en bases de datos, en la mayoría de organizaciones existen otro tipo de sistemas de información a considerar, como gestores de contenidos, gestores documentales, información en redes sociales, blogs, wikis, etc. Por otro lado, en muchas bases de datos existen campos de texto libre que difícilmente pueden ser mapeados usando herramientas de transformación semántica. Para integrar este tipo de datos se puede hacer uso de cuatro mecanismos de clasificación semántica de texto abierto, ya sean documentos, comentarios en una red social, contenido multimedia o simplemente un campo descriptivo en una base de datos:

1. Anotación automática. En este mecanismo se usan metodologías de procesamiento del lenguaje natural para extraer entidades de la ontología a partir de texto libre.

2. Anotación manual. En este proceso será el usuario el que anote manualmente las entidades de negocio sobre el modelo ontológico.

3. Anotación semiautomática. Este mecanismo es una combinación de la anotación automática y la manual. En este modelo, el proceso de anotación automática realiza una recomendación de cuáles deberían ser las anotaciones del campo de texto libre. Sobre esas recomendaciones el usuario decide cuáles se van a aplicar y cuáles no.

4. Etiquetado semántico. Este modelo consiste en definir consultas sobre el almacén semántico cuyo resultado generará tantas anotaciones como recursos sean devueltos.

Todas las anotaciones son almacenadas en RDF.

4.1.3.1 Anotación automática

Este modelo emplea técnicas de procesamiento del lenguaje natural para extraer entidades de conocimiento desde el texto en formato abierto. El enfoque propuesto permite que el almacén de datos pueda integrar los contenidos sin necesidad de ser supervisados por una persona. El método consiste en procesar el contenido y la información semántica generada se almacena automáticamente en el repositorio. Este modelo está inspirado en la metodología propuesta en [166] y usa el marco de trabajo GATE [167]. El aspecto negativo de esta metodología es que la tasa de éxito (por ejemplo, el número de entidades de conocimiento extraídas del texto comparadas con el número real de entidades de conocimiento disponibles) estará por debajo del cien por cien, es decir, es probable que no anote todo el conocimiento del contenido. Además, también existe la posibilidad de que genere falsos positivos. Por lo tanto, este modelo no aprovecha toda la información semántica que puede ser gestionada.

4.1.3.2 Anotación manual

Este modelo de anotación consiste en que el usuario añada información semántica a cada uno de los contenidos. Este enfoque también puede ser usado para etiquetar imágenes u otro tipo de contenido multimedia que no sea texto.

El proceso de anotación manual sería el siguiente:

1. El usuario encuentra que el contenido se refiere a una o varias entidades del modelo semántico.

2. El usuario busca en la ontología esa o esas entidades.

3. La aplicación anota el contenido y lo almacena en el repositorio RDF.

Para facilitar el proceso de anotación los usuarios pueden usar ODS para encontrar los conceptos, instancias, relaciones y propiedades de las ontologías.

4.1.3.3 Anotación semiautomática

Este modelo de anotación se basa en los resultados del mecanismo de anotación automática. En este caso, los resultados generados por parte del proceso automático no se guardan directamente en el modelo semántico sino que el usuario debe validar que la anotación sea correcta. Con este enfoque, el usuario puede confirmar la anotación, eliminarla o modificarla para adecuarla al contenido que se pretende representar.

Con esta herramienta, el proceso de anotación podría ser completado con la anotación manual en el caso de que el usuario lo considerase oportuno.

4.1.3.4 Etiquetado semántico

Se define el proceso de etiquetado semántico como la construcción de una o más consultas semánticas que devuelvan instancias que anoten a un recurso. El resultado de este proceso es la creación de un conjunto de tripletas $\langle x, y, \langle z \rangle \rangle$, donde x es el recurso que se quiere anotar, y es la relación concreta que se va a usar para anotar y z es el resultado de la consulta.

Este modelo es útil para etiquetar semánticamente un contenido, como las entradas de un blog, imágenes, etc. En este caso se utiliza ODS como herramienta que permite generar consultas semánticas, cuyo resultado permitirá etiquetar cualquier contenido. ODS hace uso de las ontologías del dominio para sugerir a los usuarios la siguiente palabra a introducir dentro de una oración. Por ejemplo, si se ha escrito un comentario en una red social financiera sobre empresas de telecomunicaciones españolas, se podría emplear ODS para expresar que el comentario se anota con una consulta que devuelve todas aquellas empresas de telecomunicaciones con sede en España, creando tantas etiquetas semánticas como resultados devuelva la consulta, en un único paso.

El texto introducido en ODS está totalmente alineado con las ontologías del dominio, lo que supone algunas limitaciones:

1. La herramienta usa una entrada cerrada, basada en los conceptos y relaciones de la ontología. El usuario no puede etiquetar los contenidos en texto libre que sería la forma más natural.

2. Las ontologías usadas para definir el etiquetado pueden carecer de conceptos y relaciones que sean relevantes para el usuario.

Para solucionar el primer inconveniente se podrían usar alguno de los modelos de anotación que se han comentado anteriormente. En este caso, en vez de etiquetar el contenido con los resultados de una consulta SPARQL, se tendrían que etiquetar uno a uno usando los conceptos, relaciones, propiedades e individuos de la ontología. Por ejemplo, en el caso que se ha comentado anteriormente, el usuario escribiría *?Tele?* y el anotador le recomendaría todos aquellos términos de la ontología que contengan esa palabra, y así sucesivamente.

Para paliar el segundo inconveniente se ha desarrollado un editor de ontologías que permite la incorporación de nuevos conceptos, relaciones e instancias. Esta modificación de la ontología puede requerir una validación antes de que esos elementos pasen a estar disponibles para otros usuarios. Con este enfoque, cuando los usuarios no encuentran en ODS los términos que ellos esperan para hacer la consulta, pueden entrar al editor y modificar o añadir información semántica.

Una característica a destacar de esta herramienta es que cada etiquetado semántico puede reutilizarse. Cada vez que un usuario genera una consulta SPARQL usando ODS ésta puede ser almacenada como un nuevo concepto de la ontología y crear los resultados de la consulta como instancias de éste. Por ejemplo, el usuario anterior, una vez que ha definido con ODS la clase de empresas de telecomunicaciones con sede en España puede reutilizar esa etiqueta simplemente indicando el concepto, sin necesidad de volver a construir toda la consulta y sus diferentes filtros.

4.1.4 Clasificación de la información en criterios de evaluación

Conceptualmente, se define "criterio de evaluación" como un punto de referencia con el que medir la idoneidad o el rendimiento de un determinado recurso (persona, actividad, material, etc.) en un determinado momento de su vinculación a una organización. Semánticamente, un criterio de evaluación es una instancia de un concepto ontológico que representa la unidad mínima de conocimiento en una organización. En la figura 4.6 se pueden ver las relaciones semánticas de este concepto. Además, se puede usar *owl:subClassOf* para determinar que un criterio es una especialización de otro. Por ejemplo, se podría definir que "hoja de cálculo" es una subclase de "ofimática".

Este método es útil para clasificar el conocimiento de una organización en un modelo que posteriormente permita evaluarlo. Con esta metodología,

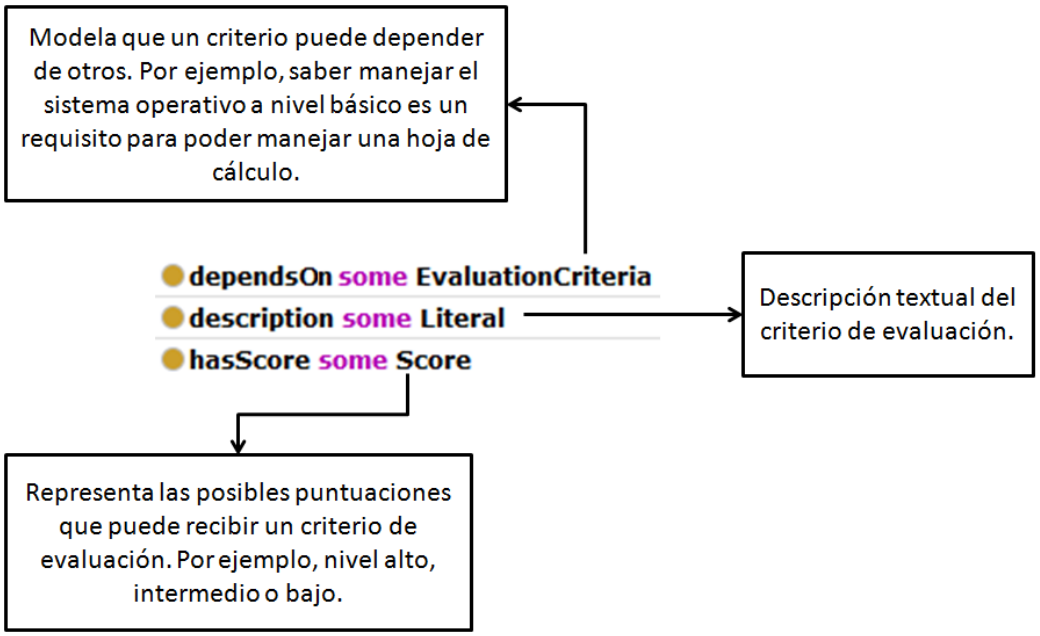


Figura 4.6: Relaciones semánticas de un criterio de evaluación

en vez de anotar sobre cualquier entidad de la ontología, se anota sobre un modelo semántico que representa una serie de criterios que pueden ser evaluables.

Esta metodología permite la generación de mapas de criterios de evaluación de las entidades de negocio de cualquier organización. Por ejemplo, se podrían definir mapas de competencias para los puestos de trabajo de una empresa, definir las características necesarias que debe cumplir un espacio de trabajo para la prevención de riesgos laborales, cuáles son los estándares que marcan la calidad de un proceso clínico, etc.

La metodología presentada permite definir el catálogo de criterios de evaluación de una o varias organizaciones, que se convertirá en el modelo de conocimiento que se usará para evaluar y planificar qué se necesita para mejorar. Sin el concepto de criterio de evaluación, la granularidad del modelo no sería adecuada, ya que se trabajaría directamente con los recursos, cuando lo normal es que un mismo recurso pueda ser evaluado con diferentes criterios dependiendo de las circunstancias. Este enfoque ofrece una gran flexibilidad a la hora de definir el glosario de criterios de evaluación. Permite que éste sea plano, jerárquico o incluso en forma de mapa, existiendo diversos criterios que pueden estar relacionados entre sí sin la necesidad de que esa relación sea jerárquica. A continuación se muestran los diferentes mecanismos de cla-

sificación del glosario de criterios de evaluación:

- **Por tipología.** Los criterios de evaluación pueden clasificarse en base a unos tipos (conocimientos, habilidades, características, etc.). Estos tipos pueden a su vez estar jerarquizados teniendo la posibilidad de crear subtipos que permitan afinar aún más la clasificación de los diversos criterios. La creación de estos subtipos también permitiría la definición de niveles de conocimiento, es decir, un recurso puede tener un determinado nivel sobre un criterio de evaluación. Por ejemplo, alto, medio o bajo.

- **Clasificación jerárquica.** Los criterios de evaluación pueden clasificarse jerárquicamente entre sí. Por ejemplo, el criterio Ofimática podría tener subcriterios como editor de textos, hojas de cálculo, etc. Esta jerarquización también ahorra trabajo a la hora de gestionar los mapas de evaluación, ya que al vincular un criterio padre, todos sus hijos también son vinculados. Al igual que en la clasificación anterior, este tipo de clasificación también permite definir niveles de conocimiento.

- **Etiquetado.** Todos los criterios de evaluación pueden etiquetarse semánticamente usando las herramientas que se han citado anteriormente. Gracias a este etiquetado, la herramienta puede encontrar relaciones entre criterios que no han sido definidas implícitamente.

- **Relaciones.** Los criterios de evaluación pueden estar relacionados entre sí mediante relaciones no jerárquicas. Los gestores de la plataforma pueden definir tipos de relación a través de las cuáles se conformen vínculos entre criterios que son equivalentes, o que implican alguna relación de transitividad, etc.

Como se puede observar, el modelo es totalmente flexible y permite que cada organización defina su glosario de criterios tan complejos como considere. Lógicamente, un glosario de criterios complejo, aunque sea más difícil de gestionar, facilitará la obtención de mejores resultados en el proceso de evaluación y planificación.

El modelo de anotación semántica de un criterio de evaluación sobre un recurso permite saber que éste cumple una serie de funciones y qué puntuación hay que obtener en ese criterio para que la valoración del desempeño de las funciones sea positiva. Semánticamente, la anotación de un recurso se define en base a instancias de la clase "*DesirableLevel*" cuyas relaciones se pueden observar en la figura 4.7.

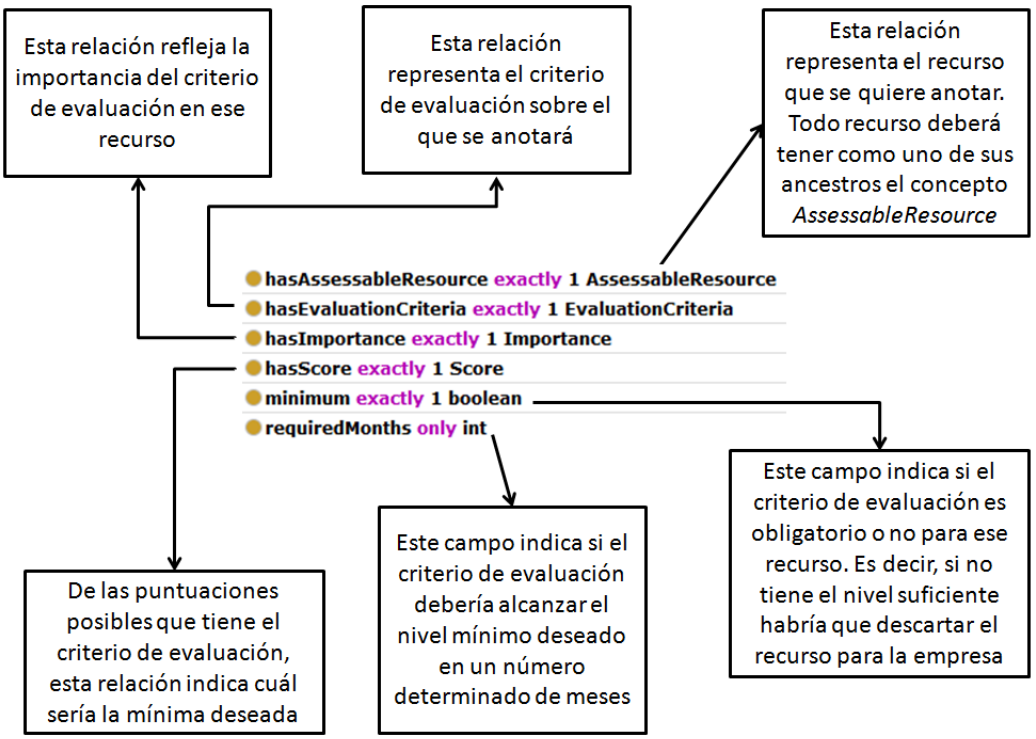


Figura 4.7: Modelo de anotación en criterios de evaluación

4.2 Generación de cuestionarios semánticos

En muchas ocasiones las organizaciones necesitan disponer de la suficiente flexibilidad en sus sistemas de información para poder añadir nueva información a los mismos sin necesidad de que esos cambios requieran de desarrolladores de software que los lleven a cabo. Muchos de los ERP del mercado cuentan con soluciones para solventar estos problemas [168]. El problema de este tipo de herramientas es que, aunque permiten la introducción de esos datos personalizados, rara vez permiten su explotación o su interoperabilidad con otras herramientas, ya que la carencia de significado de esa información impide su procesamiento automatizado.

Para solverntar este problema, se ha definido una metodología que permite extender la información que se almacena de un recurso directamente sobre motores de persistencia semánticos. Para ello, se ha definido un modelo ontológico en el que se representan los conceptos básicos (será extendido para cada recurso concreto):

- *Report.* Esta clase representa el conjunto de información que se va a

añadir a un recurso. Se compone de propiedades y relaciones definidas por los gestores de datos para cada caso de uso concreto.

- *Stage.* Esta clase representa la etapa o fase en la que se encuentra un recurso. En cada uno de estos estadíos se podrán almacenar uno o varios *Report*.

Se define "cuestionario semántico" como un concepto representado por la tripleta $\langle\langle C, S\rangle, op, \langle R, M\rangle\rangle$, donde C es una instancia del concepto *Report* que agrupa una serie de propiedades y relaciones S que añadirán información a través de una *owl:objectProperty op* a un recurso concreto R en un momento concreto M de su ciclo de vida (instancia del concepto *Stage*).

En consecuencia, este modelo permite la definición de cuestionarios que completen la información de cualquier entidad de negocio de la organización y que directamente la incorporen al modelo semántico, pudiendo ser explotada como el resto de la información. Esta metodología consta de cuatro elementos principales:

- **Cuestionarios.** Los cuestionarios sirven para definir las entidades que van a completar la información de un determinado objeto de negocio.
- **Máquinas de procesos.** Las máquinas de procesos ayudar a saber cuándo y cómo se deben rellenar esos cuestionarios, es decir, van a indicar el ciclo de vida que sigue esa información a la hora de alimentar el sistema.
- **Traducción semántica.** La combinación de entidades de negocio, máquinas de procesos vinculadas a éstas y cuestionarios que completan su información generará una ontología que servirá como esquema para manipular y explotar la información recogida usando estos elementos.
- **Motor de ejecución semántica.** Este módulo permite generar formularios Web que crean y actualizan la información de cada uno de los cuestionarios semánticos, siguiendo las reglas definidas en los campos del cuestionario y las máquinas de procesos.

En la figura 4.8 se puede ver la arquitectura de esta herramienta. A continuación se describen cada uno de esos elementos.

4.2.1 Generación de cuestionarios

La generación de cuestionarios consiste en definir los campos necesarios para completar la información concreta que se desea incorporar. Los tipos de campo que se pueden usar en este esquema son: números (enteros y reales), fecha

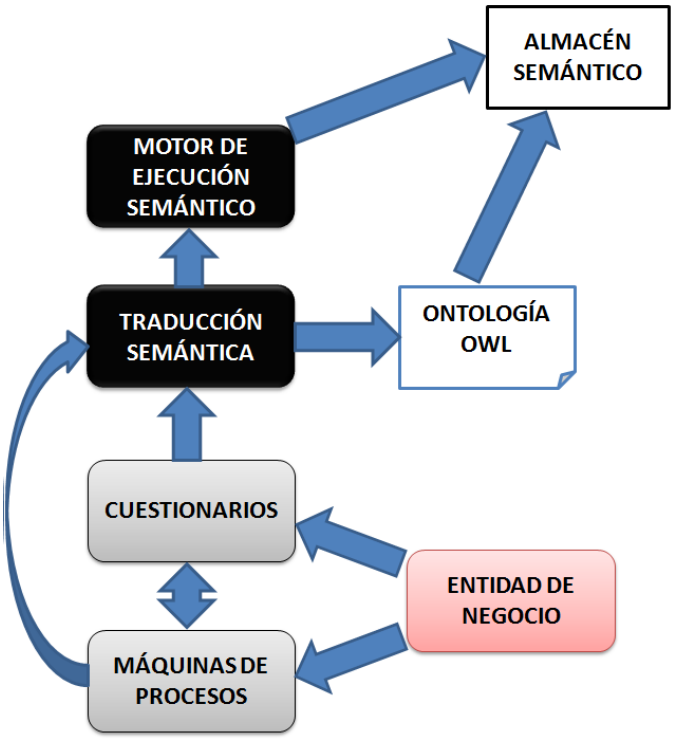


Figura 4.8: Arquitectura de los cuestionarios semánticos

y hora, texto (enriquecido o que cumpla una expresión regular), booleano y seleccionar entre una o múltiples opciones.

De estos campos, se puede destacar el seleccionable, que permite que el usuario defina nuevas taxonomías para rellenar el valor del campo o que se aproveche de conceptos ontológicos ya definidos previamente. También se puede relacionar a conceptos de ontologías del almacén de datos semántico. Cuando se definen campos seleccionables a partir de elementos taxonómicos, éstos tienen asociados valores numéricos que permitirán la generación de modelos cuantificables de los cuestionarios. Por ejemplo, si se quisiese modelar un examen tipo test, se puede indicar que todas las respuestas no válidas tienen un valor de cero y que las válidas tienen un valor de uno.

Otro aspecto a destacar es que los campos pueden ser reutilizados en múltiples cuestionarios, lo que permitirá estandarizar esta información para poder compartirla y compararla.

Cuando se asocian campos a un cuestionario existe la posibilidad de generar una serie de reglas (que tendrán su correspondencia en el modelo semántico generado):

- Reglas de cardinalidad. Indican la cardinalidad mínima y máxima del elemento a almacenar en ese campo. La cardinalidad puede ser un valor fijo o puede venir definido por los valores que se hayan definido en otros campos. Por ejemplo, si en un cuestionario definimos un campo numérico denominado número de hijos, posteriormente se puede definir el campo nombre del hijo tantas veces como el valor que se haya rellenado para ese campo. Otro ejemplo podría ser un campo con la pregunta de si usted fuma o no, y que en el caso que responda que fuma sea obligatorio rellenar el número de cigarrillos diarios que consume.

- Reglas de rango. Indican cuál será el rango de valores que puede tener una determinada variable.

- Reglas de formato. Son expresiones regulares a satisfacer por el valor a almacenar del campo. Estas reglas pueden ser de utilidad para guardar una dirección de correo electrónico, un número de teléfono, etc.

Una vez que se han definido los cuestionarios y los campos asociados, existe la posibilidad de definir diferentes reglas para generar campos automáticamente. Por ejemplo, si en un cuestionario se define un campo de estatura y otro de peso, se podría crear un nuevo campo denominado "índice de masa corporal" que se calcula a partir de los elementos anteriores. Es decir, se pueden generar campos ocultos a la hora de introducir los datos que, sin embargo, luego se podrán explotar al calcularse automáticamente a partir de los valores de otros campos.

Cada cuestionario definido se asocia a una o varias entidades de negocio a las que completará la información. Por ejemplo, en un hospital, se podría definir un cuestionario de calidad de vida para la entidad de negocio "estadío de la enfermedad" de un paciente.

4.2.2 Máquinas de procesos

Las máquinas de procesos se definen siguiendo el estándar de máquinas de estado [169], es decir, presentan unos estados iniciales, unos estados finales y unos procesos intermedios por donde transita la información. Estas máquinas de procesos se relacionan con la entidad de negocio que se quiere completar, es decir, van ligados a cada una de las instancias de esas entidades. Una misma entidad puede tener más de una máquina de procesos o tener varias entrelazadas entre sí.

Para configurar cada uno de los procesos se especifica la siguiente información:

- Transiciones entre procesos. Las transiciones entre procesos son los caminos que la información puede seguir a partir de un proceso origen. También se usan estas transiciones para hacer mapas de procesamiento, es decir, configurar cada uno de los procedimientos por lo que tiene que pasar una determinada entidad de negocio. Por ejemplo, para la recogida de muestras clínicas se pueden definir mapas de procesamiento que representen cómo se va a procesar la muestra recogida para generar los diferentes productos que serán almacenados.

- Cada proceso tiene una o varias personas responsables de realizar las tareas que correspondan en esa transición de la información.

- Para cada proceso se pueden definir diferentes alertas a múltiples niveles:

  - Alertas de tiempo. La información lleva demasiado tiempo detenida en el mismo proceso.
  - Alertas de completitud de la información. La información incluida en cada uno de los procesos no cumple con los requisitos mínimos especificados.

- En cada uno de los procesos se identifican cuáles van a ser los cuestionarios que se han de rellenar. Por ejemplo, en un programa de cribado clínico, en una fase inicial del reclutamiento se podría usar un cuestionario para almacenar los datos de la analítica sanguínea y, a partir de las variables recogidas, recomendar cuáles son las posibles transiciones y consecuentemente cuáles serían los datos que se deben seguir completando.

En la figura 4.9 se puede ver un ejemplo ilustrativo de una máquina de procesos. En la figura se puede ver que hay un proceso inicial, varios finales y una serie de procesos intermedios que están unidos por las transiciones. Como ya se ha comentado, en cada proceso puede haber uno o varios responsables, pueden existir alertas y los cuestionarios que se deben rellenar en esa fase.

4.2.3 Traducción semántica

El proceso de generación de estos elementos que pueden añadir información y lógica de negocio a cualquier entidad o proceso del sistema de información finaliza con la generación automática de una ontología en formato OWL que representa todo el dominio.

El proceso de traducción consta de los siguientes pasos (véase figura 4.10):

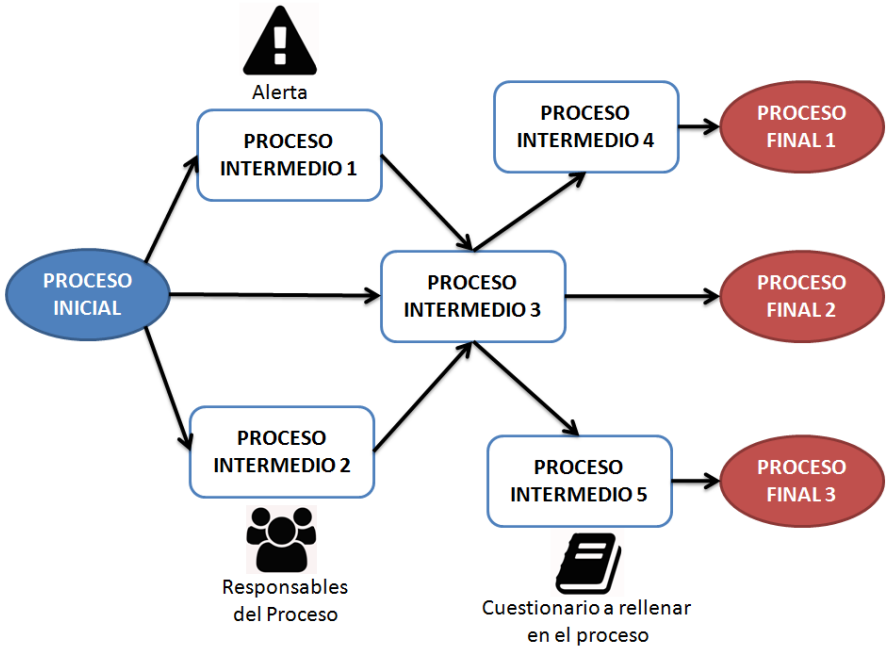


Figura 4.9: Ejemplo de máquina de procesos

1. Generación de los conceptos que representan las instancias de negocio que se van a usar. Por ejemplo, si se va a ampliar la información almacenada de una muestra, en el modelo ontológico tendrá representado el concepto OWL "Muestra".

2. Generación de los conceptos que representan los procesos por los que puede transitar la instancia de negocio. Además de generar los procesos, genera las relaciones y usa un concepto intermedio que representa cuándo un objeto de negocio se encuentra en un proceso determinado. Otro de los aspectos importantes que genera son las reglas de transición entre procesos.

3. Por último, genera tantos conceptos como cuestionarios se hayan definido para cada uno de los procesos. El objeto de negocio, el proceso y la información se vinculan con el concepto intermedio que se ha comentado anteriormente. Cuando el cuestionario tiene campos a seleccionar entre múltiples opciones también generará conceptos taxonómicos en OWL con los que se relacionará el cuestionario. La traducción también mantendrá las reglas de integridad de la información, como la cardinalidad y el rango.

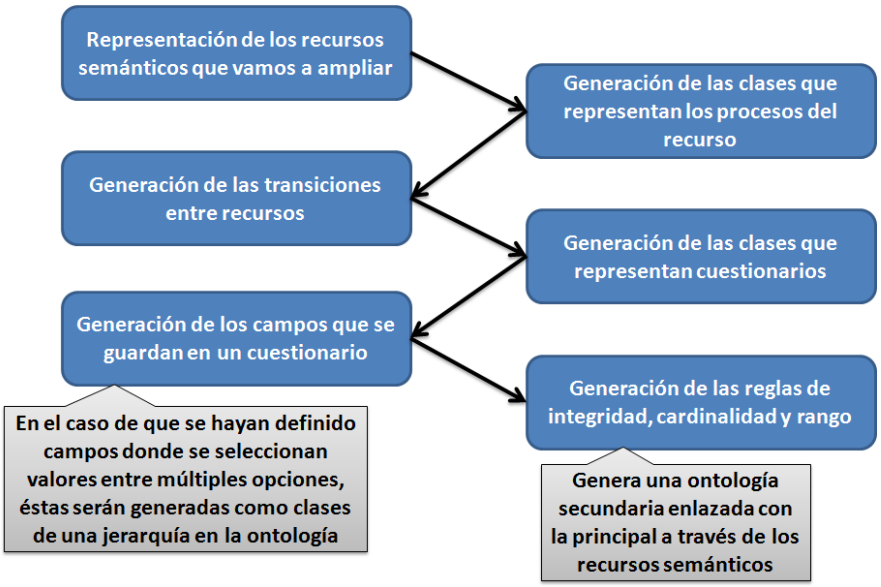


Figura 4.10: Traducción semántica

La ontología generada se almacenará en un gestor documental con control de versiones. Este control de versiones será necesario para explotar información que no esté modelada en la última versión de la ontología pero de la que se registraron datos previamente. Es decir, el usuario tiene la posibilidad de volver a una versión anterior del modelo ontológico y automáticamente recuperar los datos RDF almacenados para ese modelo.

4.2.4 Motor de ejecución semántico

El motor de ejecución es el encargado de gestionar los procesos de inserción y actualización de los datos en los cuestionarios. También se encarga de gestionar las fases del ciclo de vida del recurso. El motor de ejecución tiene dos modalidades:

- **Modo borrador.** El motor de ejecución no comprueba que los datos introducidos cumplan las reglas de integridad de la información. Por ejemplo, si se ha definido un nuevo campo fecha que tiene cardinalidad mínima de uno y no se ha introducido, el sistema no dará error y almacenará esa información. Sí que comprobaría errores del tipo si la fecha está en un cierto rango, si es un número entero positivo, si cumple una expresión regular, etc.

- **Modo normal.** El motor de ejecución comprueba que todos los datos han sido introducidos siguiendo las reglas semánticas definidas.

En ambas modalidades, los datos son almacenados en el repositorio semántico. La diferencia principal en el almacenamiento es que cuando los datos están en borrador, se anotan como tal para que a la hora de usar razonadores para comprobar la integridad de la información no se tengan en cuenta.

4.3 Modelo de evaluación de activos de conocimiento

A continuación se va a describir cómo, usando los elementos anteriores, se puede generar un entorno de trabajo que permite evaluar todos los activos de conocimiento de la organización. Esto permitirá conocer si el conocimiento de la organización es el que realmente se necesita, cuál es el nivel de adquisición de conocimiento del personal de la entidad, y para detectar fortalezas y debilidades.

En la figura 4.11 se puede ver la arquitectura del modelo de evaluación de activos de conocimiento propuesta. Uno de los procesos clave de este modelo es la definición de anotaciones semánticas sobre el catálogo de criterios de evaluación. Gracias a esas anotaciones se pueden definir los mapas de conocimiento para los recursos de la organización que se desea evaluar. Por ejemplo, esa anotación podría servir para definir mapas de competencias para los puestos de trabajo de una organización.

Otro de los procesos clave consiste en la definición de indicadores y de los rangos de los mismos, lo que permitirá saber cuál es el estado actual de la organización. Para poder relacionar los indicadores con la información semántica es importante que en la definición de esos indicadores se establezca qué criterios de evaluación pueden afectar sus métricas.

Los dos procesos anteriores deben revisarse y repetirse cada cierto tiempo para poder comparar la evolución del conocimiento de la organización.

Gracias a la clasificación y catalogación del conocimiento usando criterios de evaluación, éste puede ser evaluado y comparado con él mismo en otros momentos en el tiempo, o incluso con otro conocimiento similar generado por otro departamento u organización. Gracias a las múltiples herramientas descritas anteriormente, se puede recurrir a varios modelos de evaluación dependiendo de las características concretas de cada caso. Este modelo se ha definido de forma genérica, pudiendo adaptarse a múltiples entornos de evaluación (vea la sección 2.6). Esa adaptación la realiza el usuario cuando

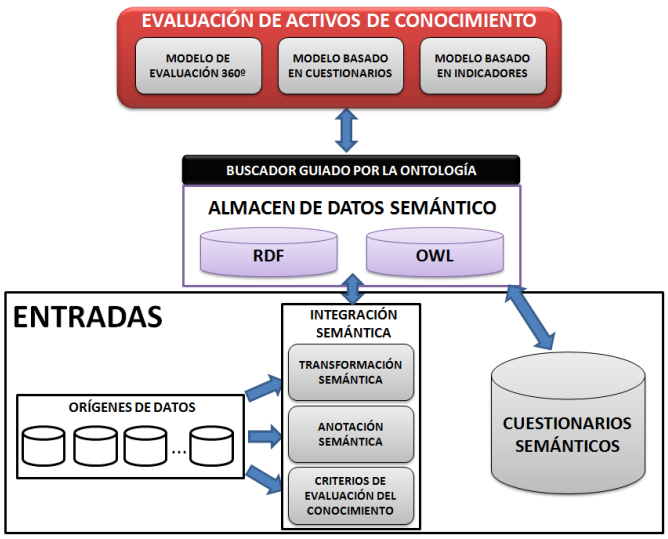


Figura 4.11: Evaluación del conocimiento

modela su catálogo de criterios de evaluación y cuando define el proceso de cálculo de los distintos indicadores.

En la figura 4.12 se puede ver un extracto de la ontología de criterios de evaluación. A continuación se van a describir las principales clases:

- *EvaluationCriteria.* Esta clase representa la unidad mínima de evaluación. Como ya se ha comentado, este concepto permite definir relaciones y jerarquías entre criterios.

- *Score.* Representa las diferentes puntuaciones que puede recibir un criterio de evaluación. Por ejemplo, para un criterio de evaluación denominado "ofimática" se podrían tener conceptos *Score* como "nivel bajo", "intermedio" o "alto". Es importante que se describa la puntuación de cada criterio para que los evaluadores puedan tener unas instrucciones claras de qué puntuación están dando y por qué.

- *AssessableResource.* Esta clase es la raíz de la jerarquía de activos que se pueden evaluar.

- *Importance.* Esta clase representa la importancia del criterio de evaluación con un valor numérico. Está definido como una clase OWL porque es importante describir el significado de cada valor numérico. Por ejemplo, se podrían definir conceptos como "imprescindible", "necesario" o "recomendable".

- *DesirableLevel.* Como ya se describió en el modelo de anotación con criterios de evaluación, esta clase permite establecer relaciones entre los recursos y los criterios de evaluación. *DesirableLevel* almacena información sobre la puntuación mínima requerida que debe tener un recurso sobre un criterio de evaluación en un momento concreto. Este concepto permite la construccion de mapas de criterios de evaluación del conocimiento para cualquier recurso.

- *Evaluator.* Esta clase modela al evaluador de la convocatoria de evaluación. El evaluador puede ser una persona, el cálculo de un indicador o el resultado de un cuestionario semántico que, a su vez, estará asociado a la persona que lo ha realizado. Esto quiere decir que habrá criterios de evaluación que serán puntuados por el criterio de personas o por la propia actividad de la organización.

- *EvaluationCall.* Esta clase representa una convocatoria de evaluación. La convocatoria establece un periodo de evaluación, los recursos que van a ser evaluados y quiénes serán las instancias de la clase *Evaluator*.

- *ResourceEvaluation.* Esta clase representa la puntuación que un *Evaluator* da a un criterio de evaluación en el contexto de una convocatoria de evaluación.

- *ResourceScore.* Esta clase calcula la media de la puntuación de las diferentes evaluaciones recibidas por un recurso para cada uno de los criterios de evaluación.

A continuación se detallan cada uno de los modelos y en un apartado final se discute cuáles podrían ser los más apropiados, dependiendo del problema a solucionar.

4.3.1 Modelo de evaluación 360º

Este modelo está pensado para medir recursos intangibles de la organización, tales como el conocimiento de los trabajadores, la operatividad de los procesos, el liderazgo de los mandos intermedios, la actitud de los trabajadores, etc. [151]. También puede ser útil para evaluar recursos tangibles de los que se carecen de datos para generar indicadores de su rendimiento en la organización.

En la figura 4.13 se puede ver el proceso a seguir para usar este modelo. En primer lugar, los administradores de la plataforma definirán el catálogo de criterios de evaluación del conocimiento de la organización que pretende

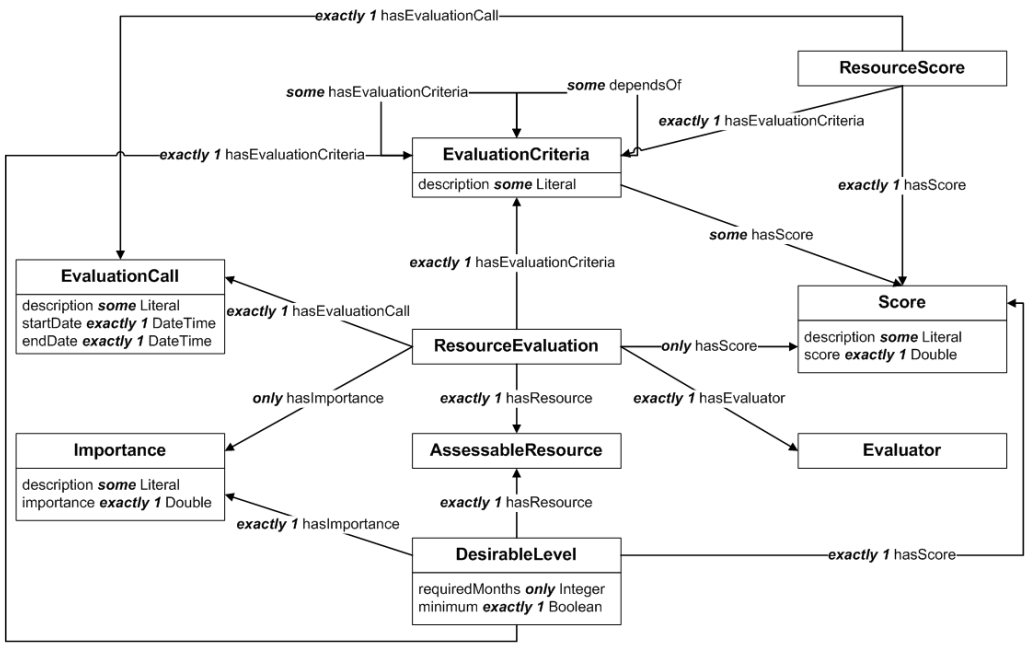


Figura 4.12: Ontología de criterios de evaluación

ser evaluado. Una vez que se dispone de ese catálogo, se usa para anotar los recursos que van a ser evaluados. Gracias a este proceso de anotación se pueden generar los mapas de criterios de evaluación en los que se podrá ver qué es lo que un recurso necesita para operar correctamente. Una vez que los activos están anotados, los administradores ya pueden configurar convocatorias de evaluación en los que para un mismo recurso habrá diferentes evaluadores, existiendo la posibilidad de que en el caso de que se estén evaluando recursos humanos, pueda existir autoevaluación. En la siguiente fase, cada evaluador evalúa el recurso siendo el resultado de este proceso la media geométrica de la evaluación de cada uno, estando los pesos definidos por el concepto *Importance*. Por último, se analizan los resultados obtenidos que son usados para redefinir el catálogo de criterios de evaluación y a su vez las anotaciones, repetiéndose de nuevo todo el ciclo.

Un aspecto importante de este modelo son los diferentes mecanismos que se han usado para que la evaluación sea lo menos subjetiva posible. El primer mecanismo consiste en que la evaluación de un recurso siempre se haga por más de un evaluador. El segundo mecanismo consiste en usar el campo *description* del concepto *Score* para que el evaluador tenga una descripción clara de por qué ese criterio debe tener esa puntuación. Por ejemplo, para un criterio de evaluación que evalúe la capacidad organizativa de un determina-

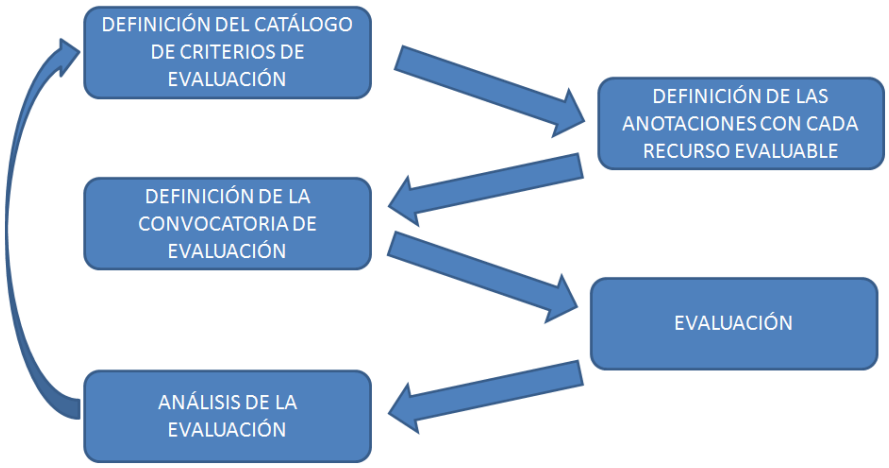


Figura 4.13: Proceso de evaluación del modelo 360º

do departamento, se podrían tener tres posibles puntuaciones: la puntuación (1) se describiría como que el departamento no realiza reuniones periódicas, la (2) como que hacen reuniones periódicas de seguimiento de los proyectos y la (3) que hacen reuniones periódicas y se redactan actas de cada una de ellas con los siguientes pasos a dar. De este modo, cuando el usuario va a evaluar, no usa el valor numérico sino que ve el descriptivo definido, restando subjetividad a la evaluación.

4.3.2 Modelo basado en cuestionarios

Este modelo está pensado para completar el modelo de evaluación 360º con la definición de cuestionarios semánticos que permitan conocer el nivel de conocimiento o el comportamiento social de un trabajador de la organización.

Al igual que en el caso anterior, se parte de que se ha definido un catálogo de criterios de evaluación y que éstos se han asociado a los recursos que se desea evaluar. Como se observa en la figura 4.14, el cambio está a la hora de convocar la evaluación. En este caso, en primer lugar hay que definir los cuestionarios que permitirán conocer el valor de la evaluación para uno o varios criterios de evaluación en un recurso concreto. Estos cuestionarios deben ser cuantificables, es decir, deben devolver una puntuación, por lo que siempre deberán ser cuestionarios de tipo test. Además, los resultados de dichos cuestionarios ayudarán a realizar una puntuación concreta en uno o varios criterios de evaluación que se asocien a éstos.

Para poder generar los resultados con este modelo se han desarrollado

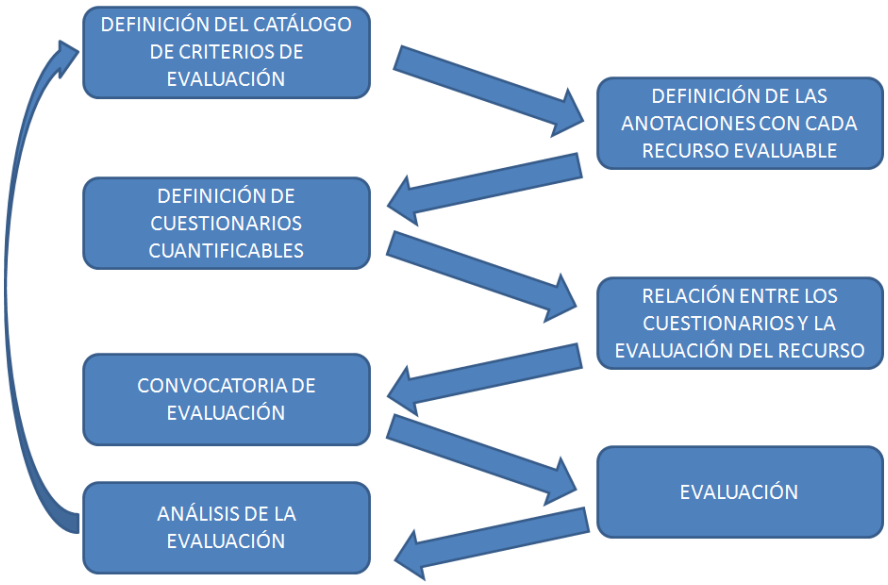


Figura 4.14: Proceso de evaluación del modelo basado en cuestionarios

dos herramientas:

- Calculadora semántica. Esta herramienta, guiada por la ontología, permite coger los valores cuantificables de los campos y operar con ellos añadiendo operaciones aritméticas. Por ejemplo, si se quiere evaluar un tipo test y hay preguntas que valen el doble que el resto, se podría usar una media geométrica asignando pesos a dichos campos.
- Definición de umbrales. Esta herramienta permite establecer umbrales para los resultados cuantificables de cada cuestionario. Los umbrales se definen como valores numéricos entre un máximo y un mínimo. Cuanto más cerca esté del valor máximo será un valor más cercano al óptimo.

Aunque este modelo es perfectamente válido para evaluar el conocimiento de un trabajador, se considera oportuno la combinación con el modelo 360º, que permitirá comprobar que además de tener los conocimientos, los aplica en su puesto de trabajo.

4.3.3 Modelo basado en indicadores semánticos

Este modelo es útil para evaluar objetivamente el rendimiento de la organización al completo o de parte de la misma. Este modelo se basa en la información del almacén de datos semántico. Para extraer los datos que se necesitan

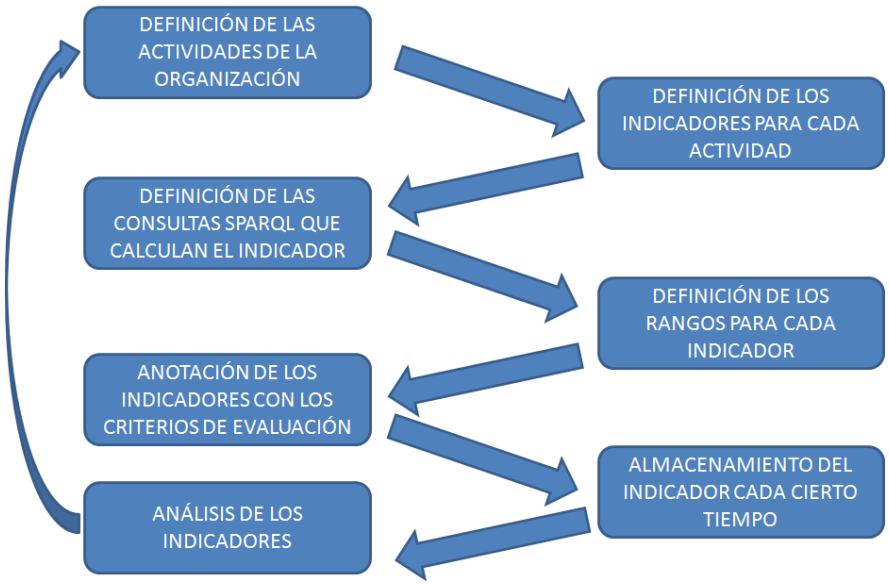


Figura 4.15: Proceso de evaluación basado en indicadores

para realizar los cálculos de los indicadores se usan consultas SPARQL definidas con ODS. Esto quiere decir que cada usuario del sistema podrá definirse sus propios indicadores e incluso compartirlos y compararlos con los de otro usuario. En la figura 4.15 se puede ver el proceso de definición de este modelo. En primer lugar se definen las actividades de la organización que se desean medir. En segundo lugar se definen los indicadores, identificando la consulta SPARQL que los calcula y los rangos que identifican si el valor medido es positivo o no. Para poder relacionar los indicadores con la evaluación de criterios, se anotan como un recurso en el que su rendimiento está relacionado con uno o varios criterios que a su vez evalúan otros recursos. Posteriormente se almacena el valor del indicador cada cierto tiempo automáticamente, para poder evaluar su evolución en el tiempo. Al final, se analizan los indicadores y se detecta si la mejora o el empeoramiento de los criterios ha influido en el nuevo valor del indicador. Esto será de utilidad para volver a definir nuevos indicadores, o añadir a éstos nuevos criterios de evaluación.

En la figura 4.16 se puede ver un extracto de la ontología[2] que modela los indicadores. Como se puede observar, *Indicator*, que es el concepto que representa al indicador, está en la jerarquía de *AssessableResource* lo que quiere decir que se le pueden asociar los criterios de evaluación que afectan al rendimiento de ese indicador. A continuación se describen las conceptos

[2]http://sele.inf.um.es/ontologies/IndicatorsV1.owl

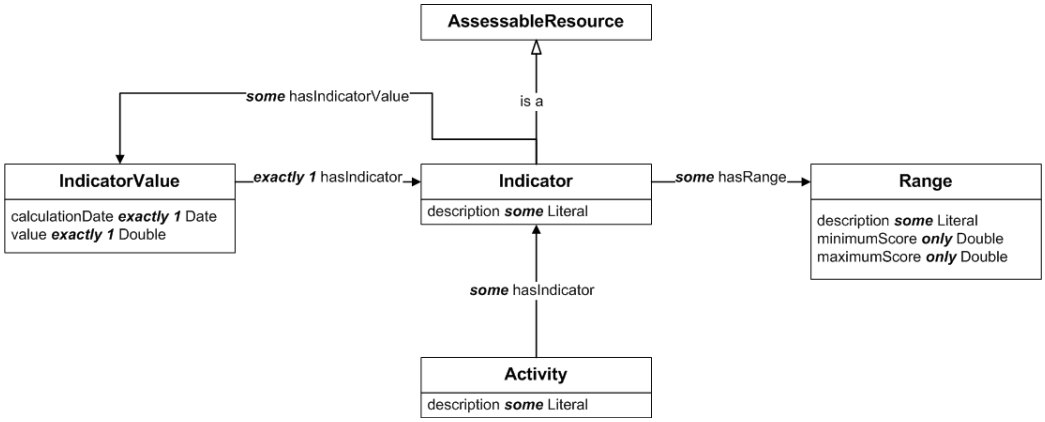


Figura 4.16: Extracto de la ontología de indicadores

que aparecen en ese extracto:

- *Activity.* Representa las actividades de la organización que van a ser evaluadas. Cada actividad tiene un conjunto de indicadores que describen su rendimiento.

- *Indicator.* Modela el indicador en sí. Cada indicador se define descriptivamente y a partir de unos umbrales que miden los valores.

- *Range.* Representa el rango de valores del indicador entre un valor máximo y uno mínimo. Además de un descriptivo del rango, también almacena un valor numérico que indica si el indicador roza la excelencia o si hay que mejorar notablemente la actividad para que el indicador mejore. Si no se define el indicador máximo se asume que tendrá ese rango cuándo el valor sea mayor o igual que el mínimo. Si no tiene el mínimo se asume que el indicador tendrá ese rango si su valor numérico es menor que el máximo.

- *IndicatorValue.* Almacena el valor del indicador en una fecha determinada. Para que el modelo sea más eficiente, cada vez que se calcula un indicador es almacenado como tal calculado. Por lo tanto, para ver un indicador mensual no se necesitaría calcular éste para cada uno de los meses, sino únicamente para el último, con el ahorro de tiempo y coste computacional que ello conlleva.

4.3.4 Modelo mixto

El modelo mixto permite combinar varios de los modelos anteriores. Para ello es importante que la definiciones de cuestionarios y de indicadores semánticos estén asociadas con criterios de evaluación.

Un ejemplo puede ayudar a aclarar este modelo. En una organización existe un puesto de trabajo cuya actividad principal es solucionar incidencias de clientes a través de una herramienta en línea. Los directivos están interesados en evaluar la eficiencia y la eficacia. Para ello se han definido los siguientes criterios: (1) manejo de la herramienta telemática para la resolución de incidencias, (2) grado de satisfacción de los usuarios de esta herramienta, (3) eficiencia resolviendo incidencias.

Para evaluar estos criterios se van a usar los siguientes modelos:

1. Evaluación del criterio 1. Se usan los cuestionarios para proponer un examen en el que se evalúe el conocimiento que cada usuario tiene al manejar la herramienta de incidencias.

2. Evaluación del criterio 2. Se definen una serie de criterios de evaluación semánticos, es decir, conceptos semánticos que clasifican la información. Por ejemplo: (1) la incidencia ha sido resuelta satisfactoriamente, (2) el trato del usuario ha sido de su agrado y (3) la rapidez de la respuesta a la incidencia. Una vez definidos esos tres criterios se puede hacer una evaluación 360º con todos los clientes.

3. Evaluación del criterio 3. Para este criterio se tendría en cuenta indicadores como el tiempo medio desde que se registra la incidencia hasta que se marca como resuelta. Este indicador estaría relacionado con los criterios de evaluación que se han usado en los puntos 1 y 2.

Gracias a esa evaluación se podría inferir fácilmente cuáles son los puntos débiles y fuertes identificando, por ejemplo, si la evaluación (1) no es del todo buena, que probablemente el hecho de que el tiempo medio de respuesta de la incidencia sea elevado se deba a que los usuarios pueden mejorar en el manejo del software informático.

4.3.5 Comparativa de los modelos de evaluación

En la tabla 4.1 se puede ver una comparativa entre los diferentes modelos de evaluación, destacando cuáles son sus ámbitos de aplicación, sus ventajas y sus desventajas.

Tabla 4.1: Comparativa de los modelos de evaluación

Modelo	Ámbito	Ventajas	Inconvenientes
Evaluación 360º	Es un modelo pensado para evaluar trabajadores, aunque el modelo lo extiende para que pueda evaluar cualquier recurso de la organización.	Este modelo permite evaluar el rendimiento de los recursos de la organización desde el prisma de diferentes personas, tanto internas como externas a la organización.	El hecho de que la evaluación sea realizada por personas añade subjetividad a la evaluación pudiendo derivarse en resultados incorrectos.
Cuestionarios Semánticos	Permite elaborar test de evaluación de conocimientos o actitudes sociales.	Es útil para que la organización tenga una cartera de los conocimientos y actitudes de los trabajadores de la organización.	El hecho de que una persona tenga mucho conocimiento en un aspecto concreto no quiere decir que lo aplique en su día a día. Otra desventaja de este modelo es que sólo es aplicable a recursos humanos.
Indicadores	Ideal para la evaluación del rendimiento de la compañía.	El cálculo de este modelo de evaluación es totalmente objetivo ya que se calcula a partir de los propios datos de la empresa	No tiene en cuenta el conocimiento intangible para realizar la evaluación.

4.4 Perfil semántico de una entidad de negocio

Se define el perfil semántico de un recurso como un subconjunto de sus propiedades, unidas al perfil semántico de un subconjunto de las instancias a las que está directamente relacionado. Esto quiere decir que la definición de perfil semántico es recurrente y se calcula usando la siguiente fórmula:

$$PS(r) = S(d) \cup S(PS(io)) \tag{4.1}$$

donde $PS(r)$ simboliza el cálculo del perfil semántico de un recurso r, $S(d)$ representa un subconjunto de sus propiedades, y $S(PS(io))$ representa un subconjunto de los perfiles semánticos de las instancias relacionadas i a través de la relación o.

Este cálculo permite generar nuevas representaciones a diferentes niveles semánticos de la información, y es muy útil a la hora de realizar análisis concretos sobre el almacén de datos.

Conceptualmente un perfil semántico es un conjunto de relaciones y propiedades que definen a una entidad de negocio para resolver un problema de explotación concreto. Es decir, son resúmenes semánticos de la información que realmente es importante explotar en un caso concreto, y que facilita tanto la eficiencia del análisis como la comprensión de los resultados. Además, los perfiles semánticos permiten identificar grupos de entidades que comparten algunas propiedades, y son útiles para comparar y estudiar estos grupos. Las ontologías son de especial interés para crear estos perfiles, ya que permiten la selección y agregación de entidades desde una perspectiva conceptual.

La definición del perfil semántico de una entidad se realiza usando ODS. En una primera fase se definen cuáles van a ser las consultas SPARQL que van a devolver el subconjunto de propiedades y relaciones que mejor describen la entidad de negocio para el caso concreto de explotación. Posteriormente esas consultas son ejecutadas para cada una de las entidades, y los resultados generan una nueva información semántica más plana, es decir, se genera una caché de la información semántica más importante que describe a esa entidad.

Un aspecto a destacar de la definición del perfil es la posiblidad de definir líneas de tiempo de un recurso, es decir, este modelo permite definir la información de una parte del perfil de la entidad de negocio en un momento concreto de tiempo, generando una colección de los valores o relaciones que tenía el recurso enlazada cronológicamente. Por ejemplo, cuáles han sido la terapias aplicadas a un enfermo crónico durante los primeros 10 años de su enfermedad.

Los perfiles semánticos permiten agrupar entidades que tengan criterios comunes. Esta agregación genera un nuevo concepto ontológico cuyos axio-

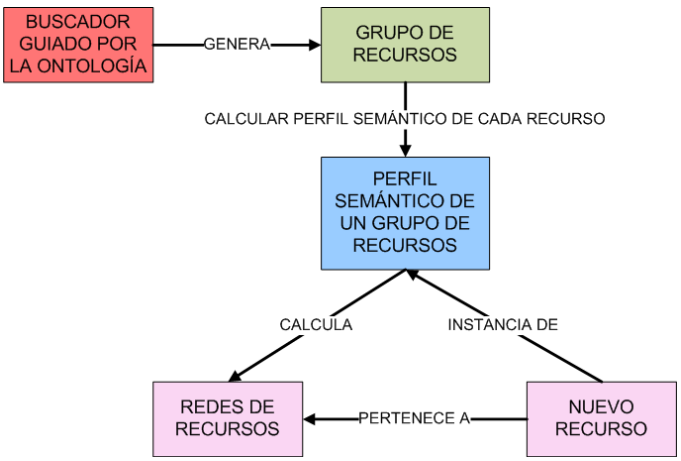


Figura 4.17: Perfil semántico agregado

mas quedan identificados por esos criterios. Esto permite que cuando registran nuevas entidades, éstas se clasifiquen automáticamente como instancias de ese concepto.

Como se puede ver en la figura 4.17, se usa ODS para hacer una búsqueda de aquellas entidades de negocio que tienen unas propiedades comunes. Para cada una de esas entidades se define el perfil semántico y se genera un nuevo concepto en la ontología que engloba a éstas. Cada vez que una entidad nueva se almacene en el sistema, si cumple las condiciones establecidas se incluirá como instancia del concepto. Estas asociaciones permitirán generar redes entre los diferentes grupos a través de sus perfiles, lo que a su vez habilitará la generación de algoritmos de clasificación a partir de funciones de similitud semántica, o de redes bayesianas construidas sobre las distribuciones de probabilidad existentes entre las diferentes propiedades de los recursos de la red.

Para entender mejor este concepto se ha definido el siguiente ejemplo: en una universidad se gestionan proyectos de investigación que tienen una serie de datos como el equipo investigador, la facultad o facultades que participan (incluyendo sus departamentos), las partidas presupuestarias concedidas, la entidad financiadora, documentación asociada, facturas, personal contratado y producción científica generada (publicaciones, tesis, patentes, etc.). A la hora de explotar la información se quiere saber cuánto dinero ha costado la producción científica generada. Para generar este indicador se tienen dos opciones:

1. Cada vez que se desee calcular el indicador, se genera una consulta

con ODS que suma todos los gastos, incluyendo facturas y personal contratado. Por otro lado, se genera otra consulta con ODS que calcula el número total de producción científica generada. Dividiendo los resultados de la primera por la segunda se tendría el resultado global.

2. Generar un perfil semántico de cada proyecto en el que se defina una propiedad que incluya los gastos generados. Para ello se genera la consulta con ODS que suma los gastos y crea la información semántica de los resultados en el almacén semántico. Esto implica que, además de la información generada de las fuentes de datos, los proyectos incluirán un campo directo con la cantidad de dinero invertida. Por otro lado se puede generar un perfil semántico para la producción científica en el que haya una nueva propiedad en la que se almacena el coste analítico de esa producción. Con esos dos perfiles semánticos sería muy fácil para los usuarios generar una única consulta con ODS que indique el coste total por producción o por proyecto.

Como se ve a través de esta segunda opción, la primera vez que se calcula el indicador va a ser un proceso más costoso, ya que el usuario debe definir dos perfiles semánticos antes de poder calcular el indicador. Sin embargo, para las siguientes veces dispondrá de la información de forma directa, siendo más fácil para el usuario y más eficiente para el sistema el cálculo de los resultados.

Siguiendo con el ejemplo anterior, un perfil semántico agregado podría ser un nuevo concepto ontológico que clasifique la producción científica por el coste que tiene, o que clasifique los proyectos por el gasto de cada producción científica generada. Para ello, se generaría una nueva consulta con ODS que explotara los resultados de los perfiles semánticos de cada recurso, y que posteriormente generaría ese nuevo concepto siguiendo esas reglas. Los perfiles semánticos agregados son muy útiles para la construcción de algoritmos de clasificación según la similitud entre sus individuos. Dicha similitud se calcula a partir del perfil semántico de cada individuo, mejorando la eficiencia del algoritmo.

4.5 Modelo de explotación

A continuación se van a describir las diferentes herramientas que se han desarrollado para explotar la información contenida en el almacén semántico.

4.5.1 Buscador semántico

Esta herramienta hace uso de ODS para poder realizar búsquedas sobre el modelo de datos semántico. Los resultados de esas búsquedas pueden enlazarse con otros recursos con los que estén relacionados. Es importante destacar que cada uno de los filtros puede ser almacenado para ser parametrizado y reutilizado posteriormente.

El buscador semántico también permite definir nuevas clases en el modelo semántico, es decir, se pueden definir nuevos conceptos y poblarlos con individuos a partir de los resultados de la consulta. Ésta queda almacenada para ser ejecutada cada cierto tiempo, actualizando la información nueva que se vaya produciendo en el proceso de negocio de la organización. Por ejemplo, si se quiere filtrar pacientes que tengan un determinado diagnóstico, como puede ser diabetes, se puede usar esa consulta para crear el concepto "Paciente diabético" que denomine a esos pacientes y poblarlo con los resultados de la búsqueda, de tal forma que la siguiente vez sólo haya que buscar instancias de esa clase para reproducir ese filtro.

4.5.2 Gestión de alertas

Las alertas se generan como consultas SPARQL agregadas, es decir, devuelven un valor numérico gracias al uso de las funciones de agregación como *count*, *avg*, etc., en las que los resultados son diferentes a los valores esperados. Por ejemplo, se podría generar una alerta cuando el resultado de una consulta sea mayor o igual que cero, esté entre un determinado umbral, o incluso comparar resultados de diferentes consultas.

Las alertas se definen en el sistema como la tripleta $\langle x, alert, u \rangle$ donde x es el resultado de una consulta SPARQL que está fuera de lo esperado, *alert* representa la descripción de la alerta y u es la persona o personas que recibirán la alerta.

Esta herramienta hace uso de ODS para generar alertas de diferentes tipos:

- Alertas de consistencia de los datos. Se genera una alerta cuando una búsqueda devuelve resultados que son inconsistentes. Por ejemplo, no pueden existir carteras de inversión para fondos que han sido dados de baja en el sistema. Otro ejemplo podría ser que no pueden existir mujeres con cáncer de próstata.

- Alertas de funcionamiento. En este caso, se pueden combinar varios filtros de búsqueda. Por ejemplo, se podría generar una alerta cuando

los ingresos de la empresa son menores que los gastos en un determinado periodo contable. También se pueden generar cuando no se han realizado las correspondientes visitas a un paciente reclutado en un ensayo clínico. También se pueden generar alertas cuando los resultados de la evaluación de los activos de conocimiento esté por debajo de los umbrales establecidos.

Además de estas alertas, configuradas por el usuario, esta herramienta tiene una serie de consultas predefinidas para generar diversas alertas de funcionamiento interno como: debe realizar una evaluación del conocimiento de algún recurso de la empresa, tiene cuestionarios semánticos pendientes de rellenar, o tiene recomendaciones del sistema de recomendación (que será descrito más adelante).

4.5.3 Cuadros de mando semánticos

Un cuadro de mando semántico se define como una o varias consultas que devuelven valores que son representados gráficamente. Los cuadros de mando semánticos se representan como $\langle\langle L, V\rangle, isDashboard, U\rangle$ donde $\langle L, V\rangle$ son los resultados de la consulta SPARQL devueltos en pares clave L y valor V. U es el usuario que ha definido el cuadro. Esto quiere decir que cada usuario de la plataforma puede definir sus cuadros de mando personalizados.

Al igual que los servicios anteriores, esta herramienta se apoya en ODS. En este método se ha incorporado al generador de consultas semánticas la posibilidad de realizar agregaciones de los datos. Dichas agregaciones pueden ser presentadas gráficamente y en formato de tablas (que pueden ser anidadas si el concepto por el que se agrupa está en alguna jerarquía de la ontología). Gracias al modelo de persistencia de las consultas SPARQL se puede navegar desde estas representaciones a un listado de los datos sin agregar, es decir, convertir cada agregación del cuadro de mando concreto en un filtro de búsqueda del buscador semántico.

Esta herramienta también permite la posibilidad de definir varias consultas agregadas y obtener la representación gráfica de su comparativa.

Por último, gracias al motor de persistencia de ODS los cuadros de mando también pueden persistirse para poder ser parametrizados y reutilizados por los usuarios cuantas veces lo deseen. Esto permite que cada usuario de la plataforma pueda personalizar sus propios cuadros de mando de la organización.

4.5.4 Módulo de recomendación

En este sección se describen los dos algoritmos que se han integrado en el marco de trabajo para poder hacer recomendaciones personalizadas a los usuarios de cuáles deberían ser sus siguientes pasos en la actividad laboral que desempeñan en su organización.

Las recomendaciones pueden utilizar la herramienta de evaluación del conocimiento y generar entornos de confianza que sean usados como variables a la hora de recomendar, es decir, la aplicación priorizará la recomendación basándose en la puntuación que se le otorga a los diferentes recursos.

El cálculo de la recomendación no se hace sobre el conjunto de la ontología, sino que se calcula con los perfiles semánticos que se decida emplear. Se usan estos elementos por eficiencia y usabilidad para el usuario.

4.5.4.1 Módulo de recomendación ontológico

Este modelo de recomendación se basa en las relaciones entre conceptos ontológicos para realizar las recomendaciones.

En este caso, el algoritmo usa los perfiles semánticos del elemento que se quiere recomendar para generar una red con otros perfiles que tengan propiedades en común. Para conocer qué perfiles se deben incluir se usará una variación de la función de similitud basada en nodos propuesta por [88]. Esta extensión consistirá en la inclusión de un parámetro que permita medir la amplitud de la red generada, es decir, el número de relaciones semánticas que pueden rodear al nodo central de la red. Además, se añade un segundo filtro que tenga en cuenta los resultados de la evaluación de los activos de conocimiento, es decir, que la función de similitud descarte aquellos recursos que no tienen una evaluación por encima de un determinado umbral.

Para entender cómo funcionaría este módulo se describe un ejemplo. Se tiene un portal Web en el que se publican películas. En esa red social, cada usuario registrado puede definir sus intereses cinematográficos, puede recomendar películas y puede puntuar las películas. Si se quisiera ofrecer al usuario un recomendador de películas, se podría usar este modelo para ofrecerle directamente aquellas películas que encajan en sus intereses y que han sido bien valoradas. Este tipo de recomendación sería muy sencilla de realizar en un sistema relacional tradicional. La ventaja fundamental que incorporarían las ontologías consiste en que, gracias a la generación de perfiles de las películas y de los usuarios, se podría recomendar, además de las películas que están directamente relacionadas con sus intereses, otras que estuvieran también bien valoradas y cuyo contenido estuviera próximo a sus intereses, de tal forma que se puede ampliar el rango de recomendación fácilmente. Además,

no sólo se podrían usar las valoraciones de las películas, sino las valoraciones de los usuarios obtenidas a partir de los comentarios que realizan. Esto permitiría ampliar el rango de recomendación que se le ofrece al usuario, incluso alertarle sobre los comentarios, publicados por otros usuarios con intereses comunes, que debería revisar.

4.5.4.2 Módulo de recomendación bayesiano

Este modelo usa un algoritmo basado en redes bayesianas para proponer las recomendaciones. Como el modelo anterior, se basa en la generación de modelos probabilísticos usando los nodos de los perfiles semánticos. Las redes de bayes tienen una limitación y es que no pueden tener ciclos dentro de la red [170]. Sin embargo, en un modelo ontológico, esos ciclos sí que pueden existir y en la red generada por los perfiles semánticos es factible encontrarse con esas situaciones.

Para solventar este problema, cuando pueden existir ciclos en los perfiles semánticos, se construye un árbol de redes sobre cada uno de los perfiles que se han generado. Por ejemplo, si se tuviera una aplicación epidemiológica de pacientes con diabetes, se tendrían almacenados los tratamientos que se han ido aplicando al paciente desde el diagnóstico. Esos tratamientos pueden estar repetidos, por lo tanto, en el perfil semántico del paciente a lo largo del tiempo podrían existir ciclos. Si se quisiese saber cuál es el tratamiento más probable para un paciente con una serie de características, el modelo recuperaría todos los pacientes con ese criterio, y usando su perfil semántico generaría un mapa de redes bayesianas con los posibles tratamientos periodo a periodo (mes, trimestre, etc.). Cuando el usuario eligiera en el primer tratamiento uno concreto entre los posibles, todos los niveles siguientes del mapa se recalcularían para recomendar cuáles son los tratamientos siguientes más probables y así sucesivamente. Es decir, conforme el usuario va indicando características concretas que pueden estar en un ciclo de la red, se va bajando en el árbol construido para evitar que el algoritmo entre en un bucle infinito.

4.5.5 Plan semántico

Se define "plan semántico" como el conjunto de tripletas $\langle x, y, z\rangle$ donde x es el criterio de evaluación cuya puntuación se desea mejorar a través del plan, y es la relación entre el criterio de evaluación y el recurso, y z es el recurso que ayudará a mejorar el criterio de evaluación. Estas tripletas se ordenan en función de la prioridad de los criterios que mejoran (gracias al concepto *Importance*) y se limitan por dos variables: tiempo necesario para llevarlo a cabo y coste.

Los recursos se clasifican según tres subconceptos que no son disjuntos entre sí: (1) oportunidades de mejora, (2) oportunidades comerciales y (3) amenazas. Las oportunidades de mejora se diferencian de las oportunidades comerciales en que las primeras suponen una inversión mientras que las segundas suponen un ahorro ya que producen beneficios. Al no ser conceptos disjuntos, un mismo recurso puede ser a la vez una oportunidad de mejora y una amenaza. Por ejemplo, si se tuviese una empresa en la que es muy importante la ofimática. Un ejemplo de amenaza que a la vez es una oportunidad de mejora podría ser la llegada de una nueva versión de un paquete de software ofimático.

Este módulo explota los resultados de la evaluación del conocimiento y usa los perfiles semánticos para generar planes graduales, es decir, planes para una persona, para un departamento, para un grupo de elementos con características comunes, etc.

El sistema de planificación tiene una versión básica y otra más avanzada. En la versión básica, gracias a la generación de los perfiles semánticos se puede analizar cuál es el estado deseable de la organización y cuál sería el estado real. En esa primera identificación, la plataforma es capaz de indicar automáticamente al usuario cuáles son las fortalezas y las áreas que se deben mejorar. Es decir, es capaz de ofrecer al usuario el análisis interno de un DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades) [171]. Este análisis cubriría los aspectos de debilidades y fortalezas, pero dejaría fuera las amenazas y oportunidades.

La versión avanzada incorpora esos dos elementos del análisis DAFO. Es decir, el modelo de clasificación semántica en criterios de evaluación permite que éstos se clasifiquen con otros recursos internos o externos que están relacionados con las capacidades de mejora de la organización. Por ejemplo, si una persona para desempeñar correctamente su puesto de trabajo debe tener conocimientos avanzados en ofimática, se puede crear un criterio de evaluación que mida el conocimiento de informática de un trabajador. Además, se puede usar ese propio criterio para anotar un recurso formativo que ayuda a manejar de forma avanzada una hoja de cálculo. Si cuando se evalúa el desempeño del trabajador se detecta que no cumple los niveles deseados, se podría generar un plan de desarrollo personal que incluyera superar esa acción formativa. Otro ejemplo de amenaza podría ser que tras evaluar un proceso de la organización se detectara que no cumple los requisitos mínimos de funcionamiento (medidos con criterios de evaluación). Esos mismos requisitos podrían ser útiles para clasificar los procesos de otras organizaciones que son más ágiles a la hora de generar productos similares.

En la tabla 4.2 se describen los diferentes escenarios que se pueden en-

contrar y qué peso recibirá cada uno. Este peso, junto con las variables de tiempo y económicas ayudarán a establecer el mejor plan posible.

Tabla 4.2: Escenarios de planificación

Análisis interno	**Análisis externo**	**Planificación**
Fortalezas	Oportunidades de mejora	El algoritmo de planificación considerará que estas oportunidades de mejora tendrán prioridad baja a la hora de generar el plan estratégico, ya que tienen un coste en un aspecto en el que la organización ya es fuerte.
Fortalezas	Oportunidades comerciales	El algoritmo de planificación dará máxima prioridad a este tipo de oportunidades, ya que van a producir beneficios en la empresa.
Fortalezas	Amenazas	El algoritmo de planificación considera que esas amenazas pueden influir en el futuro de la empresa. Las amenazas se pueden relacionar con las oportunidades de mejora, lo que aumentaría la prioridad de este caso para prevenir que esas amenazas puedan afectar a las áreas fuertes de la organización.
Debilidades	Oportunidades de mejora	El algoritmo de planificación dará prioridad máxima a estas oportunidades, ya que supondrán una mejora sustancial en la organización.
Debilidades	Oportunidades comerciales	En este caso estas oportunidades serán desechadas, ya que al no ser un área fuerte de la organización pueden suponer un problema a largo plazo si los resultados no son satisfactorios.
Debilidades	Amenazas	En este caso, al igual que pasaba con las fortalezas, el algoritmo priorizará aún más este caso al suponer una gran amenaza a la operatividad de la organización.

El peso de cada uno de los escenarios será configurable por los administradores de la plataforma, otorgando mayores diferencias dependiendo del

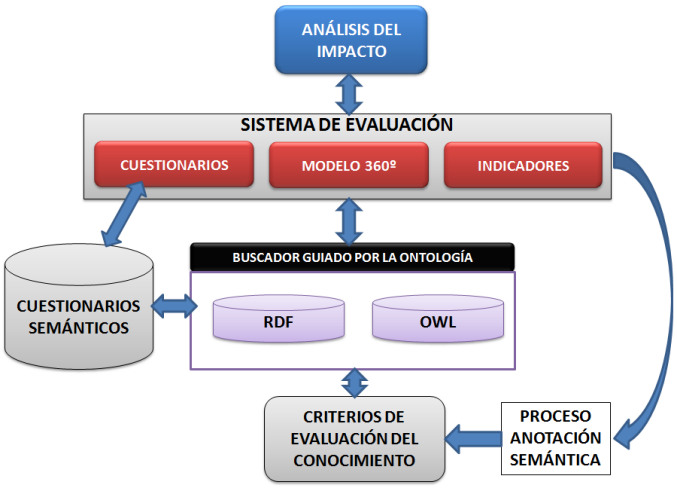


Figura 4.18: Análisis del Impacto

tipo de organización.

4.5.6 Análisis del impacto

El análisis del impacto permite medir el impacto real que ha tenido un plan en la organización. Semánticamente se representa como la tripleta $\langle R, y, \langle C_a, C_n \rangle \rangle$, donde R es el recurso o acción ejecutada, y indica si mejora o empeora al criterio de evaluación C en relación a la evaluación obtenida en el momento a frente a la evaluación en el momento actual n, siendo siempre $a < n$. Esta herramienta pretende analizar el efecto que han tenido una o varias medidas que se han llevado a cabo en una organización. El sistema de análisis del impacto se basa en la definición de indicadores, la anotación con criterios de evaluación y las evaluaciones de los activos de conocimiento realizadas. En la figura 4.18 se puede ver el esquema de funcionamiento de esta herramienta. Una de las principales características es que se necesita relacionar todos los recursos generados en el módulo de evaluación de activos de conocimiento con los criterios, lo que permitirá relacionar los resultados de cada evaluación con los mapas de criterios generados. Por ejemplo, se puede comparar fácilmente el nivel requerido de una determinada competencia, y el nivel real evaluado.

En muchas organizaciones, cuando se toma una decisión de cambio estratégico se suele ver cómo estaban los indicadores antes y después para poder analizar qué ha pasado. Con el tipo de herramienta que se proporciona en esta tesis se avanza varios pasos más allá y se responde a las siguientes cuestiones: ¿Está totalmente seguro de que el cambio estratégico es el que realmente ha

hecho que cambien sus indicadores? ¿Conoce los efectos colaterales que ha tenido ese cambio? ¿Qué debería cambiar del modelo de conocimiento después de aplicar este cambio?

La propuesta que se hace es muy sencilla, si se saben los indicadores que se desean medir, se evalúan los recursos y se relacionan los criterios de evaluación con esas evaluaciones (en el caso de evaluacion 360º no es necesario ya que la propia evaluación se hace sobre los criterios), se va a poder tener una respuesta totalmente automática a todas esas preguntas.

Se expone un ejemplo para entenderlo mejor. En una organización, cuya actividad principal es la fabricación de tornillos, se decide comprar una nueva máquina que permite fabricar más tornillos por hora. En la compra de esa máquina se incluye un curso de formación para los operarios. Al cumplirse el primer año de la compra de la máquina, el gestor mira el indicador de tornillos fabricados y ve que es mucho menor a los del año anterior, fabricados con la máquina antigua. Con esta información, el propietario de la empresa podría verse tentado a volver al modelo antiguo y perder la inversión realizada. Ahora se verá qué pasaría con el modelo propuesto. En este caso, se definirían una serie de criterios de evaluación que van a describir la nueva máquina, los operarios de ésta y el curso que reciben en su manejo, ya que ese curso ha pasado a formar parte del conocimiento necesario en la organización. Además, gracias al entorno de trabajo, automáticamente se han generado herramientas que permiten evaluar la máquina, al operario y ese curso de formación. Al relacionar esos criterios de evaluación con los indicadores esperados y cruzando esa información con la evaluación, la aplicación puede detectar automáticamente que el operario no se ha formado adecuadamente en el manejo de la máquina o que el curso de formación no es suficiente para operar con la máquina, por lo que tendría una información de gran valor para decidir si debe volver a la máquina antigua, si debe mejorar la formación de su personal o, incluso, si alguno de sus operarios puede enseñar a otros ya que su manejo del aparato es mucho más eficaz.

Este modelo se puede complicar retroalimentándose con los resultados de evaluación y los indicadores obtenidos en diversos periodos de tiempo, analizando cuáles son las áreas de mejora e incluso aquellos elementos que no tienen incidencia real en el rendimiento de los indicadores, proporcionando un entorno de gran utilidad para la ayuda a la toma de decisiones.

4.6 Soluciones de IN

A continuación se van a describir los diferentes marcos de trabajo que se han desarrollado combinando e integrando los métodos y herramientas propues-

tas. En el siguiente listado se verán las diferentes soluciones desarrolladas y cuál ha sido su motivación:

- **Red social.** Como ya se ha comentado, este dominio es uno de los retos principales de la IN [26]. En este caso se ha desarrollado una plataforma que combina los típicos datos en lenguaje natural con fuentes de datos estructuradas para ofrecer servicios de búsqueda y recomendación semántica.

- **Sistema de planificación estratégica.** La planificación estratégica en las empresas es uno de los objetivos a los que da soporte la IN [172]. En esta plataforma se ha desarrollado un sistema que, a partir de fuentes de datos estructuradas, la herramienta de evaluación de activos de conocimiento y de recursos clasificados semánticamente, es capaz de generar planes estratégicos a medida de las necesidades de cada uno de los departamentos de una organización.

- **Epidemiología.** La epidemiología es una de las disciplinas que contribuye a la mejora de la prevención y las terapias de algunas enfermedades. La epidemiología se basa en recoger registros de casos clínicos (más o menos estructurados), que posteriormente se explotan para medir distintos valores que ayuden a hacer guías terapeúticas o, se combinan con otros datos clínicos, para detectar factores de riesgo de la enfermedad [173]. Además, desde esta disciplina se suelen recomendar artículos científicos que aporten evidencia sobre cómo debe afrontarse un diagnóstico o una enfermedad. Con estas características, esta disciplina se convierte en un dominio ideal para la aplicación de técnicas de IN [26]. En este caso se ha desarrollado una plataforma que permite hacer recomendaciones clínicas, planificar servicios sanitarios, o recomendar la consulta de evidencia científica a partir de la información de bases de datos epidemiológicas y de artículos científicos o guías clínicas.

- **Cuadernos de recogida de datos (CRD).** En este solución, se ha buscado un dominio donde la heterogeneidad fuera una característica común. En el ámbito de la investigación biomédica, los CRD son herramientas para la gestión de bases de datos clínicas que cubren los requisitos de un determinado proyecto de investigación. En este caso, se ha desarrollado una plataforma que permite generar cuestionarios semánticos, almacenarlos en un almacén semántico y explotarlos usando cuadros de mando, buscadores semánticos o generadores de alertas.

- **IN para fuentes de datos multimedia.** La IN en fuentes de datos no textuales es un campo a explorar. En este caso se han usado las

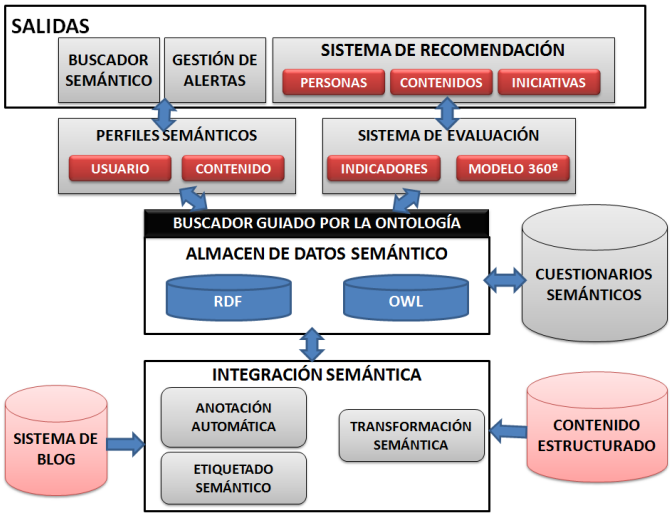


Figura 4.19: Arquitectura de la Red Social Semántica

herramientas de IN semántico descritas anteriormente para desarrollar una plataforma que permita clasificar semánticamente contenidos multimedia que, posteriormente, podrán ser comparados y analizados con herramientas como buscadores semánticos, recomendadores y cuadros de mando.

Estas soluciones software proporcionan ejemplos claros de uso de las plataformas de IN en dominios que pueden ser de interés para muchas compañías. A continuación se describen con más detalle cada una de las soluciones propuestas.

4.6.1 Red Social Semántica

En este apartado se presenta una plataforma que, usando inteligencia de negocio semántica, va a permitir la generación de entidades formales de conocimiento compartido entre usuarios que participan en una red social o en una intranet corporativa.

Como se puede apreciar en la figura 4.19, el sistema completo se sustenta en un almacén de datos semántico, en el que se almacenan los esquemas y datos del dominio. Este almacén se nutre de diversas fuentes de datos: (1) integra información de sistemas de datos basados en blogs y de fuentes de datos estructuradas, y (2) permite la definición de cuestionarios semánticos para generar información semántica directamente sobre el modelo. Sobre ese almacén se integran una serie de herramientas de IN que mejoran la gestión

de la información proporcionando servicios de valor añadido a los usuarios finales. A continuación se describen cada uno de los componentes de la arquitectura.

4.6.1.1 Almacén de datos semántico

En esta plataforma se usa la herramienta que combina representaciones del dominio en OWL y un repositorio de datos en RDF. En este caso en particular, el almacén de datos tiene cargado dos modelos concretos:

- FOAF [64]. Permite representar a los usuarios de la red social y sus relaciones.
- SIOC [65]. Permite representar los grupos sociales. Para adaptarlo a los requisitos especiales del sistema de evaluación, se ha extendido este modelo con:
  - La relación *sioc:has-favorite*, que expresa que un usuario o comentario se quiere etiquetar como favorito.
  - La propiedad *sioc:is-shared*, que indica que algunos contenidos son compartidos únicamente en un determinado grupo o que son privados.

A partir de aquí, se pueden añadir tantas ontologías como sean necesarias para describir el entorno de la red social que se quiere analizar y explotar.

4.6.1.2 Integración Semántica

El objetivo principal de este componente es integrar la información proveniente de redes sociales a partir de dos orígenes de datos concretos:

- Sistemas de texto abierto. La información que se puede encontrar en este tipo de sistemas son las entradas de un blog, comentarios, foros de discusión, documentos, etc. En este caso el proceso de ETL usará dos de las herramientas que describimos en 4.1.3.1:
  - Anotación automática. Esta herramienta permitirá anotar, usando técnicas de procesamiento del lenguaje natural, los elementos que encuentre en el texto de una forma totalmente transparente al usuario.

  - Etiquetado semántico. Esta herramienta permite generar consultas SPARQL a partir de la introducción de texto guiado por la ontología. Los resultados de esas consultas serán los que asignen semántica a ese contenido. Este proceso requiere un trabajo manual para el usuario, aunque puede realizar múltiples anotaciones a partir de la definición de una de estas etiquetas.

- Sistemas estructurados. La información que se puede encontrar en este tipo de sistemas son los datos identificativos de los usuarios, las valoraciones y, dependiendo del dominio, otra información estructurada que describa los recursos temáticos de la red. Se propone el uso de SWIT como herramienta para transformar los datos desde XML o bases de datos relacionales al modelo semántico. Se usa SWIT gracias a que la transformación se hace guiada por la ontología y por su flexibilidad a la hora de transformar diferentes fuentes de datos.

La comunicación entre los dos modelos de integración es importante, ya que se debe relacionar toda la información. Por ejemplo, si se están recuperando entradas de un blog, éstas pueden ser anotadas haciendo uso de las herramientas de anotación propuestas, a la vez que con SWIT se transforma la información de los usuarios, sus relaciones, las valoraciones de los comentarios, etc. Para conseguir esa interoperabilidad, se realiza una primera transformación almacenando la información en texto natural, que posteriormente es anotada antes de pasar al almacén de datos semántico.

4.6.1.3 Cuestionarios Semánticos

Esta herramienta permitirá completar la información de la red social directamente sobre el modelo semántico. El uso de esta herramienta no es obligatorio en este tipo de entorno, aunque podría ser muy útil para generar cuestionarios específicos sobre un ámbito concreto del dominio sobre el que versa la red social. Por ejemplo, en una red social de pacientes, podría ser interesante modelar cuestionarios sobre calidad de vida, estado de ánimo, etc.

Esta herramienta también puede ser útil para generar cuestionarios de evaluación que ayuden a puntuar semánticamente los recursos de la red social. Por ejemplo, en un entorno social de aprendizaje se podría usar este tipo de herramientas para generar cuestionarios de satisfacción de los miembros de un determinado grupo de aprendizaje.

4.6.1.4 Sistema de evaluación

Este módulo de la plataforma permite evaluar los recursos de la red social para poder generar un entorno de confianza basado en la puntuación que recibe cada recurso. Ese entorno de confianza será de gran utilidad como filtro para el sistema de recomendación, como se describirá a continuación.

En este caso se han integrado dos herramientas de evaluación:

- **Indicadores.** Este modelo consiste en la generación de indicadores concretos sobre el funcionamiento de la red social. Estos indicadores también pueden venir definidos por los cuestionarios semánticos. Un ejemplo del uso de este indicador podría ser una red social sobre cine en la que se puntúan las películas. La puntuación que dan los usuarios a la película podría ser un indicador de confianza, igual que lo pueden ser los premios que ha recibido, la crítica, etc.

- **Modelo 360º.** Este modelo consiste básicamente en que cada usuario puede evaluar a otros usuarios de la plataforma y los contenidos que ha publicado. Uno de los elementos principales de las redes sociales consiste en la posibilidad de crear entornos de confianza gracias a las valoraciones de los usuarios. Debido a que en una red social, además de las evaluaciones también es importante saber cuántas personas han evaluado, se usa la media bayesiana para calcular el indicador generado por la evaluación:

$$Media = (v/(v+m)) * R + (m/(v+m)) * C \tag{4.2}$$

  Donde R es la media aritmética de las evaluaciones que ha recibido el usuario o el contenido, v es el número de evaluaciones del recurso, m es el mínimo número de evaluaciones requeridas para tener un cierto nivel de confianza (será un parámetro de la plataforma que deberá ir ajustándose dependiendo del número de usuarios de la misma), y C es la media de todas las evaluaciones para todos los recursos.

  El modo de calcular los niveles de confianza permite balancear el peso y la influencia que cada contenido tiene respecto al resto. El primer término de la fórmula establece el peso dado a la propia media del elemento, mientras el segundo define el peso dado a todas las evaluaciones. Además, cuando un recurso recibe un número muy grande de evaluaciones, el primer término de la fórmula le dará mayor importancia. Inversamente, cuando un elemento tenga pocas evaluaciones, el segundo término de la fórmula disminuye el peso del primero.

4.6.1.5 Perfiles Semánticos

La interacción del usuario con la red social queda reflejado implícitamente en los contenidos que visita, los comentarios que hace, las entradas que publica, etc. Gracias al proceso de transformación semántico, esa interacción puede explotarse explícitamente. En esta solución, para poder identificar realmente cuáles podrían ser los intereses del usuario en función de esa interacción, se ha generado un perfil semántico a partir de aquellos elementos ontológicos más empleados en sus interacciones (comentarios, opiniones, cuestionarios semánticos, etc.). Este perfil semántico deberá calcularse con un límite para que luego el proceso de análisis de dichos perfiles sea lo más eficiente posible. Por ese motivo la plataforma tiene un parámetro de configuración que permite definir cuál será el tamaño máximo que tendrá dicho perfil para cada usuario.

Uno de los aspectos importantes del perfil semántico es que incluirá la valoración del usuario a partir del sistema de evaluación de él mismo y de los contenidos que ha publicado en la red. A su vez, gracias a la clasificación semántica que se hace de los contenidos en lenguaje natural en el proceso de transformación, también se generará un perfil semántico de cada contenido, incluyendo sus anotaciones y sus valoraciones en el caso de que las hubiese.

4.6.1.6 Sistema de recomendación

El recomendador está colocado en la cima de la arquitectura, lo que significa que depende de los resultados obtenidos en el resto de módulos. Concretamente, este módulo se alimenta de los siguientes resultados: (1) la transformación semántica de la información proporcionada por los usuarios, (2) el perfil semántico del usuario y los contenidos, y (3) la evaluación del usuario, que agrega la información de puntuación de los comentarios generados por él. En esta plataforma se han diferenciado tres tipos de recomendación:

- **Recomendación de contenidos.** El perfil del usuario modela conceptualmente los principales intereses de un usuario. A través de la recomendación de contenidos, la plataforma informa automáticamente a los usuarios de que hay un nuevo contenido que puede ser de su interés. Para hacer esto, la plataforma usa un modelo de correspondencia semántico entre los intereses del usuario y las anotaciones semánticas de los contenidos. La evaluación de los contenidos tiene un papel importante en esta herramienta, ya que prioriza en la recomendación aquellos contenidos que están valorados positivamente. Las métricas necesarias para calcular cómo se hará la función de similitud depende de cada problema concreto.

- **Recomendación de usuarios.** La clave fundamental de una plataforma social es la posibilidad de crear redes entre los recursos. Una red social es una estructura social construida sobre personas que están conectadas por algún tipo de propiedad [119]. Esta plataforma incluye un módulo de recomendación de usuarios que permitirá generar grupos sociales a partir de la compatibilidad entre los perfiles semánticos de cada uno de ellos. En este caso, al igual que en el anterior, también se tiene en cuenta la valoración de esos usuarios.

- **Recomendación de iniciativas.** Además de recomendar contenidos o recomendar usuarios, la plataforma permite recomendar otro tipo de recursos disponibles en la red social. Por ejemplo, en una red social sobre música, además de recomendar entradas de los blogs o miembros de la red con gustos musicales similares a los del usuario, se podrían recomendar canciones o álbumes que puedan ser de interés a éste y que a su vez tengan una buena evaluación. Para ello, se usan funciones de similitud semántica entre los perfiles semánticos del usuario y del recurso. La parametrización de esa función dependerá de cada dominio concreto.

4.6.1.7 Otros servicios

Gracias a la construcción del almacén semántico de la red social, los usuarios tendrán la posibilidad de usar buscadores semánticos avanzados y podrán configurar alertas sobre el funcionamiento de la plataforma (por ejemplo si un miembro de la red en el que está interesado ha publicado una nueva entrada, si hay un recurso disponible que pueda ser de mi interés, etc.)

4.6.2 Plataforma para la Planificación

La planificación estratégica es un ejercicio clave de las organizaciones. Uno de los análisis que se suele realizar antes de realizar una planificación estratégica es el análisis DAFO [174].

Las soluciones de IN se han identificado como herramientas de gran utilidad para hacer un plan estratégico, ya que recogen los datos de las fuentes de información y ofrecen una vista del estado actual de la empresa.

En este caso, se integran diferentes herramientas de la propuesta de IN semántica para que cualquier organización pueda identificar fortalezas y debilidades, no sólo de la actividad de la empresa sino también de los activos de conocimiento. Esta plataforma permitirá generar planes estratégicos adaptados a cada recurso de la organización y además, una vez generado un nuevo

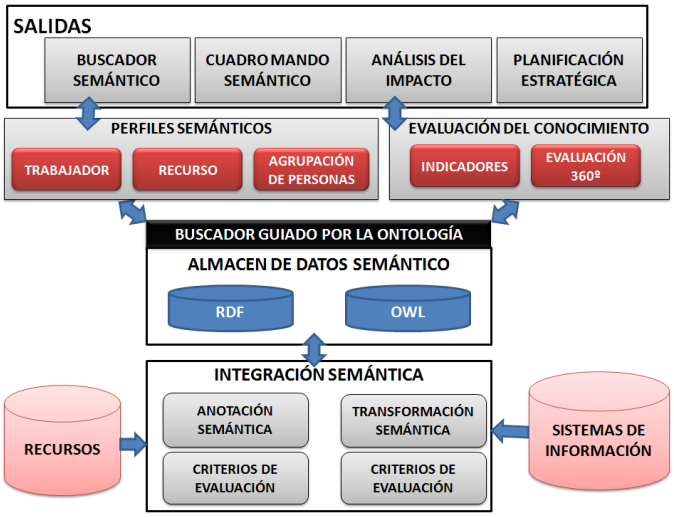


Figura 4.20: Arquitectura de la Planificación Estratégica

plan, será capaz de medir el impacto de éste en el conjunto de la entidad.

Como se puede apreciar en la figura 4.20, la plataforma se apoya en el almacén de datos semántico en el que se almacenarán las diferentes ontologías y la información del dominio. Este almacén se nutre de diversas fuentes de datos: (1) de los datos de los sistemas de información de la organización que permitirán generar indicadores de actividad, y (2) permite la definición de recursos que pueden representar oportunidades o amenazas para la organización.

4.6.2.1 Almacén de datos semántico

El almacén de datos semántico de esta solución estará compuesto por las siguientes ontologías:

- Catálogo de criterios de evaluación que permitan definir y medir los activos de conocimiento de la organización.
- Modelo ontológico de definición de indicadores y rangos.
- La ontología u ontologías necesarias para definir el dominio concreto.

4.6.2.2 Integración semántica

El modelo de integración dependerá del tipo de organización y de planificación que se quiera construir. En la sección 4.1 se describen los principales

modelos de integración presentados en esta tesis. En la sección 2.2.5.4 se discute sobre las diferentes situaciones de transformación semántica que se pueden presentar dependiendo del volumen y la heterogeneidad de la información.

4.6.2.3 Evaluación del conocimiento

Para la evaluación del conocimiento se han usado dos de los modelos que se explicaron en la sección 4.3: el modelo de evaluación 360º y el modelo de indicadores. Es importante destacar que en este caso de uso, ambos modelos están relacionados por el catálogo de criterios de evaluación.

Gracias al modelo de evaluación 360º la plataforma será capaz de detectar cuáles son las debilidades y fortalezas de los diferentes recursos. Este tipo de evaluación también permite medir los activos de conocimiento.

Por otro lado, el modelo de evaluación basado en indicadores permite obtener una instantánea del estado real de la organización. El hecho de que estos indicadores también se relacionen con los criterios de evaluación que los condicionan permite obtener directamente una correlación entre las fortalezas y debilidades, y el funcionamiento de la organización.

4.6.2.4 Perfiles semánticos

Se han generado tres tipos de perfiles semánticos:

- Perfil semántico tipo de un recurso. Puntuación mínima deseable de ese recurso para cada uno de sus criterios de evaluación que asegura que puede desempeñar correctamente su labor. Como los perfiles tipo pueden evolucionar en el tiempo, en este esquema también se almacena un histórico de las últimas N definiciones de perfil tipo, siendo N un parámetro de configuración de la plataforma.

- Perfil semántico de un recurso. Contenido de las últimas N valoraciones, donde N es un parámetro de configuración de la plataforma. Esa variable N será la misma que la del perfil tipo, lo que permitirá comparar cada evaluación o una agregación de las mismas con los perfiles deseados, detectando fácilmente qué criterios se han alcazado, cuáles son una fortaleza para el recurso y cuáles una debilidad.

- Perfil semántico de un grupo de recursos. Agrupación de definiciones de perfiles tipo y evaluaciones para crear una agregación mayor. Por ejemplo, gracias a este perfil, a partir de las evaluaciones de las personas de un departamento, se puede obtener cuál sería la puntuación total

de todo el departamento, detectando sus fortalezas y debilidades. A su vez, usando las evaluaciones agregadas de los departamentos se podría generar la evaluación global de toda la organización.

4.6.2.5 Sistema de planificación

En este caso se usa el sistema de planificación propuesto en la sección 4.5.5. Para el análisis interno de debilidades y fortalezas se han usado los perfiles semánticos que se han descrito en la sección anterior. Éstos permiten conocer, a partir de las comparaciones con el perfil tipo, cuáles son los recursos que deben incluirse en el plan estratégico.

4.6.2.6 Análisis del impacto

En esta herramienta se usa el modelo propuesto en la sección 4.5.6. Para ello, se define un modelo de evaluación basado en indicadores que se relacione con los criterios de evaluación usados en los perfiles semánticos. Este proceso dará resultados de cuál ha sido el impacto real en la organización de un plan estratégico concreto. Además, los resultados servirán para puntuar los activos que han tenido incidencia en el plan, pudiendo aumentar o disminuir su peso a partir de los resultados.

Esta herramienta también permitirá la detección de incoherencias en la anotación semántica de los recursos. Por ejemplo, si se ha definido un criterio de evaluación de "ofimática" y viendo su evolución se observa que mejora constantemente pero los indicadores reales de la empresa no, se generará una alerta al administrador para que sepa que probablemente el criterio de evaluación "ofimática" no está directamente relacionado con los resultados de esos indicadores.

4.6.2.7 Otros servicios

Gracias a la infraestructura semántica propuesta, los usuarios podrán usar buscadores semánticos sobre los planes, generar cuadros de mando semánticos personalizados y generar alertas e informes.

4.6.3 Plataforma para el Análisis Epidemiológico

En los últimos 20 años, la epidemiología se ha convertido en una displicina fundamental para la investigación y el tratamiento de enfermedades [173]. Hoy en día existen múltiples soluciones tecnológicas que son capaces de almacenar y analizar la información de pacientes con un determinado diagnóstico

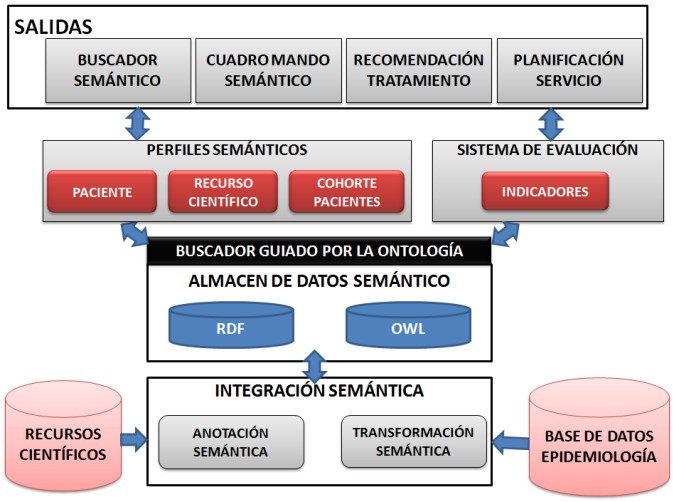


Figura 4.21: Arquitectura del Análisis Epidemiológico

[175]. Sin embargo, la ausencia de modelos comunes semánticos bien definidos es un problema cuando el usuario necesita definir análisis personalizados o enlazar con fuentes de datos externas [176].

Para solucionar este problema se ha diseñado un modelo semántico que facilite la explotación de bases de datos epidemiológicas y usando algunas herramientas ya descritas para aplicar métodos de IN que permitan explotar la información de forma dinámica y ágil.

Como se puede apreciar en la figura 4.21, el sistema completo se apoya en el almacén de datos semántico, en el que se encuentran las ontologías y la información del dominio epidemiológico. En este caso el almacén se alimenta de información proveniente de dos fuentes de datos: (1) los sistemas de historia clínica electrónica, que serán transformados a un formato semántico, y (2) los recursos científicos relacionados con el dominio.

4.6.3.1 Almacén de datos semántico

El modelo semántico se compone de los siguientes conceptos elementales (podrán ser ampliados dependiendo de la información concreta a explotar):

- **Paciente.** Se almacenan los datos pertenecientes a un paciente, como son sexo y fecha de nacimiento. Además, cada paciente tiene un histórico de las regiones geográficas donde ha vivido y durante cuánto tiempo. Cada paciente tiene una lista de diagnósticos y de terapias aplicadas a cada uno de ellos. Para todos se almacenan las fechas de diagnóstico y

de ejecución de la terapia. Por último, cada paciente tiene vinculadas las evoluciones de su enfermedad en cada visita médica.

- **Diagnóstico.** Se almacena el diagnóstico del paciente. Este concepto usa ICD 10 [177] para la clasificación del diagnóstico aunque puede ser ampliado para almacenar otros estándares.
- **Terapia.** Se almacenan las diferentes terapias disponibles. Cada terapia se relaciona con los diagnósticos para los que se puede usar. De cada terapia también se almacena el coste real de aplicación por visita.
- **Evolución.** Se almacenan los diferentes tipos de evolución que puede tener cada paciente para un determinado diagnóstico.

Este modelo puede ser extendido para cada problema concreto, pero las herramientas desarrolladas deberán disponer de esos conceptos para operar correctamente.

4.6.3.2 Integración semántica

Al igual que pasaba con la solución anterior, el modelo de integración dependerá del tipo de organización y el análisis que se quiera realizar.

4.6.3.3 Perfil semántico de un paciente

El perfil semántico de un paciente permite identificar grupos de pacientes que comparten algunas propiedades. Estos perfiles son de gran utilidad para comparar y estudiar estos grupos. En este trabajo, los perfiles han sido construidos usando entidades definidas en la ontología del dominio. En el caso de una base de datos epidemiológica se puede ver mejor la utilidad de la definición de perfiles semánticos. En un modelo epidemiológico típico, se encontrarían las fechas en las que se practica una determinada terapia a un paciente. Si se quisieran generar grupos basados en las tres primeras terapias aplicadas a un determinado diagnóstico, se tendría que ir recorriendo cada diagnóstico, ordenar las fechas, y filtrar todos aquellos que como mínimo tengan tres. Gracias a la definición de perfil semántico se podría añadir al modelo la información del orden de la terapia (no estaba incluido en el modelo original). Debido a ello se podrían comparar las terapias con orden 1, 2 y 3 sin la necesidad de usar las fechas en que se hicieron. Ni siquiera haría falta controlar el número de terapias, por lo que se optimizaría notablemente la eficiencia y la usabilidad de la plataforma para el usuario final.

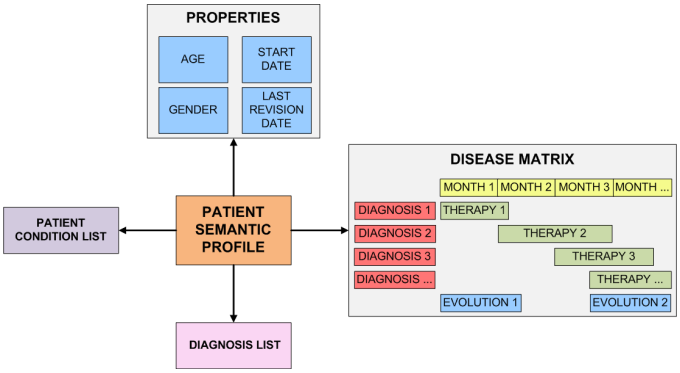


Figura 4.22: Perfil semántico de un paciente

El perfil semántico del paciente se ha definido como una serie de eventos que le han ocurrido. En concreto, diagnósticos, terapias y evolución de la enfermedad. Recuperando estos datos la plataforma es capaz de representar las terapias aplicadas por meses. En la figura 4.22 se puede ver como cada diagnóstico tiene asociado un cronograma con las terapias y evoluciones del paciente, ambas ordenadas por el mes concreto en el que se practicaron. Por ejemplo, para un paciente con cáncer de pulmón sew visualizaría un cronograma con las terapias aplicadas: cirugía, quimioterapia, etc., para cada periodo de tiempo. Además, gracias al perfil, el usuario podría identificar la evolución del paciente después de cada una de las terapias. De igual manera, en el perfil del paciente se almacena sexo, edad, condiciones sanitarias que tenía al ser registrado en la base de datos y un listado de otros diagnósticos.

4.6.3.4 Perfil semántico de una cohorte de pacientes

En este perfil se ha definido una representación en formato de cronograma tanto de terapias como de evoluciones para una cohorte de pacientes, es decir, para una serie de perfiles semánticos de pacientes que tienen propiedades en común. En este caso, para generar la cohorte es necesario que se defina uno o varios de los diagnósticos del paciente. Como se puede ver en la figura 4.23, el buscador guiado por la ontología será la herramienta que permita definir las propiedades de esos pacientes. Una vez que se ha realizado la selección, se usan sus perfiles semánticos para agregar la información de cada paciente.

Para generar una representación gráfica de este tipo de perfil, además de mostrar los criterios de inclusión, se elabora una matriz con las distribuciones de probabilidad de cada una de las terapias y evoluciones del paciente. Gracias a esa matriz de distribuciones, se puede aplicar uno de los algoritmos de

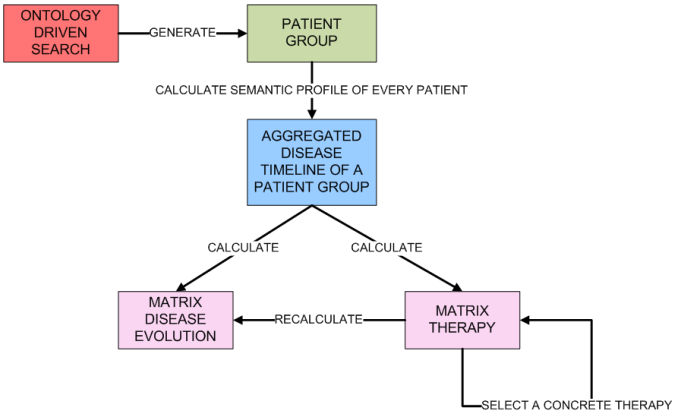


Figura 4.23: Perfil semántico de una cohorte de pacientes

recomendación propuesto basado en redes bayesianas. La entrada en la red se produce dinámicamente cuando el usuario elige una de las opciones, tanto de terapia como de evolución. En ese momento la matriz se vuelve a recalcular mostrando al médico cuál será la terapia y evolución más probables para el paciente y así sucesivamente.

Una vez generadas las cohortes de pacientes, éstas pueden ser usadas como herramienta de ayuda a la decisión, creándose automáticamente en la ontología como una nueva clase cuyos axiomas cumplen los definidos por los criterios de inclusión. Cada vez que un nuevo paciente fuese dado de alta en el sistema con esos criterios, la aplicación podría recomendar automáticamente a qué cohorte pertenece y cuáles serían sus terapias más probables o cuál podría ser la evolución de la enfermedad, basándose en el conocimiento disponible.

4.6.3.5 Recomendador de tratamiento

Este recomendador semántico está basado en redes bayesianas. El funcionamiento sería el siguiente:

1. El usuario selecciona una cohorte de pacientes con una serie de criterios haciendo uso de ODS.

2. La plataforma calcula el perfil semántico de cada uno de ellos y genera una representación gráfica de las distribuciones de probabilidad de cada terapia con su evolución durante el periodo de enfermedad. Debido a que las terapias y las evoluciones pueden repetirse a lo largo de la

enfermedad se usa una representación de árbol de redes bayesianas, que va ramificándose conforme se afina la búsqueda.

3. El usuario puede elegir qué terapia va a aplicar o cuál es la evolución que espera y la plataforma automáticamente reconstruye toda la representación gráfica para que el usuario pueda ver claramente cuál será la evolución más probable de ese paciente.

Esta herramienta puede ser usada como un asistente para conocer cuáles son las decisiones más comunes que se toman para un paciente con unas determinadas características o también puede ser usada como un sistema de ayuda a la formación de profesionales.

4.6.3.6 Planificación de los servicios clínicos

El esquema anterior puede usarse masivamente para la planificación de los servicios clínicos. Para ello, en vez de trabajar con una cohorte concreta de pacientes, se utilizan todos los perfiles, generando un grupo con todos los pacientes registrados en el sistema que lleven un determinado número de ciclos de tratamiento (ese número vendrá determinado por cada problema concreto, por lo que es un parámetro de configuración de la plataforma). La herramienta generará un árbol de redes bayesianas (al igual que en el caso anterior) y calculará el tipo, número y coste de terapias más probables que tendrá que aplicar durante un determinado periodo de tiempo.

Esta información ofrecerá una idea a los servicios médicos implicados de cuál será la actividad clínica más probable en los próximos meses o años.

4.6.3.7 Otros servicios de explotación

Gracias al modelo semántico, los usuarios de la plataforma podrán usar buscadores semánticos y generar cuadros de mando bajo demanda.

4.6.4 Cuaderno de Recogida de Datos (CRD)

Los cuadernos de recogida de datos (también llamados CRD) son una herramienta empleada para la monitorización de estudios de investigación clínica en los que se reclutan pacientes. Un CRD es un conjunto de formularios (electrónicos o en papel) que se van rellenando para cada paciente en cada una de las visitas o ciclos en los que se hace algún tipo de actividad clínica sobre éste. También pueden ser utilizados para la descripción de muestras biológicas que se recogen a esos pacientes, que luego pueden ser usadas en el ámbito asistencial o en investigación.

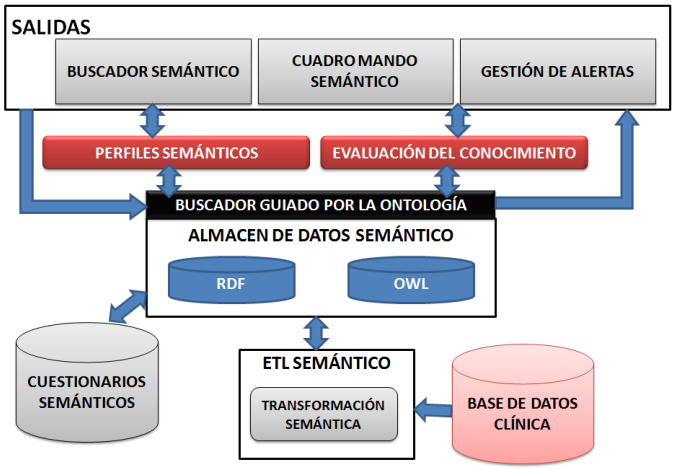


Figura 4.24: Arquitectura Cuaderno de Recogida de Datos

Una de las características más importantes de los CRD es su heterogeneidad, ya que cada estudio tiene su propio cuaderno. Para alcanzar una idea del volumen, en [178] aparecen las estadísticas a 31 de julio de 2015 del número de ensayos clínicos registrados en EudraCT [179], alcanzando la cifra de 44.621 ensayos realizados únicamente en Europa (desde que existe este registro, mayo de 2014). Esto quiere decir que en el último año se han puesto en marcha, como mínimo, 44.621 CRD en los estados miembros de la Unión Europea.

Otra de las características de un CRD es su necesidad de integración con sistemas asistenciales como la Historia Clínica Electrónica, ya que uno de los requisitos fundamentales de un CRD es que la información que gestiona esté alineada con la historia del paciente.

En la figura 4.24 se puede ver la arquitectura propuesta. En este caso se pueden ver los componentes principales que serán usados para que los usuarios puedan definir y explotar un CRD en un entorno totalmente integrado. Como se ha comentado, un CRD requiere integración con información clínica asistencial. Por ese motivo, se usan metodologías de transformación para incoporar esa información al almacén semántico. Además, se usan los "Cuestionarios Semánticos" para la definición de cada uno de los cuadernos que se rellenarán para cada paciente o muestra clínica. Opcionalmente, los gestores del CRD podrán definir perfiles semánticos y usar herramientas de evaluación de activos de conocimiento para generar representaciones semánticas que permitan optimizar la explotación, tanto de los cuadernos en sí (perfiles semánticos), como del proceso de recogida de los datos (evaluación

de activos de conocimiento). Por último, esta plataforma ofrecerá tres servicios: (1) buscador semántico de pacientes y muestras, (2) cuadros de mando personalizables y (3) gestión de alertas en el proceso de reclutamiento de pacientes.

4.6.4.1 Almacén de datos semántico

El almacén de datos semántico estará compuesto por una plantilla ontológica que modela semánticamente los conceptos necesarios del CRD entre los que se destacan:

- **Paciente (*Patient*).** Paciente reclutado en el estudio de investigación clínica.
- **Proyecto (*Project*).** Estudio de investigación clínica en el que se reclutarán pacientes, se definirán protocolos y se recogerán muestras.
- **Protocolo (*Protocol*).** Protocolo al que se somete al paciente reclutado.
- **Estadío (*Stage*).** Según el protocolo elegido, el paciente pasará por una serie de estadíos, fases o ciclos en su paso por el estudio clínico.
- **Donación (*Donation*).** Donaciones de muestras que hace un paciente en el marco de un proyecto clínico. En las instancias de esta clase se almacena el paciente que dona y las diferentes fechas de recogida y almacenamiento.
- **Muestra (*Sample*).** Muestras recogidas de cada paciente. Este concepto genera diferentes instancias a partir del procesamiento de cada muestra original, que genera una muestra destino. Por ejemplo, si la muestra original recogida en la donación era sangre, representará si de esa sangre se ha extraído plasma, ADN, capa blanca, etc.
- **Cuaderno (*Report*).** Representa los diferentes cuestionarios que se han modelado para recoger los datos. Un cuaderno se puede rellenar cuando el recurso clínico (paciente o muestra) está en un protocolo concreto en un estadío concreto. "Cuaderno" tiene dos subclases para reflejar qué cuadernos son de pacientes y cuáles de muestras.
- **Taxonomía (*Taxonomy*).** Como ya se ha comentado cuando se describieron los "Cuestionarios Semánticos", algunas de sus preguntas pueden tener como respuesta una taxonomía de respuestas definida explícitamente para un caso concreto o reusada de otras ontologías o de otros

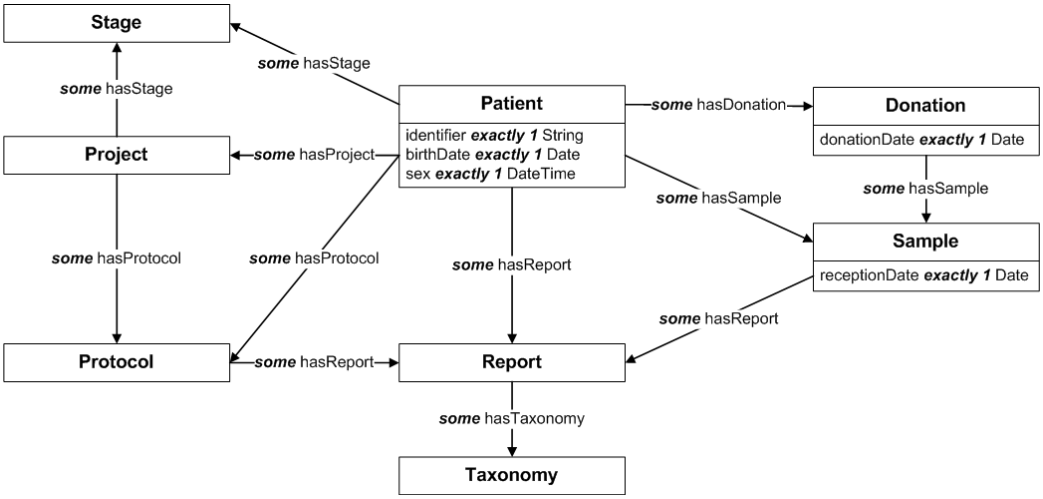


Figura 4.25: Extracto de la ontología del CRD

cuestionarios semánticos. Este concepto modela la raíz de todas esas taxonomías. Esto quiere decir que si se quiere definir que un campo de un cuestionario será rellenado con un concepto de otra ontología, ésta debe ser importarda y su concepto relacionarse como subclase de Taxonomía.

En la figura 4.25 se puede ver un extracto del modelo ontológico que se sigue para el desarrollo del CRD semántico.

4.6.4.2 Integración semántica

El objetivo principal de este componente es integrar los datos provenientes de los sistemas de información sanitarios con las variables definidas en los cuestionarios semánticos que se vayan a definir. En el caso concreto del CRD, únicamente se integra información de sistemas estructurados. Se propone el uso de SWIT como herramienta para la transformación de este tipo de información gracias a que la integración puede hacerse guiada por la ontología y por su flexibilidad a la hora de transformar datos de orígenes diferentes (XML y relacional).

4.6.4.3 Cuestionarios semánticos

Los cuestionarios semánticos serán la herramienta principal del CRD. Estos cuestionarios serán modelados siguiendo los principios expuestos en la sección 4.2. A la hora de generar el modelo de explotación, cada máquina de procesos

definida se convertirá en un *Protocolo*, y cada uno de los estados se mapea como un *Estadío*.

Esta herramienta ofrece una gran flexibilidad, ya que cualquier investigador puede crear sus cuadernos de recogida de datos aplicados al reclutamiento de pacientes y a la recogida de muestras biológicas. Usar un modelo semántico para almacenar y explotar la información facilita la reutilización, comparación y entrelazado de los datos. Además, permite que se pueda configurar cualquier tipo de CRD sin la necesidad de un equipo de expertos en TIC.

4.6.4.4 Perfiles semánticos

En esta plataforma los perfiles semánticos son de gran utilidad para la definición de representaciones semánticas que permitan explotar más fácilmente la información que realmente interesa. Además, podrán ser útiles para generar grupos de pacientes o muestras que tengan cierto grado de similitud en función de sus perfiles semánticos.

4.6.4.5 Evaluación de activos de conocimiento

En este caso, la evaluación de activos de conocimiento consistirá en la definición de criterios de calidad durante el proceso de reclutamiento. Como ya se ha comentado, un CRD, además de almacenar cuestionarios clínicos, también almacena el proceso a través del cuál se van a almacenar, es decir, las visitas, ciclos o estadíos del paciente en los que se recogerá la información. Esos procesos se pueden asociar a criterios de evaluación como, por ejemplo, calidad de las muestras recogidas, formularios de satisfacción del paciente, etc.

Esta evaluación va a dar información, no sobre los datos recogidos del paciente, sino sobre cómo está funcionando el proceso de reclutamiento y seguimiento de cada individuo.

4.6.4.6 Servicios de explotación

Como se ha comentado, la plataforma permite la explotación de todos los datos generados. Para ello se hace uso de ODS como herramienta para la definición de consultas SPARQL. Los servicios disponibles en esta plataforma son:

- **Buscador semántico.** Este servicio permite que cualquier usuario realice búsquedas de pacientes, muestras y donaciones de muestras por cualquiera de las propiedades que se hayan generado en el cuaderno

de recogida de datos, o que se hayan recuperado de las bases de datos clínicas.

- **Cuadros de mando semánticos.** Este servicio permite que cualquier gestor de datos del proyecto pueda configurarse su propio cuadro de mando bajo demanda usando todo el conjunto de propiedades recogidas en los diferentes formularios semánticos.

- **Gestión de alertas.** Este servicio permite que cualquier gestor pueda definirse alertas de funcionamiento del proceso de reclutamiento. Por ejemplo, si se está reclutando a mujeres embarazadas y se sabe que hay que recoger una muestra en la semana 20 de embarazo y otra en la 32, se puede configurar una alerta para que a partir de la primera visita avise automáticamente de que hay que hacer una segunda 12 semanas después.

4.6.5 IN Semántica en Contenidos Multimedia

La indexación de contenidos multimedia es una tarea muy importante en dominios donde parte del conocimiento se almacena en este tipo de formato. En la mayoría de soluciones de almacenamiento de imágenes como Flickr, Instagram, etc., las imágenes se describen con campos de texto basados en título, descripción y palabras clave. Esta anotación facilita la clasificación y la búsqueda del contenido, pero la ausencia de representaciones formales de esas anotaciones dificulta que los contenidos puedan ser comparados entre sí o enlazados con otros. La Web Semántica está considerada por la comunidad científica como una herramienta útil para organizar y explotar este tipo de contenido [180].

Multitud de proyectos de investigación del ámbito biosanitario usan imágenes, por ejemplo, para proyectos relacionados con radiodiagnóstico, imagen de microscopio, etc. Para clasificar estas imágenes, los investigadores suelen hacer uso de terminologías clínicas, pero no de representaciones formales basadas en ontologías en OWL.

En esta solución se propone el desarrollo de una plataforma que clasifique semánticamente imágenes biosanitarias. Esta clasificación permite que se puedan comparar imágenes entre sí, incluso recomendar anotaciones para que este proceso sea automático en el caso de que las imágenes sean similares. Además, se pueden usar los cuadros de mando y el buscador semántico para analizar, comparar y localizar las imágenes.

En la figura 4.26 se puede ver la arquitectura propuesta. En este caso, al igual que en los anteriores, la plataforma se sustenta en el almacén de

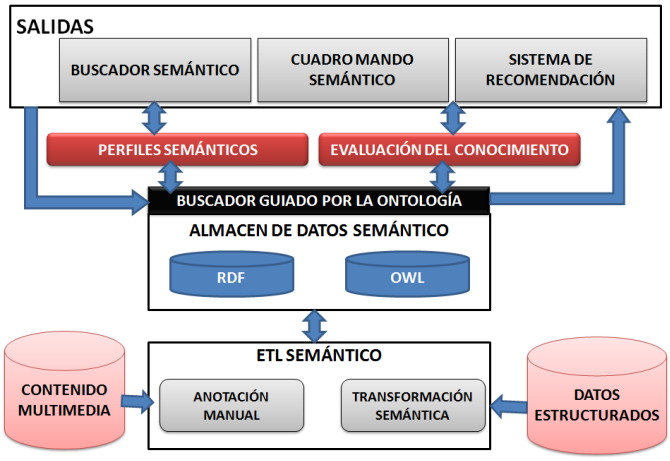


Figura 4.26: Arquitectura para la IN Semántica en contenidos multimedia

datos semántico. Este almacén estará nutrido por dos fuentes de datos: (1) los contenidos multimedia anotados manualmente, y (2) datos estructurados que sirvan para hacer dicha anotación.

4.6.5.1 Sistema de anotación

Se hace uso del modelo de anotación manual. El usuario tendrá la opción de elegir sobre qué ontología desea anotar o usarlas todas de forma combinada. En este modelo no se anotan las imágenes directamente, sino que se anotan recursos de investigación descritos por esas imágenes. Por ejemplo, la anotación de un diagnóstico a partir de una serie de imágenes no se realizaría sobre las imágenes sino sobre el paciente.

Este modelo soporta dos tipos de anotación:

- **Anotación positiva.** Esta anotación refleja que el recurso, a través de su descripción en imágenes, es anotado con un determinado elemento ontológico.
- **Anotación negativa.** En este caso el usuario espera que un determinado recurso tenga una anotación concreta, pero al revisar las imágenes no es así. Este tipo de anotación indicaría que en ese recurso no se está representando gráficamente una determinada propiedad que suele ser común, es decir, hay una excepción. También puede representar que se ha estudiado si la imagen representaba cierta caractetística y que no es así.

4.6.5.2 Buscador semántico

El buscador semántico consta de dos perspectivas de recuperación de datos, que pueden ser combinadas según el criterio del usuario de la herramienta:

- Buscador guiado por la ontología del dominio. Este buscador basa su funcionamiento en ODS. El usuario puede elegir qué concepto desea buscar y usar las anotaciones de las imágenes que los describen para crear filtros más avanzados.

- Buscador de texto abierto con clasificación de resultados. En este caso se usa una herramienta más cómoda para el usuario. El formato de esta búsqueda es parecido al de la mayoría de buscadores, es decir, hay un campo de texto y conforme el usuario va escribiendo se le recomienda cuál puede ser el texto que desea buscar a partir de la información textual almacenada en el repositorio semántico. Antes de poder buscar, el usuario debe indicar qué conceptos ontológicos está buscando, es decir, qué tipo de instancias quiere que le sean devueltas. Una vez que el sistema encuentra los resultados, los devuelve clasificados por esos conceptos y sus jerarquías.

Los dos buscadores se pueden combinar entre sí, dando lugar a potentes herramientas de búsqueda. Por ejemplo, si se quisiese buscar muestras de tejidos en las que se ve la expresión de un determinado gen, se podría usar el buscador guiado por la ontología para buscar aquellas muestras que tienen esa anotación. Además, se podría emplear el buscador de texto abierto para buscar algún texto que describa la muestra, por ejemplo, es un tejido cerebral. En este caso, el sistema devolverá los resultados de todas aquellas muestras que expresen ese gen y que en algún dato aparezca "tejido cerebral". Además, devolverá los resultados clasificados por los diferentes tipos de muestra que ha encontrado.

4.6.5.3 Recomendador de anotaciones

A partir de la información que se almacena en el motor de anotación, se ha usado la herramienta de recomendación para reusar parte o todas las anotaciones de una imagen. Esta herramienta permite establecer cuál es el número de anotaciones, equivalentes con otras imágenes, que el usuario debe insertar para que el motor de recomendación actúe. Una vez que el usuario que tiene el rol de anotador ha introducido ese número o más anotaciones la herramienta automáticamente le indica que hay imágenes que son similares y le permite visualizarlas para reutilizar sus anotaciones.

Para facilitar la reutilización de las anotaciones, se delimitan, a través de relaciones jerárquicas, cuáles son las ramas de la ontología equivalentes de tal forma que el usuario puede seleccionar todas las ramas de la ontología o aquellas que considere oportuno.

4.6.5.4 Perfiles semánticos

La definición de perfiles semánticos se realizará sobre las anotaciones que reciben los contenidos multimedia que describen uno o varios recursos de la organización. Los perfiles serán definidos en base a las propias ontologías del dominio que se hayan usado para anotar, y a través de dichas anotaciones podrán relacionarse con uno o varios contenidos.

4.6.5.5 Evaluación de similitud

En esta solución se usa la información semántica almacenada para evaluar cuál es la similitud entre dos o más imágenes. Para ello, el usuario elige una imagen de referencia con la que se compararán las demás. Después selecciona tantas imágenes a comparar como desee. Usando las funciones de similitud semántica se generan porcentajes de similitud entre imágenes. La función de equivalencia puede personalizarse indicando el número de elementos anotados en una jeraraquía que se consideran necesarios para que la imagen sea equivalente.

Esta herramienta, además de permitir que el usuario compare una o varias imágenes, ofrece la posibilidad de generar grupos de éstas con un determinado grado de similitud (este umbral también será un parámetro configurable de la herramienta). Esto permitirá generar grupos de imágenes que describan una serie de propiedades concretas que siempre se encuentran en ese tipo de contenido.

4.6.5.6 Otros servicios

Gracias al modelo semántico, los usuarios tendrán la posibilidad de usar cuadros de mando personalizables, permitiendo representar la información desde diferentes perspectivas.

Capítulo 5

Validación

En este capítulo se presentan seis casos de uso que han servido para validar las soluciones propuestas en el capítulo anterior:

- Red social semántica. La validación de este caso ha consistido en la puesta en marcha de un prototipo de red social en el dominio financiero.
- Plataforma para la planificación. En este caso se ha evaluado el desempeño del personal de un hospital y se ha planificado, en función de esa evaluación, cuáles son sus necesidades formativas.
- Plataforma para el análisis epidemiológico. La validación de esta solución se ha realizado sobre dos orígenes de datos: (1) un registro de cáncer y (2) un programa de cribado de cáncer de colon y recto.
- Cuaderno de recogida de datos. Esta solución se ha implantado para gestionar el proceso de reclutamiento del proyecto NELA [181].
- IN semántica en contenidos multimedia. Esta solución se ha implantado para anotar la expresión génica de muestras capturadas a través de un microscopio digital. Concretamente ha ayudado a gestionar el "*work package*" 6 del proyecto europeo EUCOMM-Tools [182]

A continuación se van a explicar detalladamente cada uno de los casos de validación.

5.1 SocialBROKER: Red social semántica en el ámbito financiero

5.1.1 Introducción

El dominio financiero se ha convertido en un gran entorno de conocimiento, con un gran número de compañías involucradas y con un impacto tremendo en la sociedad. Por lo tanto, existe una necesidad de establecer estrategias más precisas y potentes para la gestión de información de carácter financiero. Además, en el contexto actual de crisis financiera, este dominio se puede adaptar perfectamente a plataformas de carácter social que permitan a los usuarios beneficiarse de iniciativas generadas por expertos. En este caso de uso se ha desarrollado una ontología en el ámbito del dominio financiero. Además, se han usado diferentes módulos del marco de trabajo de IN semántica para construir una plataforma social donde los usuarios puedan gestionar su cartera financiera y compartir sus iniciativas.

5.1.2 Metodología y Herramientas

El desarrollo de este prototipo se basa en la solución de Red Social Semántica propuesta en la sección 4.6.1. En la figura 5.1 se puede ver la arquitectura real del prototipo. En este caso, se tiene una única ontología que modela el dominio financiero, y que se alimenta de la información de diferentes portales Web de índices bursátiles. En el ámbito del proceso de ETL, no se hace uso de ninguna de las propuestas de integración que se han comentado, sino que se usa un poblador de ontologías basado en los resultados de [183].

Otra de las fuentes de datos es un blog que permite la publicación de noticias económicas y financieras, y comentarios por los lectores registrados. Todos estos contenidos son clasificados semánticamente haciendo uso de dos de los modelos: anotador automático con técnicas de procesamiento del lenguaje natural y el etiquetado semántico de contenidos.

Por último, existe otra fuente de datos que viene de los cuestionarios semánticos. Para este origen de datos se ha usado para la definición de carteras financieras. Estas carteras se almacenarán clasificadas con las empresas, divisas o índices bursátiles de la propia ontología del dominio, por lo que el modelo de datos está totalmente integrado.

En la capa de evaluación del conocimiento, el modelo define varias técnicas. En primer lugar, todos los usuarios podrán valorar del 1 al 10 la calidad de las entradas de los autores. Para ello se ha usado el modelo de evaluación 360º, en el que habrá una convocatoria de evaluación permanente en la que

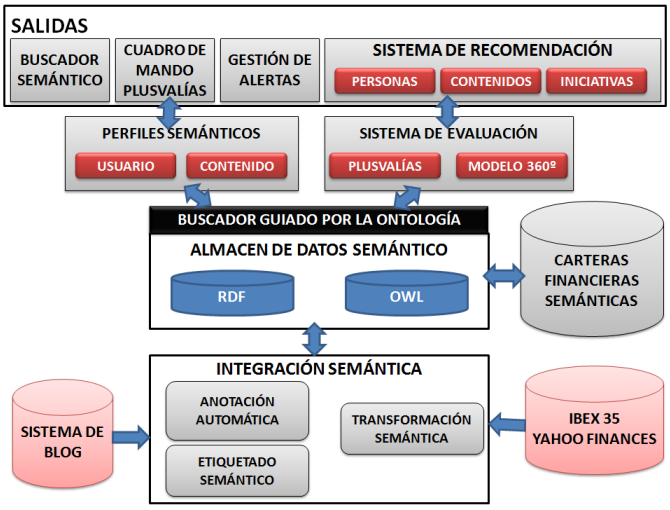


Figura 5.1: Arquitectura de SocialBROKER

todos los usuarios podrán evaluar el contenido. Otro criterio a destacar de los inversores es cómo está funcionando su cartera financiera. Gracias a que se almacenan los datos reales de los índices bursátiles y a la gestión de las carteras financieras, la plataforma es capaz de generar un perfil semántico que indique la evolución de las inversiones de cada usuario. Estas dos perspectivas de evaluación servirán para la creación de un entorno de confianza a la hora de hacer recomendaciones.

En esta solución se han definido tres tipos de perfiles semánticos:

- Perfil de una entrada del blog. Tiene dos propiedades: (1) las anotaciones realizadas por el anotador automático y por la respuesta de las consultas SPARQL definidas en el etiquetado semántico, y (2) las evaluaciones recibidas por el resto de usuarios de la plataforma. En el primer caso se identifican cuáles son los conceptos clave de la entrada y en el segundo si lo que dice es de interés o no.

- Perfil del usuario. Consiste en las n anotaciones más repetidas en sus entradas o comentarios publicados en la red social, donde n es un parámetro configurable por los administradores de la plataforma. Además, este perfil tiene la media de las puntuaciones recibidas para cada publicación que haya realizado.

- Perfil del inversor. Toma como base el perfil del usuario, al que añade la información sobre su cartera financiera y la plusvalía generada año a

año. Sólo se tiene en cuenta la evolución de su cartera durante el último año.

En la capa de explotación se destacan los siguientes servicios:

- Generación de cuadros de mando semánticos. Cada usuario puede ver cómo evoluciona su cartera financiera, incluyendo las plusvalías de cada inversión.
- Sistema de recomendación para el asesoramiento en inversiones a partir de la información de los perfiles semánticos de los inversores. Este sistema, básicamente, cruza los resultados del perfil semántico del inversor que está usando la plataforma con los perfiles de los miembros de la red que más dinero han ganado y que mejor valorados están por otros usuarios, de tal forma que si en el perfil del usuario actual se tiene el concepto "Telecomunicaciones" y hay un inversor que está ganando dinero en empresas de telefonía, se recomendará que se invierta en ella, ya que está dentro de sus intereses y actualmente está generando beneficios.
- Sistema de alertas. Notifica al usuario que hay una nueva entrada que podría ser de su interés basándose en comparaciones con los perfiles semánticos de cada uno. En este servicio se comparan los perfiles del usuario con el de la entrada del blog y si está bien valorada se le envía un correo electrónico al usuario informándole de que esa entrada podría ser de su interés.
- Sistema de recomendador de seguimiento de inversores. Es una extensión del sistema de alertas. La aplicación recomienda entradas o directamente hacer un seguimiento de las entradas de inversores basado en la similitud semántica entre sus respectivos perfiles.
- Todos los usuarios disponen del buscador semántico guiado por la ontología para realizar búsquedas avanzadas sobre las diferentes entradas y comentarios.

5.1.3 Resultados y evaluación

Los principales resultados de este caso de uso han sido la creación de una ontología que modela el dominio financiero, el desarrollo de un prototipo basado en la "Red social semántica", un estudio de rendimiento y una comparativa con otras plataformas similares analizando sus ventajas e inconvenientes.

Figura 5.2: Extracto de la ontología financiera

5.1.3.1 Ontología financiera

En los últimos años se han desarrollado múltiples ontologías en el ámbito financiero. Para este caso de uso, se ha desarrollado una nueva ontología basada en las existentes. La ontología[1] contiene un total de 123 clases, 86 subclases, 72 propiedades, 16 relaciones, 87 axiomas de restricción y ha sido definida usando OWL-DL (vea figura 5.2). La ontología cubre cuatro conceptos financieros principales:

- El concepto *Mercado Financiero* (*Financial_ markets*) representa los mercados que permiten comprar y vender activos financieros como divisas, acciones, valores, etc. Se han definido diferentes tipos de mercado como: "Mercados de renta variable" (*Variable_ income_ market*), "Mercados de divisas" (*Currency_ market*), "Mercados derivados" (*Derivative_ market*) y "Mercados de renta fija" (*Fixed_ income_ market*).

- El concepto *Intermediario Financiero* (*Financial_ intermediary*) representa, entre otras cosas, las entidades que suelen invertir en los mercados financieros. Por ejemplo: bancos, compañías de seguros, brokers, etc.

[1] http://sele.inf.um.es:9080/finanzas.owl

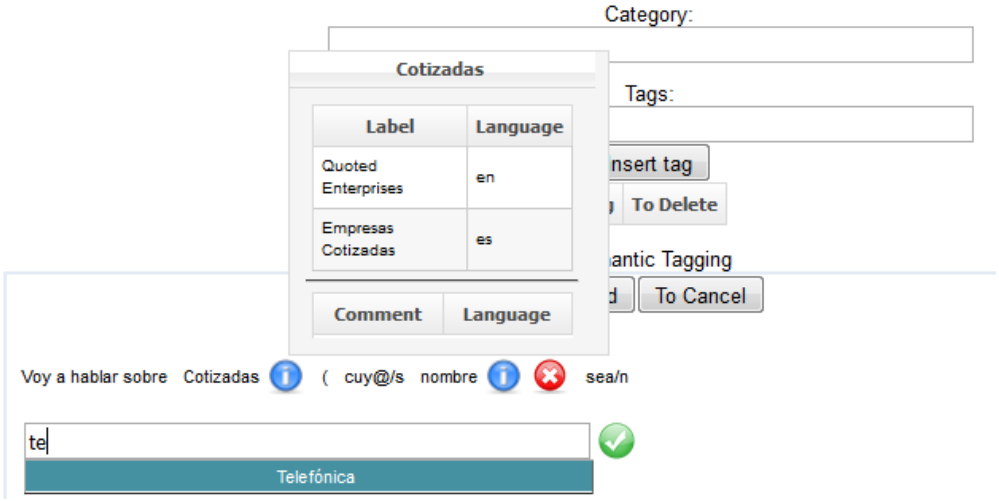


Figura 5.3: Captura de pantalla del etiquetador semántico

- El concepto *Activo* (*Assets*) representa todo elemento en el que se puede invertir (empresas como General Electric, Microsoft pertenecerían al concepto "Compañía" (*Company*), "dólar americano" o "euro" serían subclases del concepto "Divisa" (*Currency*).

- El concepto *Legislación* (*Legislation*) representa a aquellas organizaciones encargadas de supervisar un mercado financiero (banco central europeo, reserva feredal, etc), y las regulaciones y leyes que pueden ser aplicadas a este dominio.

5.1.3.2 Prototipo funcional

A continuación se describen las principales herramientas de la plataforma[2]:

- Herramienta de etiquetado semántico. En la figura 5.3 se puede ver cómo un usuario puede etiquetar un contenido a partir de la generación de consultas SPARQL guiadas por la ontología. En este caso se etiquetan sobre empresas cotizadas cuyo nombre empiece por "Te". Al situarse sobre el icono de la *I* aparecen las etiquetas de la clase. Este modelo permite, además de etiquetar con el modelo semántico, generar un gran número de anotaciones de términos que puedan estar relacionados.

- Recomendador de inversores e iniciativas. En la figura 5.4 se puede ver una imagen en la que se aprecia cómo el sistema es capaz de recomen-

[2]http://sele.inf.um.es:9080/SocialBROKER/

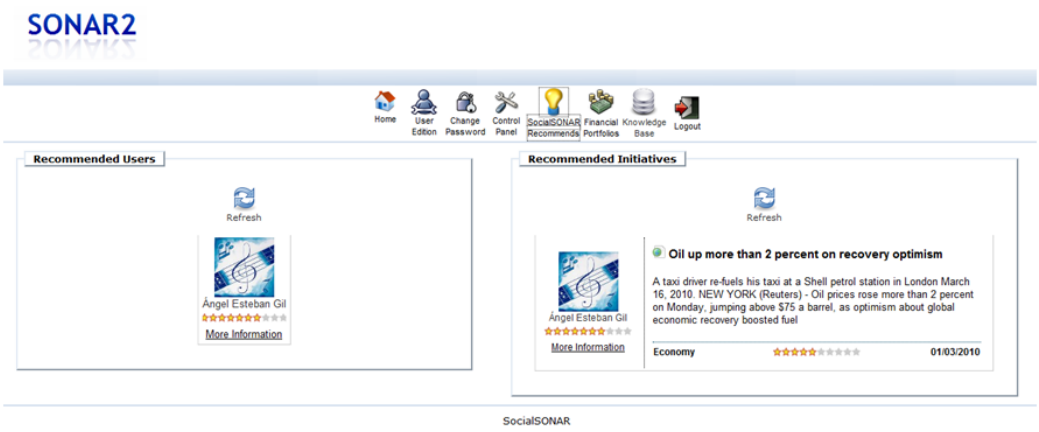


Figura 5.4: Captura de pantalla del recomendador semántico

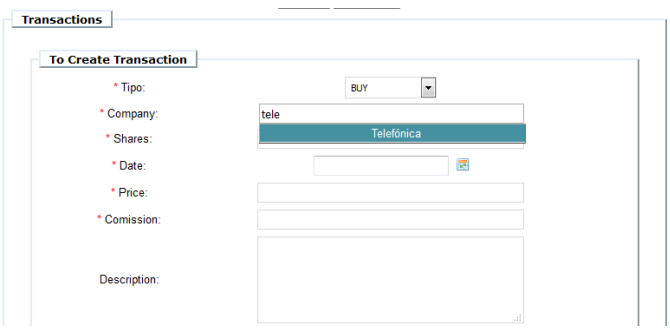


Figura 5.5: Captura de pantalla de la cartera financiera

dar usuarios e iniciativas que estén bien valorados. Esta es una de las aportaciones principales de este modelo.

- Carteras financieras. Se definen mediante los cuestionarios semánticos. Como se puede ver en la figura 5.5, el usuario define su cartera financiera buscando empresas cotizadas que empiezan por "tele".

5.1.3.3 Evaluación

Los resultados de esta plataforma han sido evaluados siguiendo cuatro parámetros: consumo de recursos, niveles de rendimiento, comparación con otros servicios financieros como Google Finances[3] o Yahoo Finances[4], y la compa-

[3]http://www.google.com/finance
[4]http://finance.yahoo.com/

ración con otras redes sociales con otras temáticas como IMDB[5].

Para realizar las dos primeras evaluaciones se ha usado un servidor con 4GB de RAM y un procesador Intel Pentium Dual T3400 con 2.16 GHz y un sistema operativo Linux de 64 bits. Las tecnologías concretas para desplegar el prototipo han sido: Apache Tomcat 7.0.24 como servidor de aplicaciones, Virtuoso [184] como servidor RDF y la OWL API [185] como motor de acceso a OWL.

En el parámetro de **consumo de recursos** se han configurado diferentes pruebas para comprobar la capacidad de memoria necesaria para que funcione correctamente dependiendo del número de tripletas y del número de usuarios concurrentes en la plataforma. Los valores de carga de memoria para las variables anteriores se muestran en la tabla 5.1.

Tabla 5.1: Nª de usuarios concurrentes - Nº de tripletas RDF - Uso de memoria

Nª de usuarios concurrentes	**Nº de tripletas RDF**	**Uso de memoria (en MB)**
10	25.000	152
10	100.000	228
10	250.000	570
10	1.000.000	2.280
25	25.000	227
25	100.000	398
25	250.000	742
25	1.000.000	2.459
50	25.000	352
50	100.000	523
50	250.000	865
50	1.000.000	2.575
100	25.000	603
100	100.000	774
100	250.000	1.116
100	1.000.000	2.826

De estos resultados se puede obtener la siguiente información sobre el comportamiento de la plataforma (véase la figura 5.6):

- El crecimiento de la memoria es lineal con el incremento de usuarios concurrentes. Aproximadamente, cada sesión de la plataforma tiene un máximo de 5MB por conexión, cuando el usuario está anotando contenidos (uno de los momentos de mayor carga en memoria).

[5]http://www.imdb.com/

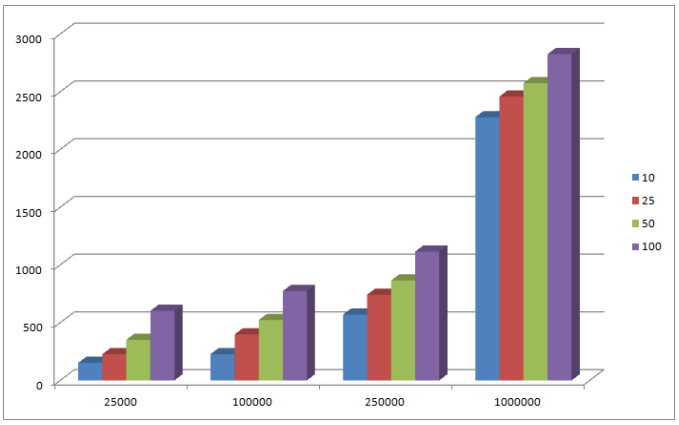


Figura 5.6: Gráfica Nº de usuarios concurrentes - Nº de tripletas RDF - Uso de memoria

- Los valores más altos de consumo de memoria son producidos por el anotador automático, ya que el método basado en GATE carga todas las etiquetas del almacén de datos semántico en memoria.

Con este modelo de datos y con la máquina empleada habría un límite real para gestionar más de un millón de tripletas.

Para el parámetro de **niveles de rendimiento** se han realizado otras pruebas consistentes en medir el tiempo necesario tanto para indexar los contenidos como para generar los perfiles semánticos. En la plataforma existen cuatro tipos de clasificación de la información:

- La clasificación semántica de las carteras financieras.
- La clasificación semántica de entradas del blog basada en técnicas de procesamiento del lenguaje natural que, en este caso, usan los Gazeteer de GATE.
- La clasificación semántica de las entradas usando los mecanismos de etiquetado semántico con ODS.
- El cálculo de los perfiles semánticos de cada usuario a partir de su cartera financiera y de los contenidos que genera en la red. Este cálculo también incluye los niveles de confianza generados a través de la puntuación de otros usuarios y de los beneficios de su cartera.

El tiempo empleado para cada uno de los tipos de clasificación anterior puede verse en la tabla 5.2 y en la figura 5.7 para diferentes números de

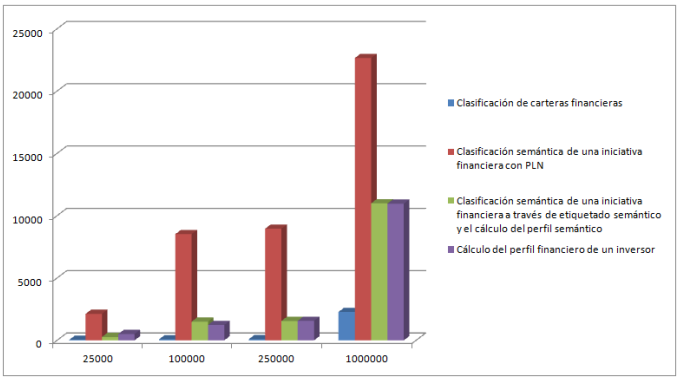


Figura 5.7: Gráfica del tiempo consumido en tareas de clasificación semántica

tripletas en el almacén. De estos resultados se pueden obtener las siguientes conclusiones sobre el rendimiento de la plataforma:

Tabla 5.2: Tiempo consumido en tareas de clasificación semántica

Tipo de clasificación	**Nº de tripletas RDF**	**Tiempo (en ms)**
Clasificación de carteras financieras	25.000	45
	100.000	81
	250.000	90
	1.000.000	2.280
Clasificación de entradas con PLN	25.000	2.125
	100.000	8.523
	250.000	8.956
	1.000.000	22.652
Etiquetado semántico de entradas	25.000	269
	100.000	1.502
	250.000	1.556
	1.000.000	10.996
Perfil semántico de un inversor	25.000	506
	100.000	1.236
	250.000	1.564
	1.000.000	10.965

- La clasificación semántica de carteras financieras es muy rápida en todos los escenarios.
- La clasificación usando GATE es el proceso más lento, sobre todo cuando hay un alto número de tripletas en el almacén semántico.

- El proceso de etiquetado semántico es más lento conforme el almacén es más grande, debido a que la ejecución de las consultas SPARQL es más lenta. Debido a que los perfiles semánticos también se calculan con consultas SPARQL, su rendimiento también decrece con el crecimiento de los datos en un orden polinomial.

Con esta información se puede concluir que, con la capacidad del hardware para hacer las pruebas, la plataforma se comporta correctamente sobre el millón de tripletas. En el caso de tener fuentes de datos mayores habría que ampliar el hardware, o recalcular algunas tareas asíncronamente, como son los procesos de etiquetado semántico o de cálculo de perfiles.

La comparación con otras redes sociales de gestión de carteras financieras como Google Finances o Yahoo Finances ha deparado las siguientes ventajas y desventajas:

- Ventajas:
  - SocialBROKER permite la creación de entornos de confianza entre los usuarios gracias a los sistemas de evaluación de activos de conocimiento (usuarios y contenidos de la red).
  - SocialBROKER genera perfiles semánticos con los intereses de los usuarios. Estos perfiles permiten la generación de modelos de recomendación de contenidos o seguimiento de inversores.
  - SocialBROKER es capaz de recomendar iniciativas financieras a partir del perfil semántico de los inversores.
- Desventajas:
  - Las otras plataformas ofrecen herramientas preconfiguradas para la explotación de carteras financieras. Además, disponen de más fuentes de datos, lo que les permite dar más información económica a sus usuarios.
  - Tanto Google como Yahoo recuperan noticias financieras de otros portales, los clasifican y son capaces de hacer recomendaciones a los usuarios.
  - Las otras plataformas ofrecen otros servicios no financieros con los que la plataforma no puede competir.

A continuación se detallan las principales similitudes y diferencias de SocialBROKER con otra semejante en otro ámbito como es IMDB.

- Similitudes:
  - Ambas plataformas crean entornos de confianza a partir de las evaluaciones de los usuarios, es decir, confían en hacer recomendaciones de contenidos que estén bien valorados por los usuarios.
  - Ambas plataformas definen el concepto de perfil de usuario, aunque en SocialBROKER ese perfil se define automáticamente a partir de la interacción del usuario con la plataforma.
  - Ambas plataformas recomiendan contenidos en base al perfil del usuario.
- Diferencias:
  - SocialBROKER realiza una clasificación semántica de todos los contenidos, facilitando la gestión de los mismos.
  - SocialBROKER recomienda a otros usuarios por la similitud entre sus perfiles, teniendo en cuenta los niveles de confianza.
  - IMDB publica sus propios contenidos clasificándose correctamente en el origen. La clasificación de esos contenidos no es semántica.

5.2 Planificación semántica de la formación continua de un hospital

5.2.1 Introducción

La formación continua es, a la vez que un derecho y una obligación de todos los trabajadores, un objetivo estratégico clave para cualquier empresa o institución. En efecto, cada vez resulta más evidente la importancia del conocimiento como recurso y en particular en la sociedad actual caracterizada por rápidos y profundos cambios que exigen considerables esfuerzos de adaptación de todas las organizaciones y niveles sociales y productivos. La formación se convierte así en una exigencia que permite considerarla más como una inversión que como un gasto o un trámite a cubrir. Los planes de formación en el ámbito de las empresas y organizaciones adquieren con este enfoque una especial relevancia dada la coincidencia en sus intereses con los de los propios trabajadores. Eficiencia y calidad deben caminar juntas con oportunidad laboral y promoción profesional.

En esta plataforma se han usado las técnicas del modelo de IN Semántica para la detección de fortalezas y debilidades de un hospital (Hospital

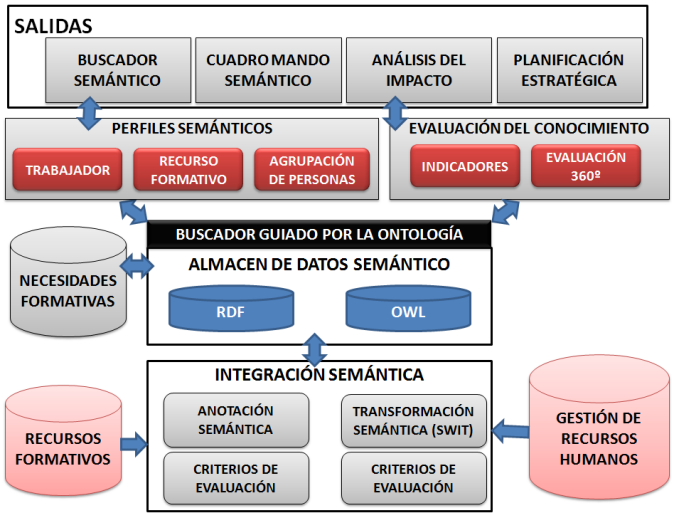


Figura 5.8: Arquitectura de la plataforma de planificación de la formación

Clínico Universitario Virgen de la Arrixaca) para cada trabajador y servicio. Ese análisis permite la automatización del proceso de diseño de planes de formación. Con ello se consigue ofertar una formación adaptada a las necesidades reales, definiendo las líneas de actuación, las acciones formativas, los recursos necesarios y también la recomendación de planes de formación personalizados.

5.2.2 Metodología y Herramientas

El desarrollo de esta plataforma se ha basado en la solución de planificación semántica descrita en la sección 4.6.2. En la figura 5.8 se puede ver la arquitectura de la plataforma.

Para llevar a cabo la puesta en marcha de esta herramienta se han definido cuatro fuentes de información:

- **Definición de puesto de trabajo.** Cada estructura organizativa definirá los requisitos formativos de los puestos de trabajo de su entidad en base a las competencias requeridas.

- **Definición del perfil laboral del trabajador.** Cada usuario de la plataforma definirá su vida formativa, su puesto de trabajo actual y sus aspiraciones a otros puestos, en el caso de que éstos existan. Esta definición permitirá conocer el perfil formativo de cada persona.

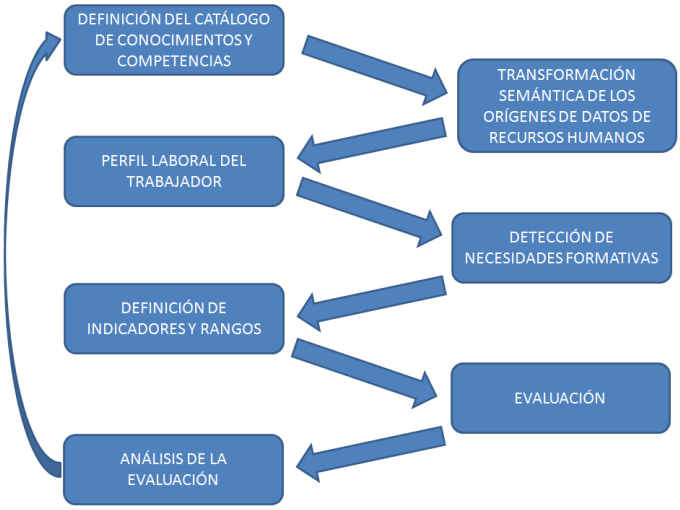


Figura 5.9: Fases de la detección de necesidades formativas

- **Los propios intereses formativos del personal.** Cada persona tiene unos intereses formativos personales, que en ocasiones no se hacen explícitos, por lo que la organización no los puede cubrir convenientemente. Además, los mandos intermedios también podrán definir los conocimientos que deberían adquirir los integrantes de su equipo. Esta definición se realizará siguiendo la jerarquía de la organización.

- **Panel de indicadores.** La plataforma recogerá indicadores de actividad de la organización relacionados con la implantación de los planes de formación. Esta definición permitirá que se pueda medir el impacto que ha tenido esa formación en la entidad.

En la figura 5.9 se pueden ver las fases necesarias para generar toda la información en el repositorio semántico.

En primer lugar, la organización debe definir cuáles son las competencias y conocimientos necesarios para el correcto funcionamiento de sus servicios públicos. Para esta definición se hace uso del modelo de clasificación basada en criterios de evaluación, donde cada criterio se corresponderá con una competencia o conocimiento. En segundo lugar, se recupera la definición de puestos de trabajo del sistema que se use para la gestión de los recursos humanos de la empresa. Además de recuperar los puestos de trabajo, también se recupera la adscripción de cada uno de ellos, es decir, a qué departamento o servicio están asociados. De esta forma se representa en el modelo semántico el organigrama de la compañía y qué puestos hay en cada uno de los nodos

de ese organigrama. Además, también se puede identificar a los trabajadores y el puesto actual que tienen. Para convertir todos esos datos en información semántica se hace uso de SWIT.

A partir de ese momento se puede empezar a clasificar los puestos de trabajo anotándolos con los criterios de evaluación, es decir, definirlos en base a las competencias o conocimientos requeridos. Al hacer esas anotaciones la plataforma es capaz de generar un mapa de competencias de la organización.

En una tercera fase, se empieza la definición del perfil laboral del trabajador. De las dos primeras fases, ya se ha recuperado quién es y cuál es su puesto de trabajo actual. En esta fase, será el propio trabajador el que tenga la opción de completar su vida laboral con otros puestos de trabajo que haya tenido en otras compañías, su vida académica y sus aspiraciones de ascenso dentro de la organización actual. Al igual que en el caso anterior, el trabajador deberá clasificar esta información en base al catálogo de competencias y conocimientos definidos por su empresa. Es importante comentar que, dependiendo de la estructura jerárquica de la organización, a cada trabajador se le asociará un peso específico. Este peso se tendrá en cuenta a la hora de evaluar las necesidades formativas de cada persona.

En una cuarta fase, se usa la herramienta de cuestionarios semánticos para que los usuarios rellenen una encuesta de forma que se conozca cuáles podrían ser los intereses formativos del personal. En este modelo, dependiendo del perfil del trabajador, se puede definir cuáles son sus intereses formativos o cuáles son las necesidades formativas de las personas que dependen operativamente de cualquier mando intermedio o directivo de la empresa. Cada pregunta que se realiza en estos cuestionarios también se anota con el catálogo de competencias. El peso de la anotación se tendrá en cuenta a la hora de analizar esos intereses. Por ejemplo, si un responsable (con peso 4) cree que en su departamento de 7 personas es importante que aprendan a gestionar proyectos, y sin embargo los 7 trabajadores de ese mismo departamento (con peso 1) consideran que deben mejorar su formación en inglés, la aplicación balanceará los resultados dando prioridad a los trabajadores (ya que (7*1) >(1*4)). De esta forma, cuando la aplicación genera los planes de formación, también es capaz de indicar cuáles serían los destinatarios potenciales, ya sea porque lo requiere su jefe, o porque un número importante de ellos está de acuerdo en que necesitan mejorar alguna de sus habilidades.

A continuación, la plataforma permite que se definan los indicadores operativos de la compañía. Esos indicadores también se anotan con las competencias o conocimientos que participan en sus resultados. Gracias a esta fase se podrán evaluar los activos de conocimiento del hospital y detectar cuáles son sus fortalezas y debilidades. Posteriormente se anotan, en función de los

criterios de evaluación, los recursos formativos de los que dispone el hospital. Todos los contenidos se anotarán como oportunidades de mejora, de tal forma que todos tendrán la misma prioridad. Por lo tanto, los planes de formación se basarán en el peso de las propias competencias y en la valoración de los propios recursos formativos para priorizar uno u otro. Es importante destacar que todos los planes generados tendrán dos variables que los limitan: (1) el coste de cada acción formativa, y (2) el tiempo de dedicación necesario para llevarla a cabo.

A continuación se describen los principales módulos de explotación de la información generada en la fase de evaluación.

5.2.2.1 Módulo de generación de planes de formación

Este módulo permite la generación de planes de formación adaptados a las necesidades de cada perfil o persona de la organización. En este módulo la planificación puede realizarse desde dos puntos de vista:

1. **Competencias necesarias.** La aplicación simplemente devuelve aquellas competencias que los usuarios no tienen validadas correctamente y el responsable del plan de formación elije los cursos que pueden satisfacer esos requisitos.

2. **Plan de formación.** Se realiza una fase adicional de anotación semántica con el catálogo de competencias. Esta fase consiste en anotar los cursos que el hospital imparte o a los que tiene acceso. Al hacer esa anotación, la aplicación es capaz de generar automáticamente aquellos cursos que prioritariamente tiene que hacer cada uno de los trabajadores de la organización.

Para poder hacer una comparativa entre qué necesita el hospital y lo qué sabe el trabajador se han definido dos tipos de perfiles semánticos:

1. **Perfil tipo.** Este perfil representa los puestos de trabajo como el conjunto de criterios de evaluación que necesita (competencias y conocimientos). Además de esa anotación, también se especifica en qué nivel, es decir, por ejemplo, si tiene que tener un nivel alto en el manejo de procesadores de texto. En la tabla 5.3 se puede ver la consulta SPARQL que calcula el perfil semántico tipo de un trabajador identificado por la variable *sourceId*. *DesirableLevel* es el concepto básico para saber qué competencias tienen que ser evaluadas y cuál es su nivel esperado.

2. **Perfil del trabajador.** Este perfil se define como la puntuación para cada competencia o conocimiento recibida de la evaluación 360º. En la tabla 5.4 se puede observar la consulta SPARQL que calcula el nivel real de un trabajador para cada competencia. En este caso la clase *ResourceEvaluation* es la que indica la evaluación. Al poder tener múltiples evaluaciones para un mismo recurso, se utiliza la función agregada de SPARQL que calcula la media (*avg*).

Tabla 5.3: Consulta SPARQL para el cálculo del perfil tipo

```
PREFIX ec: <http://www.imib.es/ontologies/EvaluationCriteria>
SELECT ?score ?competency WHERE {
?desirableLevel rdf:type ec:DesirableLevel .
?desirableLevel ec:hasAssessableResource ?resource .
?sourceId ec:sourceId ?x . FILTER (?x = 'sourceId') .
?desirableLevel ec:hasScore ?scoreConcept .
?scoreConcept ec:hasScore ?score .
?desirableLevel ec:hasEvaluationCriteria ?ec .
?ec ec:description ?competency }
```

Tabla 5.4: Consulta SPARQL para el cálculo del perfil del trabajador

```
PREFIX ec: <http://www.imib.es/ontologies/EvaluationCriteria>
SELECT avg(?score) ?competency WHERE {
?resourceEvaluation rdf:type ec:ResourceEvaluation .
?resourceEvaluation ec:hasAssessableResource ?resource .
?sourceId ec:sourceId ?x . FILTER (?x = 'sourceId') . ?resourceE-
valuation ec:hasScore ?scoreConcept . ?scoreConcept ec:hasScore
?score .
?resourceEvaluation ec:hasEvaluationCriteria ?ec .
?ec ec:description ?competency }
group by ?ec
```

Gracias a la definición de estos dos perfiles, una función sencilla de similitud entre éllos va a dar como resultado en qué necesita formarse el trabajador.

5.2.2.2 Módulo de evaluación del impacto

Este módulo permite evaluar el impacto de cada plan de formación generado y ejecutado. Como ya se ha comentado, se pueden definir tantos indicadores

como sea necesario. Los datos que se cruzan provienen de (1) esa definición de indicadores y la recogida de sus valores, y (2) de la evaluación 360º de los trabajadores. Este modelo de evaluación también puede usarse para las propias acciones formativas, aportando información sobre el índice de satisfacción del alumnado.

Este módulo puede ayudar a identificar si el plan de formación realmente ha tenido éxito. La información que se obtiene de esta herramienta se divide en:

- La medida del indicador es mejor que la vez que se calculó antes de impartir el curso, y a su vez la evaluación de las competencias relacionadas con ese indicador también es mejor. La herramienta devuelve directamente que el curso impartido es apropiado y que ha tenido un impacto positivo en la organización.
- La medida del indicador es mejor que la anterior y la evaluación de las competencias relacionadas con ese indicador es igual o más baja. En este caso la herramienta identifica que hay un posible error al anotar el indicador, y que es posible que esa competencia o conocimiento no tenga incidencia en el cumplimiento de los objetivos asociados a dicho indicador.
- El indicador es peor que el del cálculo anterior y la evaluación de las competencias relacionadas con ese indicador es mejor. En este caso sucede algo parecido al anterior, aunque la herramienta también avisa de que puede existir algún factor externo que no ayude a que ese indicador mejore.
- El indicador empeora y la evaluación de los competencias también es más baja. En este caso, la plataforma asume que el curso impartido no es bueno y que se debería de buscar otro proveedor de formación.

5.2.3 Resultados y evaluación

5.2.3.1 Plataforma funcional

A continuación se describen las principales herramientas de esta plataforma[6]:

- Herramienta de visualización de mapas de competencias. Permite visualizar gráficamente los mapas de competencias de un determinado

[6] http://www.ffis.es/Competencias

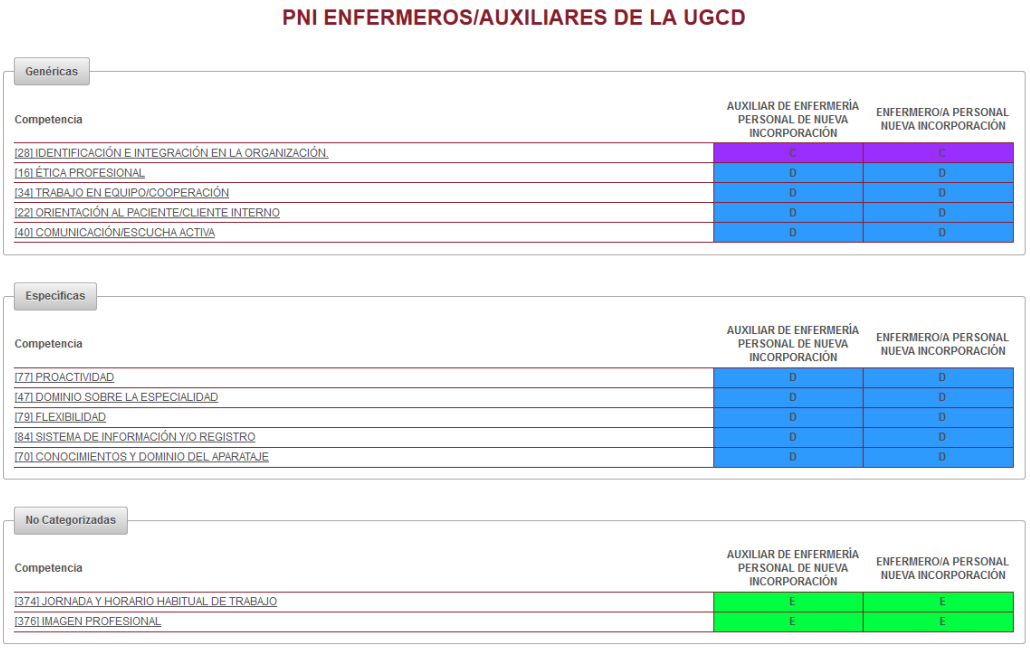


Figura 5.10: Mapa de competencias

puesto de trabajo. En la figura 5.10 se puede ver una captura de pantalla del mapa de competencias de un puesto de trabajo. Por código de colores se ve la importancia y cuál es el nivel deseado.

- Herramienta de explotación de la evaluación. Permite analizar gráficamente los resultados de la evaluación. En la figura 5.11 se puede ver la comparación de un trabajador con su perfil tipo. En la figura 5.12 se puede ver una comparativa de un grupo de trabajadores del mismo servicio.

- Herramienta de planificación. Permite hacer planes bajo demanda basados en la evaluación del usuario. El hospital no disponía de contenidos formativos oficiales, pero sí disponía de un catálogo de cuáles deberían ser esos contenidos. Por ese motivo se optó por anotar esos ítems de contenidos, que únicamente tenían título y objetivos, y que la plataforma ayudara a detectar cuáles eran más importantes a desarrollar. En la figura 5.13 se puede ver una captura de un plan de formación para jefes de unidad de gestión clínica. En ella se ve un extracto del plan recomendado y, para cada contenido del curso, cuáles son las competencias que adquiere el trabajador.

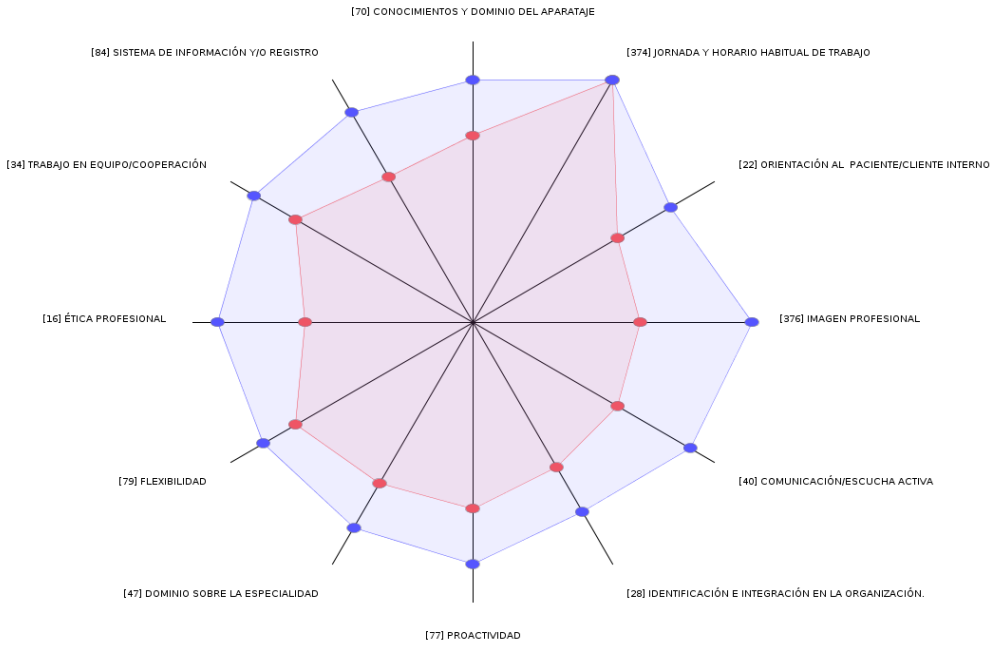


Figura 5.11: Explotación de la evaluación de un trabajador

5.2.3.2 Evaluación del uso de la plataforma

Esta plataforma se puso en marcha en el Hospital Clínico Universitario Virgen de la Arrixaca (HCUVA) en 2013 y sigue estando operativa. A continuación se listan los principales resultados:

- **Competencias.** Se ha definido un catálogo de competencias con 252 elementos, de las cuales 111 son conocimientos, 36 son competencias y 105 son comportamientos asociados a esas competencias. Los comportamientos serán los criterios de evaluación que se usan para puntuar una competencia. Es decir, la evaluación de la competencia será la media de la evaluación de cada uno de los comportamientos asociados.

- **Departamentos/servicios.** Se han definido 83 departamentos, servicios y unidades del hospital.

- **Puestos de trabajo.** Se han definido 19 puestos de trabajo diferentes en el hospital.

- **Personas.** Se han evaluado a 377 personas del hospital. La media del perfil tipo es de 4,33 sobre 5 y la puntación total obtenida ha sido de

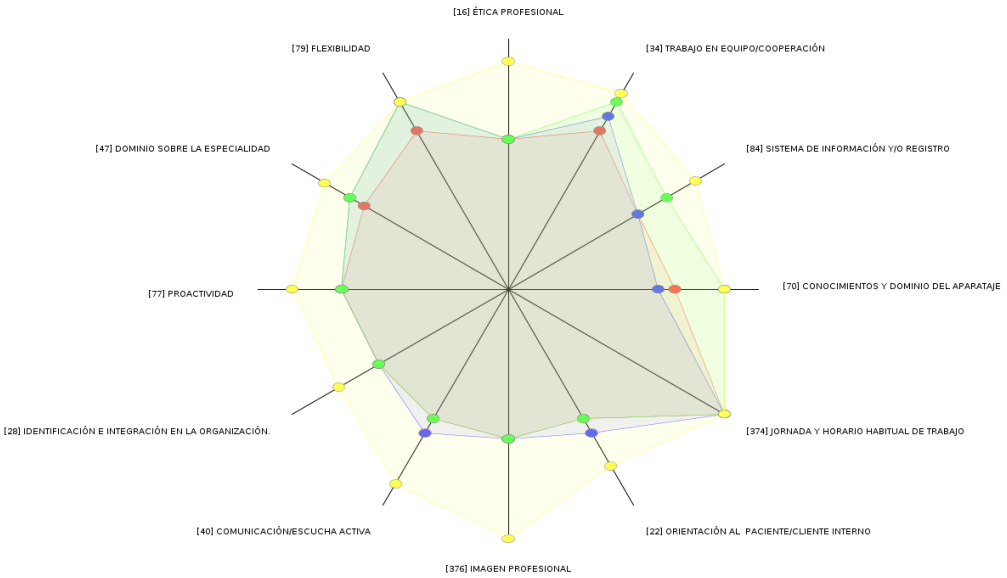


Figura 5.12: Comparación de la evaluación entre trabajadores

3,95. Esto quiere decir que un número importante del hospital necesita mejorar en algunas de sus competencias para llegar, como mínimo, al perfil tipo.

Gracias a la evaluación se han propuesto 18 actividades formativas para cubrir las necesidades del personal del hospital. Además, la aplicación ha sido capaz de identificar a los potenciales alumnos de esas actividades. En este caso concreto los posibles destinatarios de las actividades serían 131 trabajadores. Catorce de los diecinueve departamentos que se han evaluado necesitarían de un plan especial de formación.

Actualmente, el almacén de datos semántico contiene 100.384 tripletas. En la tabla 5.5 se puede ver la evolución de la fuente de datos. El número de tripletas generado en el año 2013 es mucho mayor que el resto porque se cargó la mayoría de información de la base de datos de recursos humanos y, además, se definieron los catálogos de competencias y conocimientos. Las competencias, mapas de recursos humanos y convocatorias de evaluación han sido registradas por los cinco administradores de la plataforma que pertenecen al departamento de recursos humanos del hospital. En la figura 5.14 se puede ver una gráfica con la evolución del número de tripletas comparado con el número de evaluaciones.

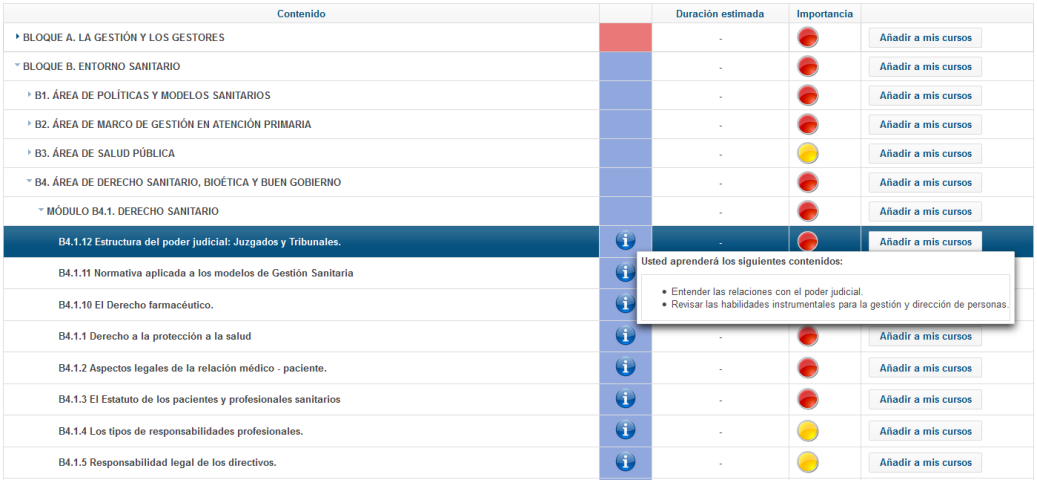


Figura 5.13: Herramienta de planificación de la formación

Tabla 5.5: Evolución del almacén semántico

	2013	**2014**	**2015**	**TOTAL**
Tripletas nuevas	76.325	18.034	6.025	100.384
Evaluaciones nuevas	156	178	43	377
Nº de usuarios nuevos	804	506	210	1.520

Como resultado de este caso de uso se destaca que, a partir de los resultados obtenidos, a los que el propio evaluado ha tenido acceso, se solicitaron dos traslados de servicio, ya que las personas evaluadas estaban totalmente de acuerdo con su valoración y consideraban que no tenían los conocimientos suficientes para trabajar en dicho servicio. Por otro lado, las evaluaciones que se han realizado son a personal temporal, ya que los sindicatos se han opuesto a la evaluación del personal estable del hospital.

5.3 SECARE: Explotación semántica de un registro de cáncer

5.3.1 Introducción

Los registros de cáncer son importantes para la investigación y para mejorar la calidad del tratamiento de esta enfermedad [186; 187]. Muchas soluciones tecnológicas permiten gestionar y analizar un registro de cáncer (por ejem-

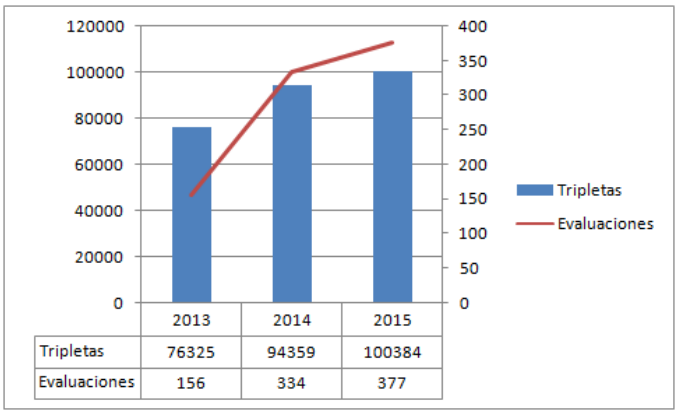


Figura 5.14: Gráfica de evolución del almacén semántico

plo METRIQ[7], OncoLog Registry[8] o CNEXT[9]). Sin embargo, la ausencia de modelos semánticos bien definidos se convierte en un problema cuando los usuarios quieren hacer análisis específicos, enlazar con fuentes de datos externas, o comparar la información de diferentes registros entre sí [188]. En este caso de uso se ha desarrollado una ontología que modela un registro local de cáncer. Además, se han empleado diferentes módulos de el marco de trabajo de IN semántica para desarrollar una plataforma Web donde los clínicos pueden ver, analizar y comparar los datos de un registro.

5.3.2 Metodología y Herramientas

La solución propuesta en la sección 4.6.3 ha sido aplicada en un registro local de cáncer (en la figura 5.15 se puede ver la arquitectura). Basado en el análisis de requisitos [189; 190] se ha desarrollado una ontología que modela un registro epidemiológico local de cáncer. Esta ontología es el núcleo principal del sistema, sobre el que se construye todo el modelo de gestión y explotación de la información. Utilizando SWIT como motor de transformación semántica se ha integrado una base de datos relacional con el esquema de un registro local de cáncer al almacén semántico.

Como servicios de explotación intermedios, la plataforma permite la definición de perfiles semánticos de pacientes y de grupos de pacientes con propiedades comunes o cohortes. También permite la definición de indicadores

[7]http://www.elekta.com/healthcare-professionals/products/elekta-software/cancer-registry.html

[8]http://www.oncolog.com/?cid=7

[9]http://www.askcnet.org/

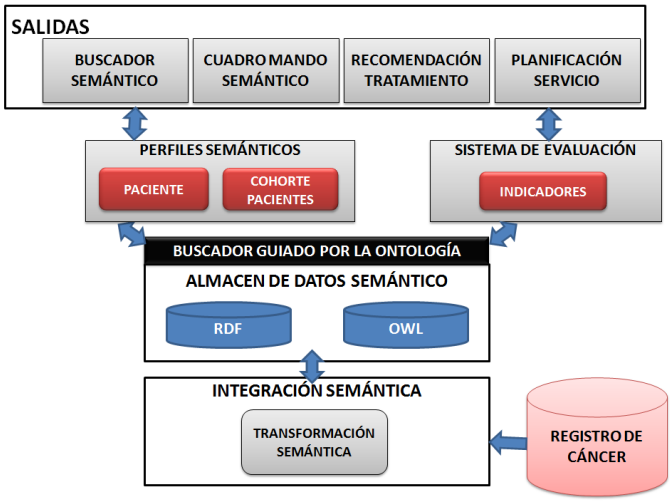


Figura 5.15: Arquitectura tecnológica de SECARE

de funcionamiento del sistema.

El desarrollo de una plataforma Web permite que los usuarios formulen consultas a través de una interfaz gráfica con ODS. Los resultados de dichas consultas pueden ser representados visualmente en diversos formatos, permitiendo la generación de cuadros de mando personalizados. Además, las complejas líneas de tiempo de un paciente o de una cohorte pueden representarse gráficamente de una forma clara y estructurada, gracias a la generación de perfiles semánticos.

El módulo de explotación se compone de cuatro herramientas:

- Buscador semántico. Esta herramienta permite hacer búsquedas avanzadas sobre todo el modelo de datos haciendo uso de ODS.

- Cuadro de mando semántico. En este caso, cada usuario puede generar cuadros de mando bajo demanda usando las características de agregación de ODS.

- Recomendación de tratamiento. El recomendador de tratamiento permite definir los criterios clínicos de un paciente y a partir de ellos generar un filtro con los pacientes de la base de datos que cumplan esos criterios. Calculando el perfil semántico de cada uno de ellos y generando un árbol de redes bayesianas el usuario puede ir viendo cuál va a ser el tratamiento más probable y la evolución en cada uno de los casos. Además, el usuario tiene la opción de seleccionar terapias o evo-

luciones y ver cómo se recalcula toda la red, indicándose las opciones más probables en cada caso.

- Planificación del servicio. En este caso no se hace uso del sistema de planificación semántico que se ha comentado previamente. Es decir, no hay evaluación de los activos de conocimiento, ni definición de recursos. Se usa el recomendador de tratamientos para un paciente como herramienta que ayude a predecir cuál será la carga real del servicio en un número concreto de meses. Para ello, el modelo predice cuáles serán los tratamientos y evoluciones más probables para los pacientes que hay en el sistema, aportando a los gestores del centro sanitario una información que puede ser muy importante a la hora de planificar los recursos que va a necesitar un determinado servicio. Realmente sería un recomendador de tratamiento para todos los pacientes activos de la base de datos. Este sistema no tiene en cuenta a pacientes nuevos, es decir, no prevee la inclusión de nuevos pacientes a la hora de estimar la carga de trabajo de los servicios.

5.3.3 Resultados y evaluación

5.3.3.1 Ontología de un registro local de cáncer

En los últimos años se han desarrollado diversas ontologías que modelan el cáncer, o algún tipo en concreto [191; 192; 193]. Para este trabajo se ha desarrollado una versión preliminar de una ontología[10] (basada en otros trabajos preliminares) que satisface los requisitos de un registro local de cáncer. Como se puede observar, para la definición de esta ontología se han usado los conceptos básicos que se describen en el marco de trabajo propuesto, como son: (1) Paciente, (2) Diagnóstico, (3) Terapia y (4) Evolución de la enfermedad. La ontología ha sido definida en OWL-DL y contiene un total de 335 clases, 18 propiedades y 29 relaciones, con 2.581 axiomas lógicos. En la figura 5.16 se puede ver un extracto de la ontología de registro de cáncer. La ontología aborda los siguientes conceptos:

- *Patient* representa una persona con un tipo de cáncer. En este concepto se representan propiedades como sexo, edad, diagnósticos, terapias, etc.

- *Patient Condition* representa la condición de salud de un paciente en un momendo dado. Fecha, edad, peso, altura, índice Karnofsky [194], índice ASA [195] y el estado menopaúsico son algunas de las propiedades que se pueden encontrar.

[10] http://sele.inf.um.es/ontologies/cancer-registry.owl

- *Diagnosis* modela el diagnóstico de un paciente en un momento dado. En esta clase se representan relaciones con otras terminologías como el ICD10 [177], grado, severidad, estadío, terapias, estructura patológica, estructura anatómica y tipo de tumor.

- *Therapy* representa las terapias que se han aplicado al diagnóstico de un paciente en un momento dado. Diferentes tipos de terapia como quimioterapia, cirugía, medicina nuclear, etc. han sido modelados como subclases de ésta. En estos conceptos se encuentra información sobre medicación o fecha de inicio y fin de la terapia.

- *Disease Course* modela la evolución de un diagnóstico de un paciente en una revisión clínica concreta. Los tipos de evolución asociados al cáncer como "remisión completa", "progresión" o "recurrencia" han sido definidos como subclases. En estos conceptos se pueden encontrar relaciones con otros como *Patient Condition* o el médico que ha realizado la valoración.

- La ontología también incluye algunas clases que representan la clasificación TNM [196] de tumores malignos. Esta clasificación incluye jerarquías de estadificación de la enfermedad.

- *Health Classification System* es la superclase que representa todos los estándares de clasificación clínica que se ha usado. Para construir estos sistemas se han reutilizado otras ontologías. Por ejemplo, para ICD10 se ha usado la ontología publicada en [197]. Para el ICD-O [198] y el ICD10-PCS [199], se han transformado hojas de cálculo en OWL.

5.3.3.2 Prototipo funcional

A continuación se describen las principales herramientas de esta plataforma[11]:

- En la figura 5.17 se puede observar el servicio que permite que los usuarios puedan ver el cronograma de un paciente con cáncer. En esta vista, los usuarios pueden ver los detalles completos del diagnóstico y de las terapias aplicadas en cada fase de la enfermedad. Además, los usuarios disponen de dos gráficos de evolución de la enfermedad basados en el índice Karnofsky. En esta captura concreta se puede ver un extracto del cronograma terapeútico y de la evolución de un paciente con cáncer de faringe.

[11] http://sele.inf.um.es/SECARE/

Figura 5.16: Extracto de la ontología del registro local de cáncer

- En la figura 5.18 se muestra cómo se puede usar ODS para hacer una selección de pacientes con los siguientes criterios: pacientes masculinos con edades comprendidas entre 50 y 70 años diagnosticados con cáncer colorrectal que hayan recibido quimioterapia. Después de realizar ese filtro, el sistema genera la matriz de redes bayesianas que serán representadas como un cronograma con las distribuciones de probabilidad de las terapias y evoluciones de los pacientes seleccionados. Gracias a la red bayesiana, este sistema puede usarse como un simulador terapeútico. Opcionalmente, el cronograma puede ser recalculado cuando el usuario selecciona la terapia a aplicar. Esta herramienta sirve de ayuda al clínico para estimar cuál es la terapia que mejor va a funcionar en el paciente a partir de lo que ha aprendido del histórico de información que hay registrada en el almacén semántico. En la figura 5.19 se puede ver un extracto de los primeros dos meses de las terapias para un grupo de 60 pacientes.

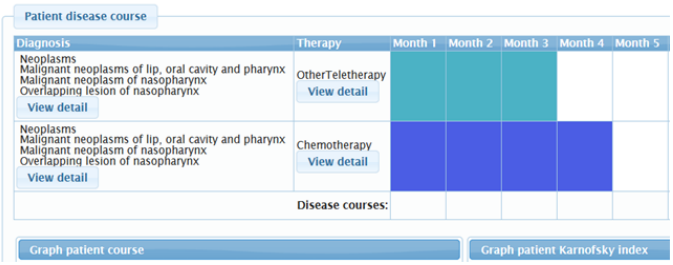


Figura 5.17: Perfil de un paciente con cáncer de faringe

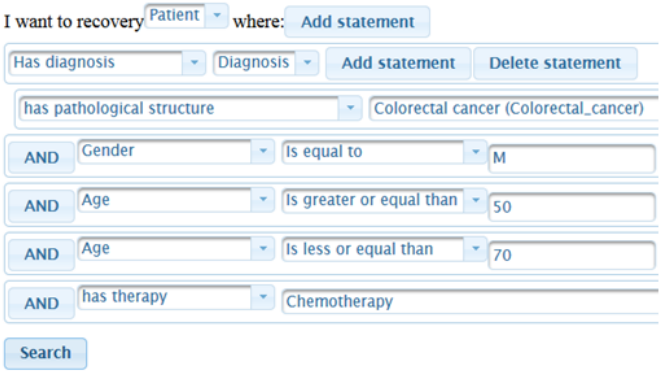


Figura 5.18: Buscador semántico basado en ODS

- En la figura 5.20 se puede ver cómo se ha configurado un cuadro de mando que permite comparar las terapias aplicadas a pacientes de un determinado tipo de cáncer con diferentes edades. En la parte de abajo del gráfico aparecen las tablas con los datos, que son enlaces al buscador semántico para encontrar esos resultados concretos. Estos cuadros de mando pueden guardarse para posteriormente ser ejecutados de nuevo. También se pueden almacenar con campos parametrizables para que posteriormente sean reutilizados definiendo esos parámetros en tiempo de ejecución.

5.3.3.3 Evaluación

En este caso de uso no se ha tenido ocasión de trabajar con datos reales, pero sí con las indicaciones y requisitos del "*Comprehensive Cancer Center of Freiburg*" (CCCF), un registro de cáncer de una ciudad alemana que tiene aproximadamente unos 5.000 casos nuevos al año. Para hacer las pruebas

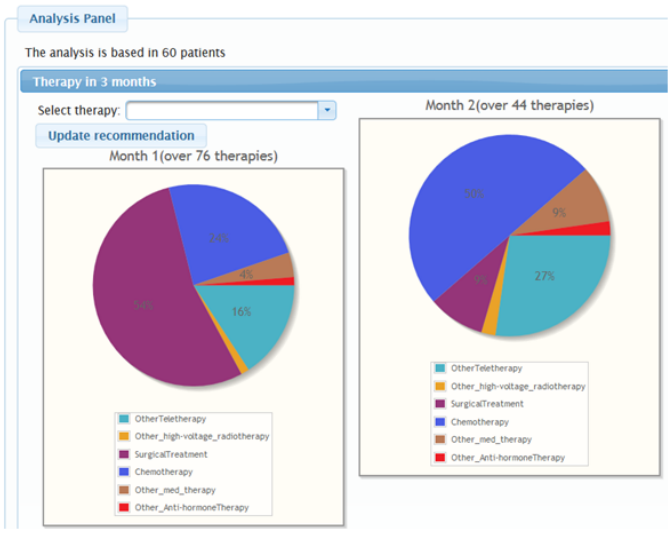


Figura 5.19: Extracto de la recomendación de tratamientos

de evaluación se han generado conjuntos de datos simulados a través de funciones aleatorizadas. Se ha construido una base de datos con el mismo esquema que usan en CCCF con 207.190 pacientes, en los que cada uno tiene uno o más diagnósticos, y tienen una o más terapias para cada diagnóstico. Para que los datos fuesen los más cercanos a la realidad posible, gracias a los modelos estudiados de CCCF, se definen una serie de reglas para que las terapias tengan sentido en relación con el tipo de cáncer, la edad y sexo del paciente.

Para realizar las pruebas de rendimiento de esta plataforma se ha usado como motor de base de datos relacional MySQL 5 y Virtuoso 7 como repositorio RDF. La máquina empleada tiene 8GB RAM y un procesador Intel Core I7-3610QM a 2.30 GHz.

En una evaluación inicial del sistema se analiza el tiempo necesario para realizar la transformación de los datos desde el modelo relacional al semántico. La media de tiempo de las transformaciones es de 32 minutos para el volumen que se ha comentado anteriormente. Debido a que el número de tripletas cargadas en el almacén semántico superaba los 25 millones se tuvo que aumentar la memoria RAM que usa Virtuoso a 2GB RAM.

En esta plataforma se han identificado claramente dos modelos de datos equivalentes, uno en el almacén semántico y otro en un modelo relacional tradicional. En la tabla 5.6 se muestra un análisis del rendimiento entre ambos enfoques. En esa tabla se puede ver que el modelo propuesto es más lento en consultas sobre una tabla, es decir, sin condiciones de reunión. Sin

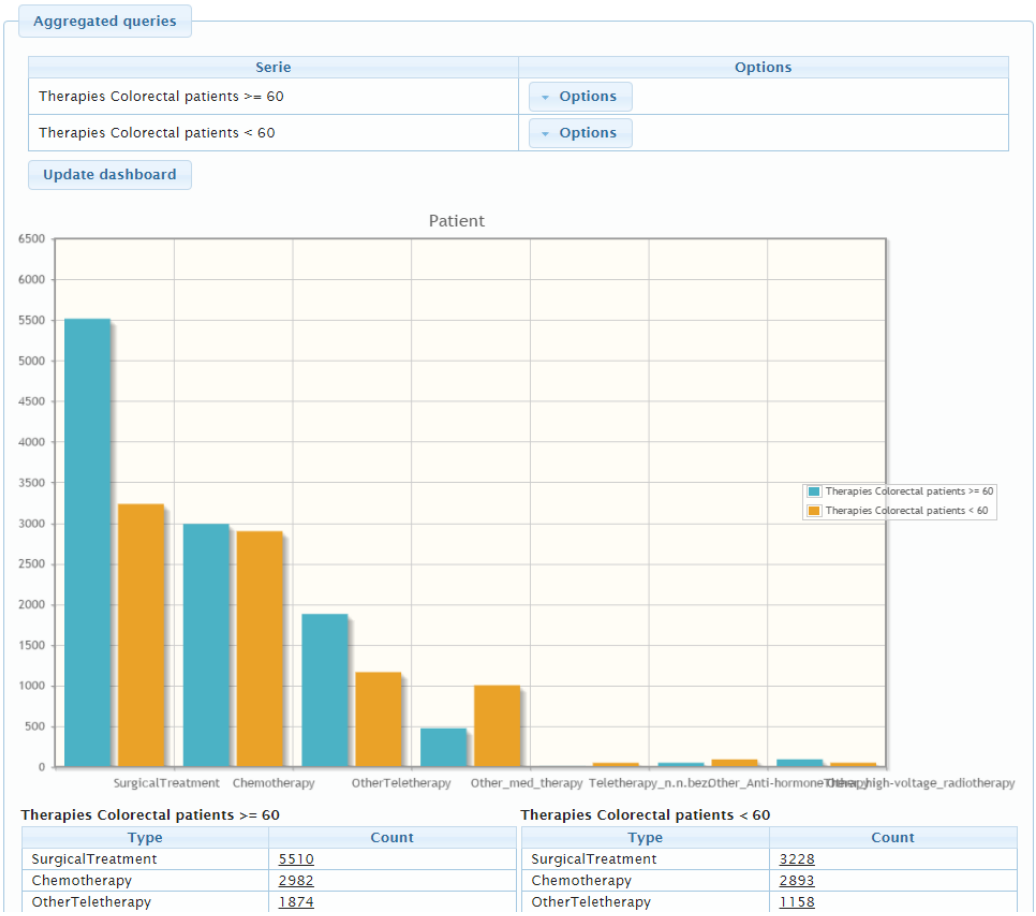


Figura 5.20: Cuadro de mando comparativo

embargo, cuando se enlaza la información, el modelo semántico es bastante más eficiente. Además, cuando se filtra en una tabla concreta, por uno o varios campos, el modelo semántico también es más rápido que el relacional.

Tabla 5.6: SECARE: Resultados de la comparativa entre SQL y SPARQL

Consulta	Resultado SQL	Tiempo SQL	Resultado SPARQL	Tiempo SPARQL
Recuperar pacientes	207.190	0,060s	207.190	0,189s
Recuperar Terapias	400.290	0,132s	400.290	0,317s
Recuperar Diagnósticos	240.088	0,070s	240.088	0,220s
Recuperar Evoluciones	108.297	0,030s	108.297	0,155s
Recuperar pacientes con diagnósticos, terapias y evoluciones	207.190	1,048s	207.190	0,204s
Recuperar pacientes que sean mujeres	105.714	0,231s	105.714	0,189s
Recuperar pacientes que sean mujeres y mayores de 60 años	62.603	0,245s	62.603	0,192s

5.4 SECOLON: Explotación semántica de un programa de cribado de cáncer colorrectal

5.4.1 Introducción

La detección precoz del cáncer de colon y recto se ha convertido en una herramienta fundamental para combatir esta enfermedad [200]. El "Programa de prevención de cáncer de colon y recto"[12] de la Región de Murcia tiene como objetivo hacer un seguimiento de personas entre 50 y 69 años para diagnosticar el nivel de riesgo que tienen en contraer este enfermedad. En este caso de uso, se ha colaborado con las personas que coordinan este programa para desarrollar una plataforma Web que explote este registro. Para ello, se ha desarrollado una ontología que modela el proceso de cribado en este tipo de cáncer. Además, se han usado diferentes métodos de la propuesta de IN Semántica para construir una plataforma que permita explotar los datos del programa y determinar qué nivel de riesgo tendrá un paciente en su siguiente revisión.

[12]http://www.murciasalud.es/pagina.php?id=123691

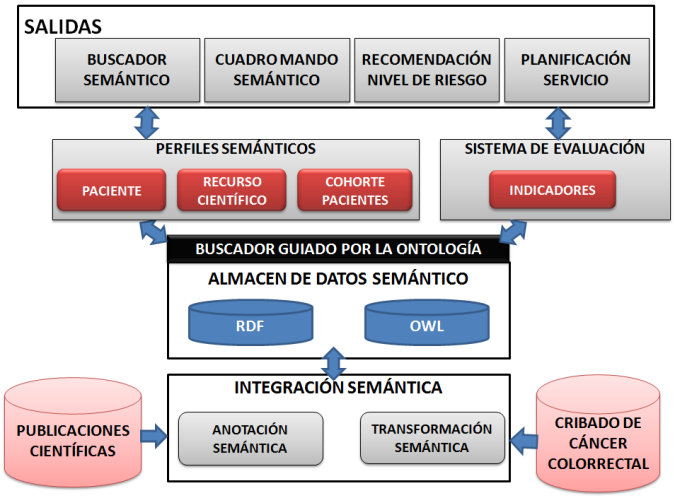


Figura 5.21: Arquitectura tecnológica de SECOLON

5.4.2 Metodología y Herramientas

El desarrollo de este prototipo se basa en la solución de Análisis Epidemiológico propuesta en la sección 4.6.3. SECOLON es una plataforma semántica que permite explotar un registro de cribado de cáncer colorrectal desde dos puntos de vista: (1) generación de informes y cuadros de mando semánticos personalizados, y (2) recomendaciones de los niveles de riesgo que tendrá un paciente en el futuro. En la figura 5.21 se puede ver la arquitectura real del prototipo. En este caso se tiene una única ontología que modela el dominio en cuestión y que se alimenta de la base de datos relacional de la que disponen actualmente, usando SWIT como motor de transformación semántico. Además, a través del modelo de anotación semántica manual también se incorporan publicaciones científicas de los servicios de digestivo de la Región de Murcia a través de la base de datos del Instituto Murciano de Investigación Biosanitaria. Estas publicaciones serán útiles para hacer recomendaciones sobre casos concretos a partir de los diferentes perfiles semánticos.

Como servicios de explotación intermedios, la plataforma permite la definición de perfiles semánticos de pacientes y de grupos de pacientes. Los perfiles semánticos se han definido de forma distinta a SECARE, ya que se ha considerado que deben tener todas las variables clínicas de cada proceso en el que se produce un diagnóstico. En la figura 5.22 se puede ver un esquema del perfil semántico de un paciente concreto. En la imagen, se observa que el perfil es un árbol de tres niveles, donde el nodo central es el diagnóstico o clasificación del paciente. Además de esas variables, también se generan va-

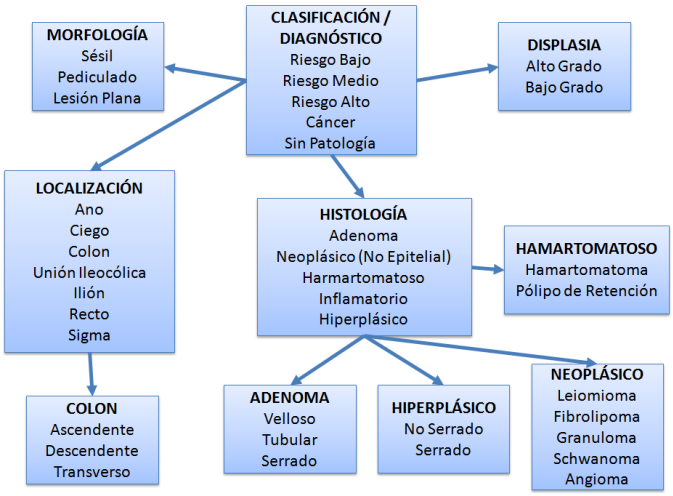


Figura 5.22: Perfil semántico de un paciente en SECOLON

riables específicas como el tamaño, número y velocidad de crecimiento de los pólipos encontrados en el colon. Como se ha comentado, también se generan perfiles de los diferentes recursos científicos. Esos perfiles se definen por las anotaciones que reciben de los profesionales sanitarios cuando se registran en el sistema. Por último, la plataforma también permite la definición de indicadores de funcionamiento del programa de prevención.

El módulo de explotación se compone de cuatro herramientas:

- Buscador semántico. Este servicio permite hacer búsquedas avanzadas sobre todo el modelo semántico haciendo uso de ODS. Los resultados de la búsqueda pueden parametrizarse y exportarse a ficheros tabulares.

- Cuadro de mando semántico. Gracias a esta herramienta cada usuario puede generar sus propios cuadros de mando para realizar sus propios análisis sobre los datos.

- Recomendación del nivel de riesgo. En este caso, sobre la agregación de perfiles semánticos de pacientes con propiedades comunes, se construye una red de Bayes que será capaz de predecir el nivel de riesgo del paciente en el futuro. Para implementar la red bayesiana se ha hecho uso de un Bayes voraz con pesos. En este caso no hay ciclos por lo que no ha sido necesario el uso del modelo de árbol de redes bayesianas, que sí se tuvo que emplear en SECARE.

- Planificación del servicio. Al igual que en SECARE, esto no es el mo-

delo de planificación semántico que se había comentado. En este caso se usa el recomendador de niveles de riesgo para predecir la carga de los servicios de digestivo, cirugía torácica y oncología a partir de los pacientes del programa de cribado. Como los pacientes están clasificados por su hospital de referencia, el modelo propuesto es capaz de hacer una predicción por centro.

5.4.3 Resultados

5.4.3.1 Ontología para el cribado de cáncer de colon y recto

En los últimos años se han desarrollado diversas ontologías que modelan el cáncer colorrectal [201]. Para este trabajo se ha desarrollado una ontología[13] (basada en otros trabajos preliminares) siguiendo los requisitos de un programa de cribado de cáncer colorrectal. Como se puede observar, al igual que en SECARE, se han usado los conceptos básicos de la solución para epidemiología, como son: (1) Paciente, (2) Diagnóstico, (3) Terapia y (4) Evolución del cribado. La ontología ha sido definida en OWL-DL y contiene un total de 364 clases, 43 relaciones, 37 propiedades y 2.100 axiomas lógicos. En la figura 5.23 se puede ver un extracto de esta ontología. A continuación se describen los principales elementos:

- *Patient* representa a una persona que está dentro del programa de cribado. En este concepto se representan la edad y el sexo, además de todos los informes de revisiones que se han realizado durante su inclusión en el programa de cribado.

- *Diagnosis* representa una jerarquía con los posibles diagnósticos del programa de cribado. Cada diagnóstico se asocia con las pruebas necesarias para obtenerlo. Es decir, hay diagnósticos relacionados con la prueba de diagnóstico colonoscopia, otros con enteroscopia, otros con la prueba de sangre oculta en heces y otros con cirugía.

- *Therapy* modela las posibles clasificaciones del paciente en lo relativo a cuándo será su próxima visita. Es decir, cuándo se le volverá a revisar y cuál será la prueba diagnóstica a realizar.

- *Screening course*. En este concepto se representa la evolución del paciente. Se indica según el nivel de riesgo que tiene para desarrollar un cáncer de colon y recto. Para este caso de uso concreto se han modelado

[13] http://sele.inf.um.es/ontologies/cancer-registry.owl

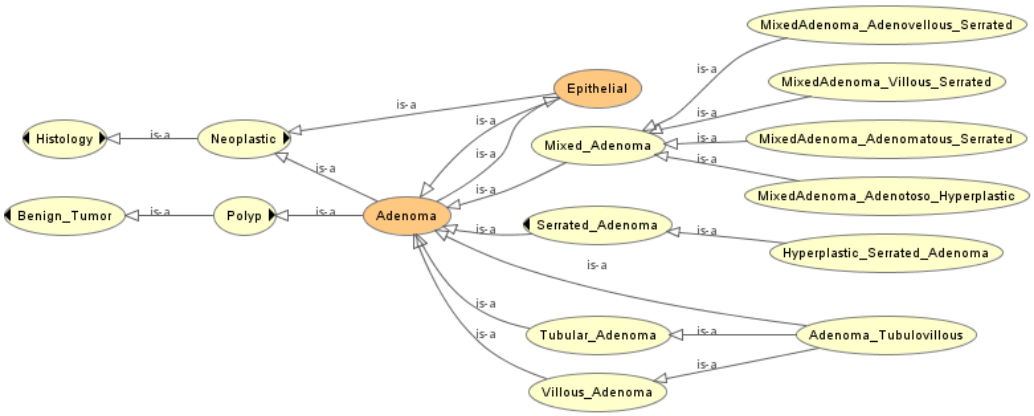


Figura 5.23: Extracto de la ontología SECOLON

las dos guías de referencia que usan en el programa: (1) la guía americana que diferencia entre bajo riesgo y algo riesgo, y (2) la guía europea que además de esos dos niveles añade un nivel de riesgo intermedio.

- *Tumor* representa una jerarquía de los diferentes tumores que se pueden encontrar en este tipo de enfermedad. La primera subclasificación que se hace son los tumores benignos y malignos.

- *Staging type* representa los diferentes estándares que se usan para estadificar la condición del paciente. Además del estándar TNM (también usado en SECARE), también se usan dos estándares más propios de este tipo de cáncer como son el estadío de Astler [202] y el estadío de Dukes [203].

- *Sample* representa las muestras que se recogen del paciente en cada una de las pruebas diagnósticas. De cada una de ellas se almacena su histología (representada por la clase *Histology*) y su localización (representada por la clase *Location*), incluyendo su tamaño en las tres dimensiones. En algunos casos, cuando la muestra es lo suficientemente grande para verse macroscópicamente, se almacena su *Macroscopic_Configuration*. Para cada muestra también se almacena su grado de displasia (representada por la clase *Dysplasia*).

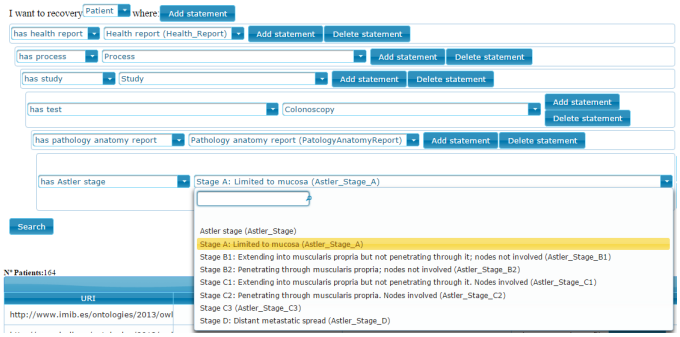


Figura 5.24: Buscador semántico SECOLON

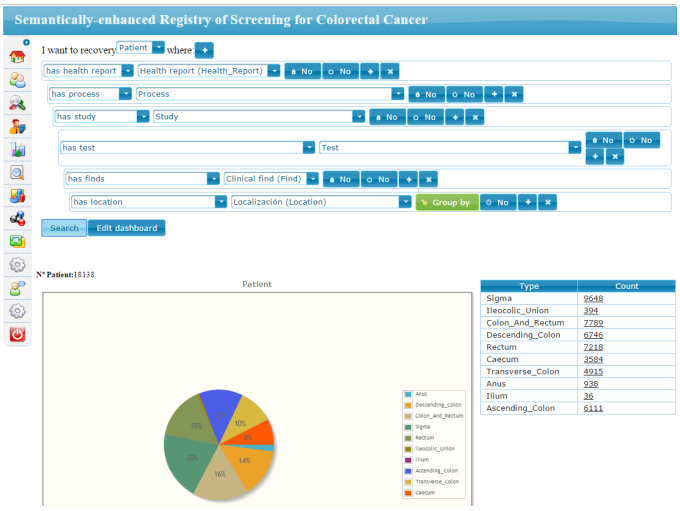


Figura 5.25: Cuadro de mando SECOLON

5.4.3.2 Prototipo funcional

A continuación se describen las principales herramientas de esta plataforma[14]:

- Buscador semántico. En la figura 5.24 se puede ver el buscador semántico que permite hacer búsquedas avanzadas sobre todo el modelo de datos. Además de poder buscar, también permite filtrar los campos a devolver y que se puedan exportar en ficheros tabulares.

- Cuadro de mando semántico. En la figura 5.25 se puede ver un cuadro de mando que permite saber las localizaciones más comunes de indicios

[14] http://sele.inf.um.es/SECOLON/

de riesgo para cada paciente del programa de cribado. Al igual que en SECARE, los usuarios tienen la opción de generar sus propios cuadros personalizados, guardarlos para ejecutarlos posteriormente, o configurarlos como cuadros de mando parametrizables donde en tiempo real se puede cambiar alguna de las variables. Otra de las opciones es que permite la comparación de consultas SPARQL desde un modo gráfico. En la figura 5.26 se puede ver una comparativa entre los tres niveles de riesgo y el diagnóstico de cáncer.

- Recomendador de nivel de riesgo. En esta herramienta se han empleado redes bayesianas usando como nodos de la red los conceptos del perfil semántico del paciente. Para calcular la predicción en el futuro se usa el 90 % de los datos de los que se dispone en ese momento en la base de datos. Haciendo uso del algoritmo WAODE definido en [204] se le asignan pesos a cada uno de los nodos para que haya variables que influyan más que otras en el resultado. El 10 % restante se usa para ajustar esos pesos, de tal forma que se obtenga el mayor acierto posible. Una vez que se sabe el nivel de riesgo, la plataforma recomienda artículos científicos relacionados con los pasos a seguir en esa fase posterior, por si el clínico quiere añadir algo más a la guía de seguimiento estándar. Esta recomendación se realiza en base a la función de similitud semántica propuesta en este trabajo. Como criterio de confianza se ha usado el factor de impacto de la revista de la publicación multiplicado por el número de citas.

- Planificador del servicio. Al igual que en SECARE, se usa el recomendador anterior para planificar la carga de los servicios implicados en el proceso de cribado. Si se quiere que el planificador funcione para cada uno de los hospitales implicados es importante añadir al perfil semántico de los pacientes su centro de referencia. De esta forma se podrá distinguir fácilmente los casos de cada hospital. También es importante que la ontología tenga reglas para separar qué pruebas diagnósticas son de un servicio o de otro. Por ejemplo, una colonoscopia será del servicio de digestivo, mientras que una intervención quirúrgica corresponderá al servicio de cirugía torácica. El resultado final de esta planificación son las pruebas diagnósticas que habrá que realizar para todos los pacientes registrados en el programa en un momento determinado de tiempo (medido en meses) predefinido por el usuario. Si además la ontología almacena el coste de dichas pruebas, el planificador devolverá el impacto económico que tendrá en los diferentes servicios. Por último, al igual que el recomendador individual, esta plataforma recomienda ar-

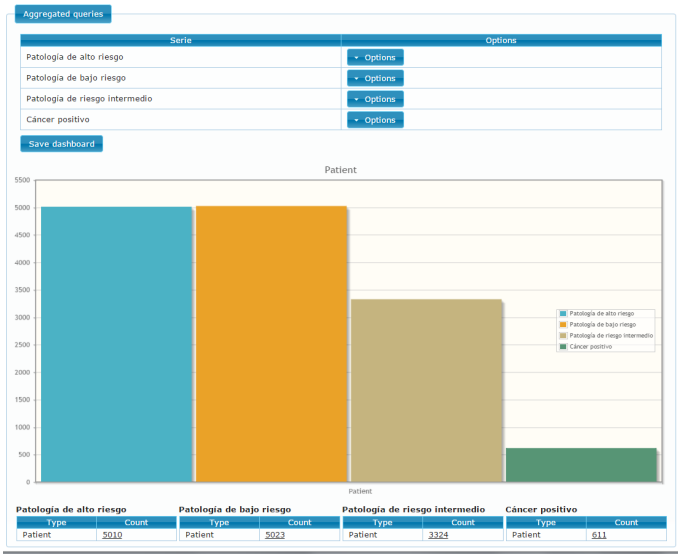


Figura 5.26: Cuadro de mando comparativo SECOLON

tículos científicos que tengan que ver con los procesos a desarrollar en el futuro.

5.4.3.3 Evaluación

Para este caso de uso se ha trabajado con datos reales. Concretamente con los datos de 322.839 pacientes reclutados desde el año 2006, cedidos por el Programa de prevención de cáncer de colon y recto de la Consejería de Sanidad de la Región de Murcia.

Para realizar las pruebas de rendimiento de esta plataforma se ha usado la base de datos relacional corporativa basada en Oracle 11 y Virtuoso 7 como almacén RDF. Para que la comparativa sea real, en el mismo hardware se ha instalado un Oracle 11 con una copia de seguridad del sistema en producción y Virtuoso 7. El servidor que se ha usado tiene 8GB RAM y un procesador Intel Core I7-3610QM a 2.30 GHz.

En una evaluación inicial del sistema se ha medido el tiempo necesario para realizar la transformación de los datos desde el modelo relacional al modelo semántico. La migración de los datos tardó 1 hora y 18 minutos. El almacén de datos semántico superó los 101 millones de tripletas, por lo que al igual que pasó son SECARE, se tuvo que aumentar la memoria RAM que usa Virtuoso a 2GB.

Al tener dos modelos equivalentes, uno en el almacén semántico y otro en un modelo relacional tradicional, se ha optado por hacer un test de rendimiento (los resultados pueden verse en la tabla 5.7). En este test, se observa que Oracle es más rápido que Virtuoso en todos los casos, pero con diferencias bastante pequeñas, por lo que, para hacer el análisis y explotación de los datos, los usuarios no notarían prácticamente la diferencia.

Tabla 5.7: SECOLON: Resultados de la comparativa entre SQL y SPARQL

Consulta	Resultado SQL	Tiempo SQL	Resultado SPARQL	Tiempo SPARQL
Recuperar pacientes	322.839	0,060s	322.839	0,189s
Recuperar procesos clínicos	330.421	0,062s	330.421	0,217s
Recuperar diagnósticos colonoscópicos	92.144	0,040s	92.144	0,096s
Recuperar niveles de riesgo	328.955	0,061s	328.955	0,192s
Recuperar pacientes con procesos, diagnósticos y niveles de riesgo	322.839	0,421s	322.839	0,491s
Recuperar pacientes que sean mujeres	132.045	0,131s	132.045	0,231s
Recuperar pacientes que sean mujeres y mayores de 60 años	80.482	0,134s	80.482	0,265s

Al trabajar con datos reales tuvo la ocasión de hacer un test de la bondad del modelo de recomendación basado en Bayes. En este caso, de los más de 300.000 pacientes, se recuperaron 1.228 que habían entrado al programa de cribado y habían tenido más de una visita, es decir, se conoce que ha pasado en el futuro con ellos. Sobre ese conjunto de datos se usan 1.000 pacientes para que el algoritmo aprendiese, a partir de los perfiles semánticos de cada uno de ellos. Los restantes 228 se han usado para usar el recomendador y determinar su capacidad de acierto. De los 228 casos acertó un 76,22 % y falló en un 23,78 %. En la tabla 5.8 se ven los resultados de este análisis. Como se puede observar, el algoritmo se desvía bastante cuando el paciente no tiene riesgo, o tiene riesgo bajo. Sin embargo, para el riesgo intermedio y para el alto, el modelo predictivo presenta grandes tasas de acierto, con curvas de ROC (método estadístico para medir el rendimiento o la efectividad de un predictor [205]) que siempre están por encima del 90 %. Este predictor, comparado con otros estudios como [206; 207; 208], ofrece unos resultados de gran valor.

Tabla 5.8: Resultados del modelo de recomendación

Recomendación	Verdaderos positivos	Falsos negativos	Curva de ROC
Sin patología	52 %	38 %	68 %
Bajo riesgo	41 %	16 %	77 %
Riesgo intermedio	69 %	1,2 %	95 %
Alto riesgo	88 %	18 %	92 %
Positivo para cáncer	71 %	9,5 %	92 %

5.5 CRD Semántico Proyecto NELA

5.5.1 Introducción

El asma es la enfermedad crónica más frecuente en niños y niñas. En las últimas décadas, el asma y las enfermedades alérgicas aparecen con más frecuencia durante la infancia. [209] Diversos factores a los que se está expuestos antes de nacer y durante los primeros años de vida pueden influir en su aparición. Entre estos factores, el papel de la nutrición en estas etapas tempranas de la vida puede ser fundamental [210].

El objetivo del Proyecto NELA [181] es investigar el impacto de la nutrición durante el embarazo y los primeros años de vida sobre la salud de los niños y niñas de la Región de Murcia. NELA está especialmente dirigido a conocer el papel que juega la nutrición durante el embarazo y en los primeros años de vida en la salud respiratoria durante la infancia, así como en el origen del asma y las alergias.

En este proyecto se usa el modelo de IN semántica para generar cuadernos de recogida de datos para dos tipos de pacientes: mujeres embarazadas y bebés. Además, debido a que se recogen muestras y se almacenan en un biobanco (almacén de muestras biológicas) también se muestran los principales resultados de la interoperabilidad semántica entre esos datos.

5.5.2 Metodología y Herramientas

El desarrollo de esta plataforma se ha basado en la solución de CRD semántico descrita en la sección 4.6.4. En la figura 5.27 se puede ver la arquitectura de la plataforma.

En este caso sólo se dispone una fuente de datos externa, el sistema de información del Biobanco donde se almacenan las muestras de los pacientes reclutados. Para transformar estos datos se emplea el motor de transformación SWIT. El resto de la plataforma se basa en la definición y explotación de cuestionarios semánticos.

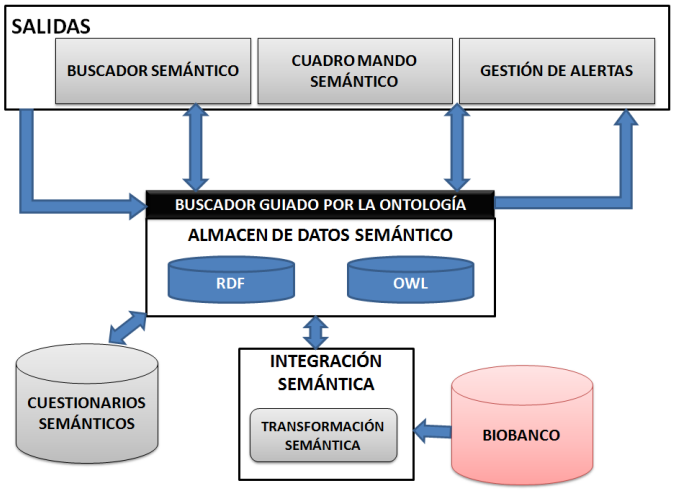


Figura 5.27: Arquitectura CRD Proyecto NELA

Gracias al modelado semántico, los usuarios de la plataforma dispondrán de un buscador semántico, un cuadro de mando semántico y la posibilidad de gestionar alertas del proceso de reclutamiento del proyecto.

5.5.3 Resultados

5.5.3.1 Plataforma funcional

A continuación se describen las principales herramientas de esta plataforma[15]:

- Definición de cuestionarios. En la figura 5.28 se puede ver un extracto de uno de los cuestionarios definidos para el proyecto.
- Definición de protocolos para el estudio. En la figura 5.29 se puede ver cómo los usuarios pueden definir cada una de las visitas o ciclos que se realizan sobre cada paciente reclutado. Además se pueden definir los cuestionarios que se van a recoger en cada uno de esos ciclos.
- Ejecución de cuestionarios. Para cada uno de los pacientes reclutados se debe indicar el protocolo. Por ejemplo, en el caso del proyecto NELA se tienen dos protocolos: uno para embarazadas y otro para los niños de éstas. En la figura 5.30 se pueden ver las visitas de un paciente determinado y los cuestionarios rellenados en cada fase. En la figura 5.31 se observa un extracto de un cuestionario cumplimentado.

[15]http://www.nela.imib.es/

Nombre del campo	Tipo	Orden	+ ★
Dirección (Calle, nº, piso y puerta)	Texto largo	0	Opciones
Municipio	Texto	1	Opciones
1. Fecha visita s20 (_ _ / _ _ / _ _ _ _)	Sólo fecha	2	Opciones
2. Fecha última regla (_ _ / _ _ / _ _ _ _)	Sólo fecha	3	Opciones
3. Código postal del domicilio de la embarazada	Número entero positivo	4	Opciones
4. ¿En qué zona del municipio vive usted?	Seleccionable	5	Opciones
5. Estado civil de la embarazada actualmente	Seleccionable	6	Opciones
6. Edad	Número	7	Opciones

Figura 5.28: Extracto definición de un cuestionario

- Panel de explotación. Desde el panel de explotación los usuarios pueden definir cuadros de mando y alertas sobre los datos del sistema. En la figura 5.32 se observa cómo se puede usar ODS para generar un cuadro de mando concreto.

5.5.3.2 Evaluación

Esta plataforma está en producción desde febrero de 2015. Desde entonces se han creado cuatro cuestionarios que han sido recogidos en dos visitas a cada paciente. En esos cuestionarios se recogen un total de 348 variables.

A continuación se destacan los datos más importantes en el proceso de gestión del CRD:

- Se han registrado 98 pacientes embarazadas.
- Se han registrado 76 cuestionarios completos para esas pacientes embarazadas.
- Se han definido tres cuadros de mando personalizados para hacer el seguimiento del reclutamiento.
- Se han definido alertas de nueva visita para cada uno de los pacientes reclutados.

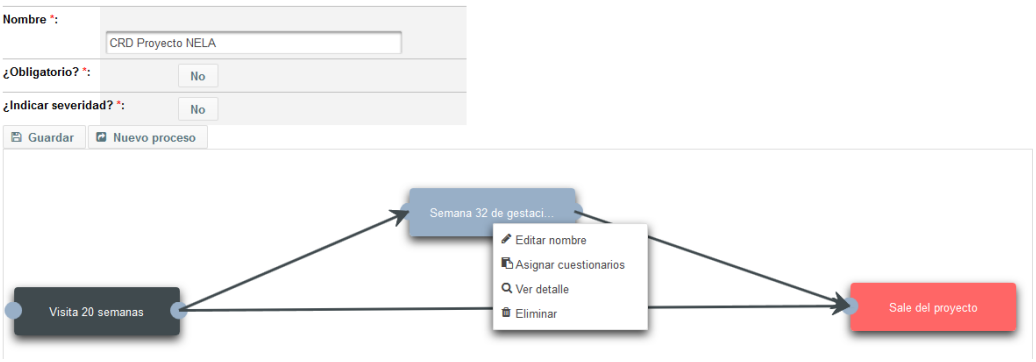


Figura 5.29: Configuración de un protocolo de un paciente del proyecto NELA

- El número total de tripletas en el almacén semántico es de 32.603.

El software ha sido usado por tres usuarios. Su principal aportación consiste en que sin usar ningún tipo de software adicional y sin la necesidad de disponer de personal especializado en TIC, los gestores de los datos del proyecto han sido capaces de configurar, poner en marcha y explotar un CRD para su proyecto de investigación.

5.6 Anotador semántico proyecto EUCOMMTOOLS

5.6.1 Introducción

El proyecto EUCOMM (Herramientas para la anotación funcional del genoma del ratón) [182] es un proyecto europeo miembro del IKMC (*International Knockout Mouse Consortium*). Su objetivo principal es generar una base de datos de mutaciones de los genes que codifican proteínas.

En este trabajo, los miembros del equipo de investigación encargados del paquete de trabajo 6, consistente en anotar la expresión génica, disponían de una bateria de imágenes provenientes de un microscopio confocal de tejido cerebral de ratones de 14 y de 56 días. Entre esos conjuntos también hay una población a la que se le inyecta "Tamoxifen" (principio activo que bloquea algunos tipos de cáncer) y otra población de control. La anotación de expresión se realiza sobre 44 genes.

El proceso de anotación consiste en usar las imágenes para describir si un determinado gen se expresa o no en un determinado ratón con unas condicio-

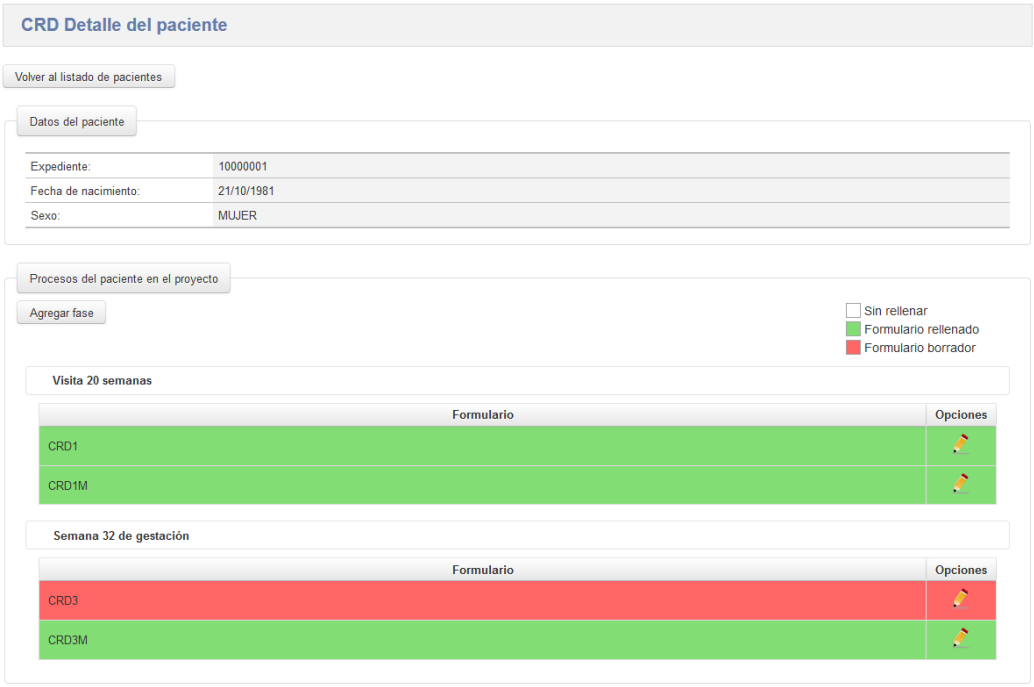


Figura 5.30: Protocolo concreto de un paciente

nes específicas. Un dato muy importante para realizar la anotación es en qué localización anatómica se expresa. Para esas anotaciones se han modelado ontológicamente la anatomía del cerebro y del cuerpo de un ratón.

5.6.2 Metodología y Herramientas

El desarrollo de esta plataforma[16] está basado en la solución de IN semántica en contenidos multimedia propuesta en la sección 4.6.5. En la figura 5.33 se puede ver la arquitectura propuesta. Como se observa, el módulo de integración de la información tiene dos fuentes de datos externas. La primera consiste en una serie de imágenes microscópicas de cada uno de los ratones del proyecto. Estas imágenes se han anotado usando el anotador manual. La segunda fuente de datos son los ratones y sus características genéricas. Estos datos serán transformados haciendo uso de SWIT.

En el almacén de datos semánticos se dispone de tres ontologías: (1) modela los ratones, (2) representa el cerebro del ratón y (3) modela el cuerpo del ratón. Una vez que se tienen las fuentes de datos integradas, el usuario

[16] http://www.imib.es/AnotadorWeb/

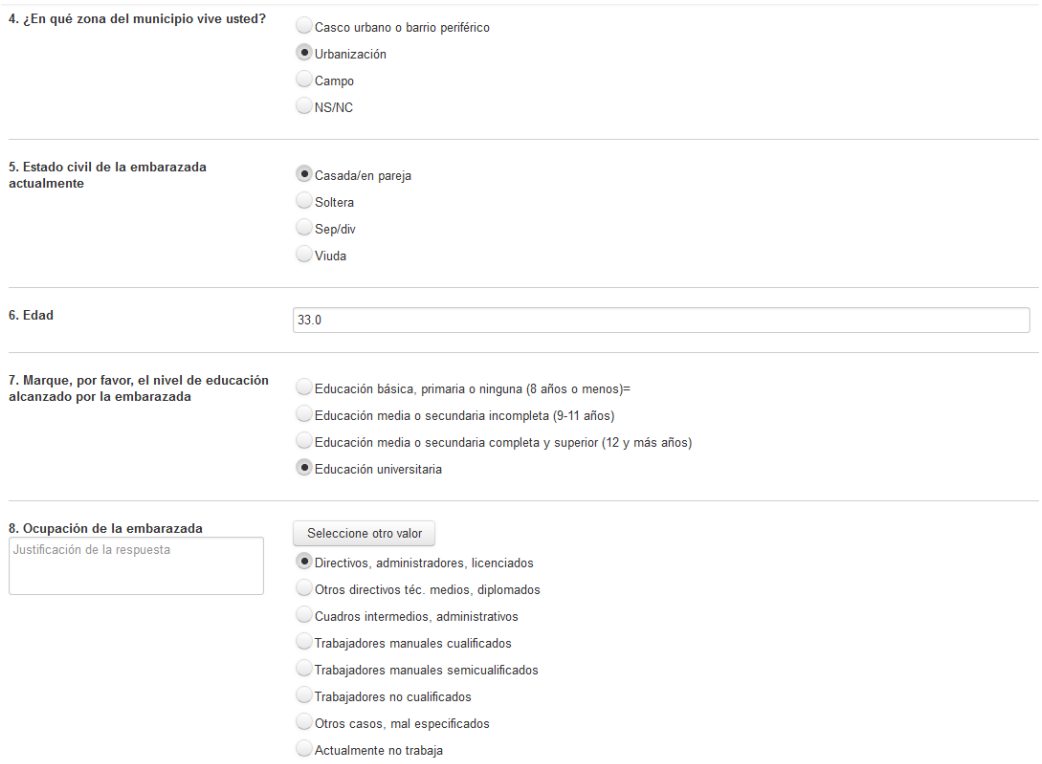


Figura 5.31: Extracto ejecución de un cuestionario

podrá definir los perfiles semánticos para los casos de análisis. En este caso se han definido dos perfiles. El primero tiene las características genéticas del ratón y su anotación con respecto a la ontología del cerebro. El segundo es similar a excepción de que en vez de la ontología del cerebro se usa la del cuerpo. Estos dos perfiles permitirán la comparación de las similitudes y diferencias entre los individuos.

En la parte de explotación, los usuarios disponen de un buscador semántico, un generador de cuadros de mando semánticos a partir de las anotaciones de cada individuo, y un recomendador de anotaciones a partir de las similitudes entre individuos.

5.6.3 Resultados

5.6.3.1 Ontologías usadas

Las ontologías que se han usado para llevar a cabo este trabajo han sido proporcionadas por el equipo investigador. El formato de esas ontologías

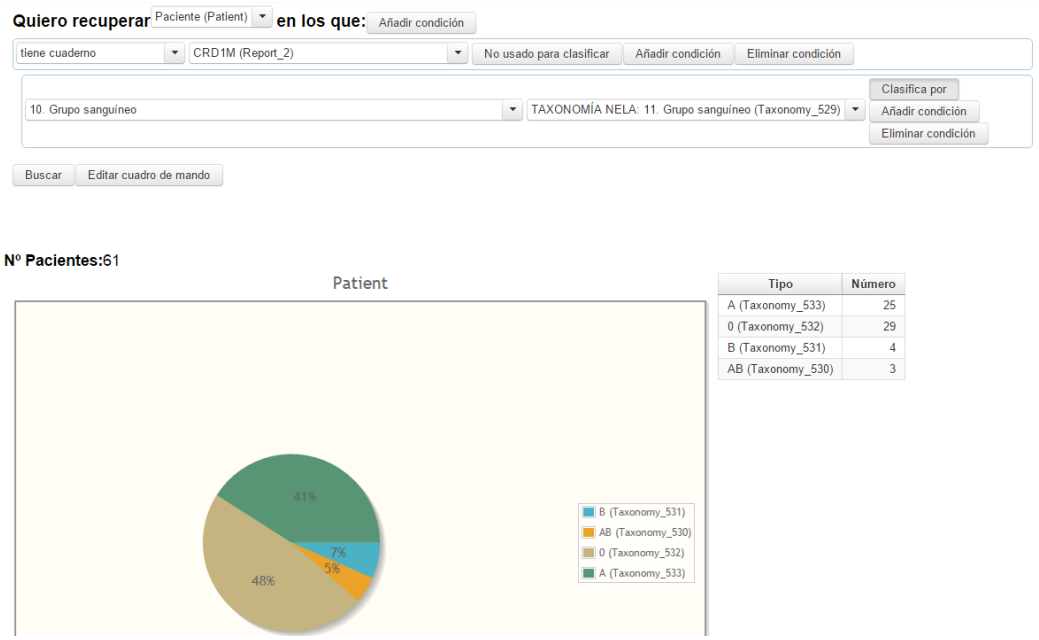


Figura 5.32: Cuadro de mando NELA

eran hojas de cálculo que fueron mapeadas con SWIT para generar el OWL necesario para anotar. Se parte de dos hojas de cálculo en las que se representa la anatomía del cerebro y la del cuerpo de un ratón. A continuación se puede ver la estructura principal de ambas ontologías:

1. El concepto *Cell_ Type* describe los tipos de célula que se encuentran en el cerebro o en el cuerpo del animal.

2. El concepto *Anatomic_ Structure* describe la localización anatómica donde se encuentran esos tipos de célula. Estos conceptos son una jerarquía de elementos ontológicos que describen que una determinada zona anatómica se encuentra dentro de otra zona mayor y así sucesivamente hasta llegar a los conceptos cerebro y cuerpo.

La ontología del cerebro del ratón tiene 78 tipos diferentes de células y 11.361 conceptos que modelan una localización anatómica. La ontología del cuerpo del ratón tiene 187 tipos diferentes de células y 5.112 conceptos que modelan una localización anatómica. Entre las dos ontologías se han definido 32.944 axiomas ontológicos.

Además de estas dos ontologías, el almacén semántico necesita de una ontología principal que integre las dos anteriores. Por este motivo se ha desarrollado una ontología que modela los ratones y su modificación genética.

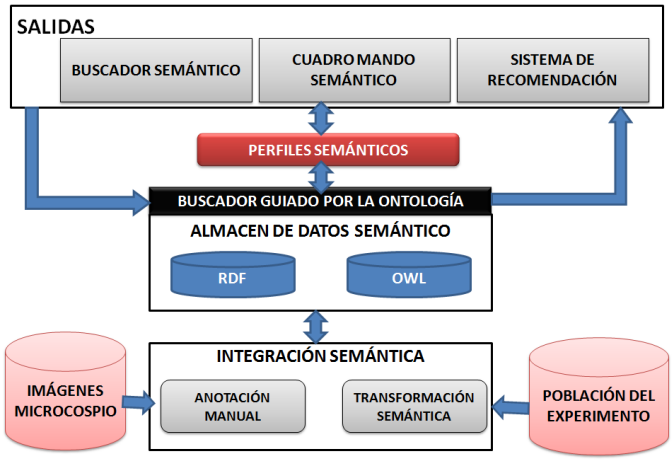


Figura 5.33: Arquitectura del anotador de proyecto EUCOMM-Tools

La clase *Mouse* define al ratón y tiene como propiedades su sexo, su tiempo de vida y un número de expediente. Además se relaciona con otras clases como "gen modificado" (representado por la clase *Gene*), "estadío" (representado por la clase *Stage*) y "sustancia suministrada" (representada por la clase *Injection*).

5.6.3.2 Plataforma funcional

En la figura 5.34 se puede ver el módulo de anotación. Como se observa, para buscar los elementos ontológicos se tienen dos buscadores, uno por tipo de célula y otro por localización anatómica. Como se aprecia en la figura, se puede anotar todo un conjunto de células, así como eliminar todas las anotaciones realizadas hasta ahora. Cuando el usuario está anotando un ratón, la aplicación va comprobando a qué individuos ya anotados se parece más. A partir de más de N (parámetro configurable) anotaciones, que compartan la expresión génica en un mismo tipo de célula en una localización anatómica concreta, la aplicación indica a la persona que anota si quiere reutilizar parte o toda la anotación de otro individuo.

En la figura 5.35 se puede ver el módulo de cuadros de mando comparativos entre ratones. Se pueden generar cuadros de mando agregados de la comparación de las anotaciones entre varios ratones seleccionados. Directamente genera dos cuadros de mando, uno para la ontología del cerebro y otro para la ontología del cuerpo.

En la figura 5.36 se puede ver el módulo de similitud entre ratones. Este servicio diferencia las anotaciones iguales de las que son diferentes por cada

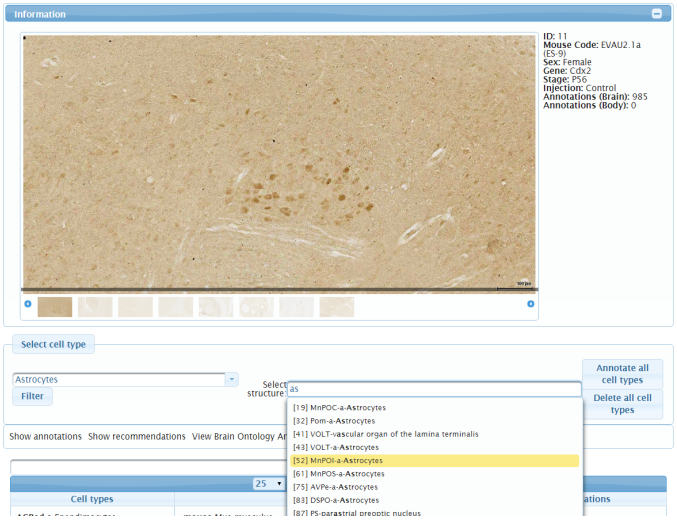


Figura 5.34: Anotador EUCOMM-Tools

individuo comparado.

5.6.3.3 Evaluación

La plataforma se puso en marcha a finales de 2013 y el proceso de anotación terminará a finales de 2015. A continuación se muestran los principales resultados obtenidos:

1. Se han registrado 137 ratones.
2. Se han registrado 964 imágenes asociadas a esos ratones.
3. Se han generado 110.570 anotaciones semánticas.
4. De las 110.570 anotaciones, 66.144 se han generado clonadas del motor de recomendación.
5. Se han generado 36 grupos de imágenes cuya similitud semántica está por encima del 90 %.

La plataforma ha sido usada en modo anotación por tres investigadores con un buen grado de satisfacción. Además, también ha sido consultado el módulo de explotación para ver la información anotada por el resto de participantes en el proyecto europeo. En total, el módulo de explotación ha recibido una media de 98 visitas mensuales, siendo mayores los accesos al módulo de

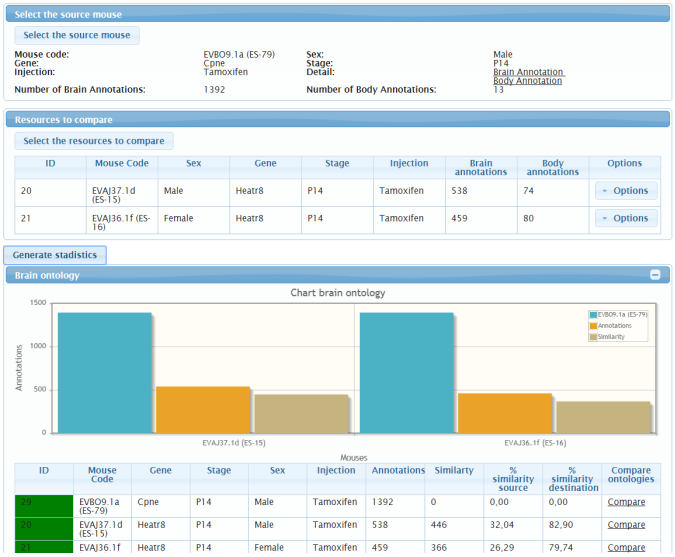


Figura 5.35: Cuadro de mando EUCOMM-Tools

búsqueda (88 %) que al de evaluación de similitud, aunque conforme vaya llegando la finalización del proyecto se espera que ese valor se equilibre ya que necesitan la función de similitud para extraer las conclusiones pertinentes.

Aún así, el hecho de que un 60 % de todas las anotaciones se hayan generado a partir del recomendador y que hayan 36 grupos de ratones cuya similitud semántica esté por encima del 90 % ha ayudado a los investigadores del proyecto a observar que sus hipótesis eran ciertas y que los genes se están expresando del mismo modo para los mismos tipos de experimentos.

	Source	Destination
Mouse code:	EVBO9.1a (ES-79)	EVAJ37.1d (ES-15)
Sex:	Male	Male
Gene:	Cpne	Heatr8
Stage:	P14	P14
Injection:	Tamoxifen	Tamoxifen
Brain annotations:	1392	538
Body annotations:	13	74

Export

Equal annotations

Number of annotations: 446

Cell types	Structure
La-pn- Projection neurons	**mouse**-Mus musculus **NP**-neural plate **F**-forebrain **SP**-secondary prosencephalon **CSP**-caudal secondary prosencephalon **Tel**-telencephalic vesicle **TelA**-alar plate of evaginated telencephalic vesicle **Pall**-pallium **VPall**-ventral pallium **VAP**-ventropallial amygdalopiriform area **VAPm**-mantle zone of VAP

Unique annotations in EVBO9.1a (ES-79)

Number of annotations: 946

Cell types	Structure
Apir-pn- Projection neurons	**mouse**-Mus musculus **NP**-neural plate **F**-forebrain **SP**-secondary prosencephalon **CSP**-caudal secondary prosencephalon **Tel**-telencephalic vesicle **TelA**-alar plate of evaginated telencephalic vesicle **LPall**-Lpall **AOB**-lateral pallium **LAP**-lateropallial amygdalopiriform area **LAPm**-mantle zone of LAP

Unique annotations in EVAJ37.1d (ES-15)

Number of annotations: 92

Cell types	Structure
r1VPCRt-pn- Projection neurons	**mouse**-Mus musculus **NP**-neural plate **H**-hindbrain **PPH**-prepontine hindbrain **r1**-rhombomere 1 **r1A**-r1 alar plate **r1Lim**-liminal part of alar r1 **r1Limm**-mantle zone of r1Lim **r1Limi**-intermediate stratum of r1Lim **r1VPCRt**-r1 part of ventral parvicellular reticular formation **r1VPCRt-n**-Neurons

Figura 5.36: Vista de similitud EUCOMM-Tools

Bloque III

Discusión, conclusiones y vías futuras

Capítulo 6

Discusión

En este capítulo se discuten las diferentes metodologías y herramientas que se han descrito en el bloque anterior. Para diferenciar los niveles en los que se ubica esta solución, este capítulo se divide en siete secciones. En las tres primeras se discute el alcance y las limitaciones de la metodología y herramientas propuestas diferenciando por las tres capas que tiene una solución de IN: integración, entrega de información y análisis. En la cuarta se discuten las diferentes soluciones de IN que se han desarrollado siguiendo los enfoques anteriores. En la quinta se realiza una comparativa con otras soluciones semánticas de IN. En la sexta se compara la propuesta con una solución comercial. Por último, se describen las expresiones de interés en el uso de esta propuesta por diferentes organismos.

6.1 Discusión: nivel de integración

Dentro de las trece características fundamentales de las soluciones de IN (véase el apartado 2.1) cuatro se encuentran dentro del nivel de integración. A continuación se revisan esas cuatro características sobre el modelo semántico equivalente propuesto:

- **Infraestructura para la IN.** La infraestructura de IN propuesta en esta tesis está compuesta por el repositorio RDF del almacén de datos semántico, y por ODS. El almacén de datos es un punto centralizado de almacenamiento semántico de toda la información, y ODS es la herramienta que permite navegar por él. La conjunción de ambos permite ofrecer un entorno centralizado donde se apoyan el resto de servicios. Para integrar la información en esta infraestructura se emplean diferentes metodologías y herramientas de integración. La principal aportación

de este trabajo en relación a esta característica está en la posibilidad de incorporar información no estructurada gracias a los diferentes mecanismos de anotación.
Ventajas de usar tecnologías semánticas. Gracias a las tecnologías semánticas se puede crear un almacén de datos que no sólo sea procesable por humanos, sino también por agentes software, lo que facilitará la automatización de las tareas de análisis y explotación.

- **Gestión de metadatos.** La gestión de metadatos la aporta el repositorio OWL dentro del almacén de datos semántico. Dicho repositorio proporciona un lenguaje común a la hora de integrar, recuperar y analizar la información. La principal característica del modelo propuesto es que no sólo se emplea un lenguaje común para analizar la información, sino que se usa en todo el ciclo de vida de la solución software. Esto supone una gran ventaja, ya que se puede analizar la información en cada una de sus fases de transformación.
  Ventajas de usar tecnologías semánticas. El uso de modelos ontológicos en la gestión de metadatos es un aspecto reconocido por las soluciones actuales, ya que en muchas de ellas se definen vocabularios comunes para la integración y la explotación de los datos. En este trabajo, el repositorio OWL, además de guiar todos los procesos, tanto de transformación, como de entrega de datos y de explotación, también será útil para aportar un mecanismo a través del cuál se pueda enlazar con recursos externos o facilitar las tareas de comparación entre departamentos u organizaciones.

- **Herramientas de desarrollo.** Gracias a que los datos son almacenados siguiendo los principios de *Linked Data*, es sencillo el desarrollo de servicios que exploten esa información. Además, ODS permite la generación de herramientas de consulta y análisis bajo demanda, sin la necesidad de personal TIC especializado para la puesta en marcha de estos servicios. Otra aportación importante a esta característica es el modelo de cuestionarios semánticos, que permite añadir nueva información, tanto en el repositorio RDF como en el repositorio OWL, generando nuevos metadatos que facilitarán los procesos de gestión y análisis.
  Ventajas de usar tecnologías semánticas. La propia definición de Web Semántica implica que la información pueda procesarse automáticamente por ordenadores. En este sentido, el hecho de disponer de un repositorio RDF siguiendo los principios de *Linked Data* ya facilita la tarea de generar servicios de explotación. Por otro lado, los modelos

relacionales no son adecuados para entornos de gran heterogeneidad, donde el esquema de datos puede variar para cada caso de uso concreto. En este sentido, la Web Semántica también puede ser de utilidad, ya que debido a su flexibilidad a la hora de representar y explotar la información puede adaptarse fácilmente a estos cambios. Prueba de ello es el modelo de cuestionarios semánticos.

- **Colaboración.** La propuesta en el ámbito de la colaboración se basa en que cada usuario de la plataforma pueda evaluar los activos de conocimiento generado y calcular o compartir indicadores. La característica de colaboración se basa en el modelo de evaluación propuesto, que puede usar tres métodos: evaluación 360º, definición de indicadores y cuestionarios semánticos. Es importante destacar que la información generada en la colaboración retroalimenta al sistema, es decir, se convierte en nueva información semántica que puede ser explotada usando todos los servicios disponibles.
  Ventajas de usar tecnologías semánticas. La integración de la Web 2.0 y la Web Semántica ha sido y sigue siendo un campo de estudio muy activo. La posibilidad de clasificar, etiquetar y posteriormente explotar los contenidos que se generan en aplicaciones sociales supone un reto para la IN. En el modelo propuesto se usan redes sociales semánticas para poder generar modelos de evaluación 360º sobre cualquier recurso del dominio. Estas evaluaciones permitirán generar entornos de confianza entre los recursos, poniendo en valor a aquéllos que están más preparados para realizar una determinada tarea. Otra ventaja de usar tecnologías semánticas es la posibilidad de usar esas evaluaciones para generar enlaces con fuentes de datos extenas o comparar recursos entre sí.

A continuación se discutirá sobre los modelos propuestos de anotación semántica, cuestionarios semánticos y evaluación.

6.1.1 Modelo de anotación semántica

En la sección 4.1.3 se han definido los modelos de anotación semántica que soporta esta propuesta:

- Anotación manual.
- Anotación automática.
- Anotación semiautomática.

- Etiquetado semántico.

A continuación se describen cuáles son las principales ventajas y desventajas de cada uno de los modelos de anotación.

La principal ventaja del modelo automático consiste en que sin la necesidad de hacer ningún trabajo adicional se puede integrar la información generada a partir del procesamiento de cualquier entidad de negocio descrita en texto abierto. Este proceso presenta dos desventajas importantes: (1) el proceso de anotación automático no anotará el cien por cien del texto y generará, posiblemente, anotaciones que no sean correctas, es decir, falsos positivos; y (2) este proceso no se puede aplicar a la anotación de información como imágenes o cualquier contenido multimedia que no tenga texto.

En el caso de la anotación manual, la principal ventaja es que la persona que está realizando la anotación es un ser humano y estará, por tanto, capacitado para entender perfectamente a qué se refiere el texto con el que está trabajando. Además, es muy útil para anotar imágenes y contenido multimedia. La principal desventaja consiste en que es un proceso muy tedioso y que en organizaciones muy grandes puede llegar a ser una tarea irrealizable debido al gran volumen de información de este tipo que generan. Otra desventaja importante es que, en dominios muy específicos, los anotadores deben ser expertos en esa materia.

La anotación semiautomática permite que el proceso de anotación no sea tan tedioso, ya que el usuario sólo tiene que validar los resultados de la anotación automática. Gracias a este proceso de validación se evita que se produzcan falsos positivos.

El etiquetado semántico puede ser usado con cualquiera del resto de modelos. Esta herramienta permite describir el texto usando palabras clave que realmente son entidades de las ontología del dominio almacenadas en el repositorio OWL. Esta herramienta usa los resultados de una consulta sobre el repositorio semántico para generar una gran cantidad de anotaciones con cada uno de ellos en un único paso. Esta herramienta es muy útil para anotar entidades de negocio que se asemejan entre sí, facilitando que una vez que se ha anotado unade ellas se pueda reusar lo que sea de interés fácilmente con el uso de SPARQL.

Dependiendo de cada problema se puede elegir el enfoque que se considere más apropiado o una combinación de los mismos.

6.1.2 Cuestionarios semánticos

Esta propuesta de generación de información semántica es muy útil en entornos en los que los modelos de datos tienen un alto componente de variabili-

dad. En estos casos, las compañías pueden hacer uso de sus ERP (algunos de ellos permiten añadir información personalizada [168]) o generar aplicaciones propias que solventen problemas concretos. En el primer caso, la información nueva podrá gestionarse desde el propio sistema, aunque la carencia de modelos semánticos que indiquen el significado que tiene esa información en las líneas de negocio de la organización impedirá su explotación automatizada. En el segundo de los casos se estará aumentando la heterogeneidad de los sistemas de información de la organización, y habrá que hacer un nuevo esfuerzo para volver a integrar esos datos en el modelo. En este caso, habrá que usar modelos de explotación específicos desarrollados a demanda.

Con el enfoque propuesto, los usuarios podrán generar modelos de datos específicos a demanda. Estos modelos estarán integrados de origen en el almacén de datos semántico, ya que añadirán información a recursos ya existentes. Sobre esos cuestionarios se generarán automáticamente aplicaciones que permitirán su gestión y su explotación. Además, al ser información del propio almacén de datos semántico, los usuarios tendrán a su disposición la posibilidad de usar toda la capa de explotación del marco de trabajo propuesto de un modo totalmente transparente.

La principal limitación de esta solución está en cómo se definen los cuestionarios. Este mecanismo no aprovecha toda la potencia de OWL para definir cualquier tipo de regla, sino que está acotado a la definición de rangos y cardinalidades específicas. La rigidez de su modelo ontológico tampoco permite que se puedan reutilizar ontologías modeladas para dominios concretos que no hayan sido definidas siguiendo su especificación. Por ejemplo, si un usuario define una ontología que permite ampliar el diagnóstico de un paciente oncológico con la estadificación TNM, debería de adaptar esa ontología al modelo de entidades y relaciones propuesto, no podría usarse directamente.

6.1.3 Modelos de evaluación del conocimiento

Cada criterio de evaluación puede evaluarse usando uno o varios modelos de evaluación de los que se han definido anteriormente, aunque hay casos concretos donde un determinado modelo es más aconsejable que otro.

El modelo de evaluación basado en indicadores semánticos es útil cuando lo que se quiere evaluar se encuentra implícito en el trabajo diario que se registra en el sistema de información. Por ejemplo, si se quiere evaluar la eficiencia en el pago a proveedores, fácilmente se puede identificar el tiempo medio de pago como indicador que indica si esa tarea se hace mejor o peor. La principal ventaja de este tipo de indicador es que es real y objetivo. Sin embargo, también tiene ciertos inconvenientes. Se podría pensar que, si los

pagos se están retrasando, se debe a que el departamento de tesorería de la organización no está haciendo correctamente su trabajo. La pregunta que se debe hacer es ¿por qué?. Puede ser porque el personal de ese departamento no está lo suficientemente formado, no está motivado en su trabajo, no tienen el equipo que necesitan, no hay buena comunicación con el banco, etc. Responder a estas cuestiones no es posible con este modelo de evaluación.

En este tipo de situaciones es donde es importante el resto de evaluaciones. Por ejemplo, para evaluar el conocimiento de una persona, los cuestionarios pueden ser una herramienta de gran valor, ya que permitirán configurar exámenes tipo test para saber cuál es el nivel de conocimiento en el manejo de una herramienta o en el tratamiento de un proceso. También va a permitir realizar encuestas que puedan ser integradas en el modelo semántico sobre clima laboral, satisfacción del puesto de trabajo, etc.

¿Dónde podrían ser útiles los modelos de evaluación 360º? Este tipo de evaluación puede usarse en combinación con los anteriores para aportar información adicional, pero tiene un gran valor a la hora de evaluar aspectos que no son tangibles como ética laboral, resolución de problemas, etc. También puede ser muy útil para evaluar la usabilidad de una aplicación informática o de un equipo. Esta evaluación es más subjetiva, ya que son los propias personas las que responden según su opinión a estos criterios. El modelo de evaluación 360º también es útil para identificar, por ejemplo, que una determinada persona, aunque sí tiene el conocimiento necesario, no lo está aplicando en su vida profesional.

Se considera que el modelo mixto es la evaluación del conocimiento que más se acerqua a la realidad, ayudando a saber qué esta pasando y cuáles son los posibles puntos débiles que se deben mejorar y cuáles las fortalezas a potenciar.

6.2 Discusión: nivel de entrega de información

A nivel de entrega de información, dentro de las trece características fundamentales (véase el apartado 2.1) aparecen cinco. A continuación se revisan esas cinco características sobre el modelo semántico equivalente:

- **Generación de informes.** Esta propuesta aporta dos soluciones en el ámbito de esta característica. La primera consiste en usar ODS para hacer los filtros que se consideren necesarios sobre el modelo de datos y representarlos en una página Web. La herramienta de generación de informes permite la inclusión de gráficas anidadas a las tablas de datos. La segunda solución consiste en la generación de alertas. Al igual

que en el caso anterior se emplea ODS para construir una consulta que devuelve una serie de resultados. Si los valores agregados de ese resultado no cumplen un determinado criterio se genera un informe de alerta que consiste en un correo electrónico con el nombre de la descripción de la alerta y los recursos que están en esa situación.

- **Cuadros de mando.** Como ya se ha comentado, ODS está preparado para la generación de cuadros de mando personalizados. El usuario puede filtrar por todos los conceptos de la ontología y representarlos gráficamente o en formato de tabla. Esas representaciones pueden almacenarse para reproducirse posteriormente cada cierto tiempo o bajo demanda por el usuario.

- **Generación de consultas *ad hoc*.** ODS es una herramienta fundamental para la generación de este tipo de consultas. Además, en muchas soluciones actuales de IN ya usan una capa de metamodelado para este tipo de herramientas, por lo que se han identificado a las tecnologías semánticas como el entorno ideal para la implementación de servicios como éste [5].

- **Integración con Microsoft Office.** No se ha aportado ninguna solución a esta característica, ya que el uso de tecnologías semánticas no iba a mejorar esa integración.

- **Buscadores avanzados.** ODS podrá ser usado para el desarrollo específico de este tipo de buscadores. Con ODS se puede describir una consulta y parametrizarla para que el usuario sólo tenga que rellenar los campos de búsqueda que se decidan, pudiendo poner en marcha infinitos buscadores avanzados por tantos criterios como se deseen sin la ayuda de personal TIC especializado.

En este nivel, la ventaja de usar tecnologías semánticas radica en el hecho de cómo se modela el conocimiento ontológicamente y cómo en una base de datos relacional. Conceptualmente, un ser humano es capaz de comprender mejor que una persona tiene un listado de direcciones de contacto, que en cada dirección de contacto hay un campo que indica a qué persona corresponde (equivalencia en un modelo relacional). El modelado semántico se parece mucho más al razonamiento humano que el relacional. Por ese motivo, formular consultas a ese tipo de almacenes también es mucho más intuitivo. Todos los servicios de entrega de la información usan ODS como motor de búsqueda. ODS permite generar consultas SPARQL guiadas por la ontología sin necesidad de tener conocimientos informáticos avanzados. La representación

semántica de las propias consultas hace que éstas se puedan serializar para ser reutilizadas en otros ámbitos o parametrizadas para generar buscadores semánticos avanzados.

6.3 Discusión: nivel de análisis

A nivel de análisis, dentro de las trece características esenciales (véase el apartado 2.1) se encuentran cuatro. A continuación se revisan esas cuatro características sobre el modelo semántico equivalente:

- **Procesamiento analítico en línea (OLAP).** La alternativa a OLAP en el enfoque propuesto son los perfiles semánticos. Gracias a la potencia de las ontologías para generar nuevas representaciones de datos, los perfiles semánticos permitirán generar "cachés" de datos que modelen el resumen de la información que se quiere analizar. La definición de estos perfiles vuelve a sustentarse sobre ODS, ya que se definen como consultas SPARQL que devuelven los datos concretos con los que se va a trabajar. Ese conjunto de datos puede almacenarse como una simple consulta o como un nuevo concepto ontológico de tal forma que esas instancias pasen a ser individuos de éste. Uno de los problemas que tiene el perfil semántico es que genera información redundante en el almacén de información. Por ese motivo habrá que decidir en qué situaciones es mejor usar un enfoque u otro. Realmente, lo que habrá que decidir es si el aumento en el volumen de datos supone una mejora sustancial en el funcionamiento del sistema que palie ese aumento de tamaño. Por ejemplo, puede que para calcular un indicador puntual no sea necesario definir los perfiles semánticos, pero en el caso de un indicador que se calcule mensualmente, sí que se necesitará que ese cálculo sea lo más eficiente posible. Otra limitación de los perfiles semánticos es que cuando se actualiza la información de las fuentes de datos, ellos no se actualizan. Para evitar este problema, las consultas que generan los perfiles se almacenan y se marcan para que se calculen de nuevo cada vez que se carga la información.
  Ventajas de usar tecnologías semánticas. En este caso, las tecnologías semánticas dan una nueva perspectiva del procesamiento analítico en línea u OLAP. Gracias a las ontologías se pueden generar diferentes representaciones de la información, lo que permitirá transformar los datos en modelos de información adaptados para un análisis concreto. Ejemplos de esto se pueden encontrar en las vistas de los modelos relacionales, aunque éstas son mucho más planas, ya que se trabaja con

una tabla de los datos. Gracias a la Web Semántica, la representación puede ser mucho más rica, incluyendo diferentes niveles de agregación de la información.

- **Visualización interactiva.** El servicio de generación de cuadros de mando permite la navegación por la representación gráfica, es decir, cuando el usuario quiere ver de dónde vienen los datos de una parte del gráfico, la aplicación es capaz de generar la consulta SPARQL que los recupera. ODS almacena las variables por las que el usuario ha hecho la agregación, por lo que el modelo sustituye esas variables por el valor real proporcionado por el usuario al hacer clic sobre la representación gráfica.
  Ventajas de usar tecnologías semánticas. Gracias al motor de persistencia de ODS se pueden almacenar consultas SPARQL manteniendo su significado. Es decir, aunque se esté generando una representación gráfica de la información, siempre se pueden cambiar los parámetros a partir de las interacciones de los usuarios para que pueda navegar al origen concreto de una parte de la representación gráfica. Para poder hacer esto mismo en un esquema relacional habría que usar algún tipo de mecanismo para anotar en la consulta cuáles son las alternativas en la agregación.

- **Modelado predictivo y minería de datos.** Para esta característica se ha propuesto el uso del sistema de recomendación. Como ya se ha comentado, este sistema se basa en dos algoritmos: similitud semántica y redes bayesianas. Esto no quiere decir que no se puedan usar otros algoritmos de minería de datos, sino que estos dos son los que mejor se adaptan a un enfoque semántico. El primero porque directamente es un algoritmo relacionado con ontologías, y el segundo porque se basa en la distribución probabilística de un grafo dirigido sin ciclos, que se aproxima mucho al modo en que se estructura una ontología, aunque éstas si pueden tener ciclos. Para solventar este problema se ha definido un método que permite la generación de árboles de redes bayesianas con cada concepto que se puede repetir cíclicamente.
  Ventajas de usar tecnologías semánticas. En esta propuesta se usan algoritmos que están muy ligados a la propia representación semántica, como son las funciones de similitud semántica o las redes bayesianas. Este segundo algoritmo se puede usar fácilmente en modelos relacionales, pero la Web Semántica va a facilitar que la estructura de las redes y del modelo de almacenamiento de datos sea el mismo gracias al uso de perfiles semánticos. Este hecho va a ayudar a que

sean los propios usuarios los que puedan configurar sus propios análisis. En el caso de la función de similitud, no existe una alternativa en el mundo relacional, aunque se podrían definir métricas de similitud entre registros de diferentes tablas. La representación ontológica permite que la similitud entre instancias sea algo inherente al propio modelo, sin necesidad de tener otro tipo de soluciones tecnológicas. Las funciones de similitud semántica son muy útiles para generar modelos de clasificación y recomendación de recursos.

- **Cuadros de mando de indicadores estratégicos.** La combinación de los cuadros de mando con el modelo de evaluación de los activos de conocimiento permite cubrir esta característica. Con el modelo de evaluación se podrían definir cuáles son las métricas cuantitativas y cualitativas deseadas en la organización, y con la generación de cuadros de mando personalizados se podría comprobar la evolución de dichos indicadores. La relación entre los indicadores y la información real permitirá que el modelo asista en la generación de planes estratégicos, gracias a que son fácilmente detectables las fortalezas y debilidades de la organización.
  Ventajas de usar tecnologías semánticas. Usar tecnologías semánticas para evaluar el conocimiento, y emplear ese mismo modelo para anotar recursos, tanto internos como externos, permite generar planes estratégicos, cuyos resultados también enriquecerán la información del almacén semántico, permitiendo que se pueda analizar el impacto de los mismos en la organización. Es importante destacar que, gracias a la representación semántica, se pueden generar planes con diferentes granularidades, es decir, dirigidos a una organización entera o a un departamento concreto.

6.4 Discusión: Plataformas de IN

A continuación se discute sobre las cinco soluciones de Inteligencia de Negocio definidas en la sección 4.6.

6.4.1 Discusión: Red social semántica

La Web 2.0 y la Web Semántica pueden verse como dos métodos diferentes para mejorar la Web actual y superar sus limitaciones. El modelo tradicional de Web, en la que grandes compañías publicaban información y los usuarios eran meros consumidores de ella, ha sido reemplazado por un modelo

más flexible en el que cada usuario produce y consume los contenidos que le interesan. Además, el valor de la Web social consiste en la agregación de muchas contribuciones individuales de muchos usuarios. En este ámbito, las metodologías de Inteligencia de Negocio pueden ayudar a analizar la información y proporcionar servicios de valor añadido a los usuarios de la red. Por otro lado, la Web Semántica proporciona un modo claro de estructurar la información de un sistema de información, reemplazando la Web actual por modelos más formales que permiten a las máquinas procesar la información de forma similar a como lo hace un humano.

Muchos investigadores líderes en este ámbito están de acuerdo en que la combinación de las mejores ideas de la Web 2.0 y la Web Semántica pueden dar lugar a aplicaciones potentes [122; 84; 123]. Esta combinación puede dar lugar a una nueva evolución de la Web denominada web social-semántica [122]: "en una web social-semántica, se pueden representar formalmente partes del conocimiento humano que podrán ser clasificadas y razonadas por herramientas de Web Semántica, pero también podrán ser registradas y mantenidas a través de las técnicas sociales y comunitarias de la web 2.0".

En la sección 4.6.1 se ha definido la arquitectura y los diferentes módulos de una plataforma que permite la explotación de cualquier tipo de red social, permitiendo la incorporación de datos estructurados y no estructurados. La principal debilidad de esta plataforma está en el procedimiento de anotación de los contenidos en lenguaje natural, ya que se puede perder información de gran valor para la explotación del conocimiento que se genera en la red social. Por ese motivo, se recomienda el uso de los cuestionarios semánticos para que se estructuren contenidos que puedan ser de gran valor en la fase de explotación. Prueba de ello, es el uso de los cuestionarios para modelar las carteras financieras de SocialBROKER (ver sección 5.1).

6.4.1.1 SocialBROKER

SocialBROKER (ver sección 5.1) representa un paso importante para alcanzar esa Web social y semántica. Combina una plataforma social con el modelo de Inteligencia de Negocio Semántico que permite clasificar y explotar la información generando servicios de valor añadido para los usuarios. Por ejemplo, se ofrece la generación de búsquedas más precisas y recomendadores más sofisticados que acercan el conocimiento que realmente puede interesar al usuario. Todas estas herramientas se han puesto en marcha en la plataforma SocialBROKER para el dominio financiero.

La principal limitación de este trabajo está asociada con el uso de una única fuente de conocimiento, la información proporcionada por los usuarios.

6.4.2 Discusión: Plataforma para la planificación

Una de las principales funciones de las plataformas de IN es ayudar a la toma de decisiones estratégicas. Para tomar esas decisiones no es suficiente con medir los indicadores de productividad de la empresa, sino que hay que evaluar todos los activos implicados, incluyendo el conocimiento.

En la sección 4.6.2 se ha definido la arquitectura y las herramientas para la generación de planes estratégicos a partir de los resultados y la evaluación de los activos de conocimiento de cualquier organización.

6.4.2.1 Planificación semántica de la formación continuada

Para validar este modelo se ha puesto en marcha una herramienta que permite generar planes de formación personalizados para cada uno de los profesionales de un hospital (véase sección 5.2). En esta plataforma se ha definido un catálogo de criterios de evaluación, se han asociado a recursos humanos, cuyo desempeño se ha evaluado. Los puntos débiles de los profesionales han servido para diseñar planes de formación en base a diversos itinerarios formativos. Esta plataforma ha sido validada durante todo el plazo de análisis, diseño e implementación por varias entidades que realizan formación.

Actualmente está implantada y operativa en el Hospital Clínico Universitario Virgen de la Arrixaca (HCUVA), donde lleva funcionando desde 2013. Durante los tres años que lleva implantada, dos personas han solicitado el cambio de servicio después de su evaluación. Además, los sindicatos realizaron una queja formal para evitar que se evaluara a personal con contratos estables del hospital, por temor a que la evaluación pudiera derivar en una bajada salarial o incluso en el despido. Por ese motivo, la plataforma únicamente se usa para evaluar al personal temporal, aunque desde la Gerencia del hospital se sigue negociando para que entiendan que lo que se persigue es una mejora en los servicios de atención al paciente, y no hacer ningún ajuste laboral.

Una de las limitaciones de este trabajo se encuentra en la anotación de los recursos formativos. Al no disponer de material digitalizado, sólo se pudo usar los epígrafes de cada acción formativa. Para anotar estos contenidos se usaron los propios criterios de evaluación de tal forma que el usuario puede ver los epígrafes y cuáles van a ser las competencias o conocimientos que va a adquirir realizando ese curso.

6.4.3 Discusión: Plataforma para el análisis epidemiológico

El resultado principal de esta línea consiste en el desarrollo de una plataforma basada en Inteligencia de Negocio Semántica que facilita el análisis y la visualización de pacientes cuyas enfermedades han sido registradas en un sistema epidemiológico. Los principales servicios que proporciona esta plataforma son: (1) representación gráfica de la enfermedad de un paciente, (2) representación gráfica de las propiedades agregadas de un grupo de pacientes filtrados con un determinado criterio de búsqueda, y (3) definición de cuadros de mando personalizables para la selección y visualización de la información. El segundo de estos servicios puede ser considerado como una herramienta que ayude a planificar cuáles van a ser los siguientes pasos a seguir en un paciente.

Esta plataforma también proporcionará servicios para que los usuarios puedan usar la potencia y precisión del motor de búsqueda semántico. Otra ventaja importante del modelado semántico del conocimiento es la posibilidad de compartir y comparar información con otros casos clínicos.

Gracias al uso de datos simulados en el primer caso (ver sección 5.3), y de datos reales en el segundo (ver sección 5.4), se ha demostrado la viabilidad de explotar este tipo de aplicaciones usando los modelos y herramientas expuestos en esta tesis. Una comparativa de rendimiento analizando la velocidad de respuesta de consultas relacionales y su equivalente en consultas semánticas ha demostrado que el almacén semántico es equivalente o más rápido en muchas de las situaciones de análisis.

Una de las principales limitaciones de este trabajo ha sido el uso de versiones preliminares de ontologías que modelaban un registro de cáncer y un cribado de cáncer colorrectal. Estas ontologías necesitan ser revisadas y extendidas, aunque han sido lo suficientemente formales para demostrar que la explotación semántica de este tipo de información es posible de forma robusta y escalable.

Otra de las limitaciones está en el proceso de planificación de los servicios. El modelo actual únicamente trabaja con los pacientes ya registrados y no tiene en cuenta la inclusión de nuevos.

6.4.4 Discusión: Cuaderno de recogida de datos semántico

La IN es especialmente útil en entornos donde hay una gran variabilidad de los datos, tal y como ocurre en los estudios de investigación clínica. En

la sección 4.6.4 se ha definido la arquitectura y las herramientas de una plataforma que permite la configuración, gestión y explotación de cuadernos de recogida de datos.

Esta herramienta ha sido validada y testeada en un proyecto real para la creación de una cohorte de pacientes que puedan tener algún tipo de enfermedad respiratoria en los primeros años de vida (véase la sección 5.5).

6.4.5 Discusión: IN semántica en contenidos multimedia

En la sección 4.6.5 se ha definido la arquitectura y los diferentes módulos de una plataforma que permite clasificar semánticamente contenidos multimedia que posteriormente podrán ser explotados a través de diferentes servicios.

En el caso de uso del proyecto EUCOMM TOOLS (véase la sección 5.6) ha servido para demostrar que la anotación manual es una herramienta importante cuando se desea clasificar contenido multimedia. En este trabajo se ha desarrollado una plataforma que permite anotar cualquier tipo de contenido con elementos ontológicos. Además, gracias a las herramientas de inteligencia de negocio semánticas se han podido explotar las anotaciones y generar servicios que aporten valor añadido. La plataforma se ha usado en un entorno real desde principios de 2014 y los resultados han sido muy satisfactorios. El módulo de recomendación ha permitido reusar más de un 60 % de las anotaciones introducidas.

La principal limitación de este caso de uso ha estado en las versiones preliminares de las ontologías. Las ontologías anatómicas del ratón usadas en esta herramienta son únicamente taxonomías que permiten jerarquizar las partes del cuerpo y del cerebro del ratón por un lado, y los tipos de célula del sistema nervioso por otro.

6.4.6 Tabla resumen de las plataformas de IN

En la tabla 6.1 se puede ver un resumen de las plataformas de IN en relación con las metodologías y herramientas que se han usado.

Tabla 6.1: Resumen de las plataformas de IN

	Red Social Semántica	Planificación Semántica	Epidemiología	CRD Semántico	Contenidos Multimedia
Anotador Semántico	✓		✓		✓
Etiquetado semántico	✓				
Criterios de evaluación	✓	✓			
Cuestionarios semánticos	✓	✓		✓	
Evaluación del conocimiento	✓	✓			
Perfiles semánticos	✓	✓	✓	✓	✓
Buscador semántico	✓	✓	✓	✓	✓
Gestor de alertas	✓		✓	✓	
Cuadros de mando	✓	✓	✓	✓	✓
Planificación semántica		✓			
Análisis del impacto		✓			
Recomendador similitud	✓	✓			✓
Recomendador bayesiano			✓		

6.5 Comparativa con soluciones semánticas semejantes

En la sección 2.2.6.1 se describe EU MUSING, un proyecto europeo enfocado a integrar tecnologías de la Web Semántica y del Procesamiento del Lenguaje Natural con metodologías basadas en reglas y enfoques estadísticos para mejorar las capacidades de adquisición de conocimiento y de razonamiento en aplicaciones de IN. Como se puede observar, las tecnologías a emplear en el proyecto son equivalentes a las que se plantean en este trabajo. Sin embargo el enfoque propuesto en esta tesis es distinto.

Los resultados de este proyecto han estado orientados al proceso de ETL, usando tecnologías de la Web Semántica para realizar esa tarea y para conocer si el proceso se ha realizado correctamente usando diferentes modelos

estadísticos. También han conseguido importantes avances en la estandarización semántica del estándar XBRL, y han propuesto varios modelos para poder integrar fuentes de datos externas a la propia compañía. Los entornos donde han aplicado sus resultados son: (1) la gestión de riesgos financieros, (2) la internacionalización de las compañías, y (3) los riesgos de las soluciones basadas en tecnologías de la información y de las comunicaciones. Como puntos débiles, no presentan una solución integrada siguiendo el esquema de [2]. Además, no emplean las tecnologías semánticas para el resto de capas de las soluciones de IN, como son los niveles de consulta de la información y de análisis. Usan sistemas de evaluación para el proceso de integración, pero no para evaluar los activos de conocimiento de la organización, no permitiendo así el uso de entornos sociales que, gracias a sus valoraciones, creen entornos de confianza a partir de los modelos predictivos que se van generando en los procesos de razonamiento o de minería de datos.

El modelo propuesto de anotación automática usando técnicas de Procesamiento del Lenguaje Natural es más sencillo que el propuesto en EU MUSING. Para solucionar este inconveniente, este proceso puede ser complementado con la anotación manual o el etiquetado semántico. Dejando de lado el plano de integración de la información en texto abierto, el proyecto EU MUSING no presenta soluciones para la integración de datos estructurados, ni para contenidos no textuales como los contenidos multimedia. En su arquitectura no presentan ninguna alternativa a OLAP para el análisis de los datos, ni ofrecen herramientas de explotación gráfica como generación de informes, cuadros de mando o buscadores avanzados.

6.6 Comparativa con soluciones de IN comerciales

En este caso se han evaluado las prestaciones que ofrece la plataforma de IN conocida como PENTAHO (http://community.pentaho.com/). En la tabla 6.6 se pueden ver los diferentes servicios de PENTAHO y la alternativa propuesta.

Tabla 6.2: Comparativa PENTAHO - IN Semántica

Servicio ofrecido	**PENTAHO**	**IN Semántica**
Plataforma de análisis analítico	Esta plataforma permite a los usuarios de PENTAHO la posibilidad de acceder a cuadros de mando, informes y a sus modelos predictivos basados en WEKA (es una colección de algoritmos para aprendizaje y minería de datos).	La plataforma propuesta permite la definición de buscadores semánticos avanzados y cuadros de mando semánticos. La principal ventaja de este modelo es que cada usuario puede tener sus propios cuadros de mando gracias a ODS, sin la necesidad de la ayuda de personal TIC. Además, el modelo semántico permite que se pueda navegar por los resultados de la agregación de datos. El modelo predictivo es más limitado que el de PENTAHO, ya que sólo tiene funciones de similitud semántica y redes bayesianas.
Integración de datos	Kettle es la solución de PENTAHO para hacer ETL sobre los orígenes de datos. PENTAHO usa un modelo de integración basado en un almacén de datos centralizado.	En el modelo propuesto no hay un modelo único para integración de la información. Dependiendo del dominio se pueden emplear diferentes modelos de anotación y motores de integración semántica. La principal ventaja es que el enfoque propuesto permite clasificar contenidos en texto abierto, documentos, y cualquier contenido multimedia. En este caso, también se dispone de un almacén de datos centralizado, aunque no es obligatorio.

Tabla 6.2: Comparativa PENTAHO - IN Semántica

Servicio ofrecido	**PENTAHO**	**IN Semántica**
Diseñador de informes	PENTAHO dispone de diseñadores de informes en diversos formatos como CSV, PDF, HTML, Excel y XML.	El marco de trabajo propuesto está bastante más limitado en este sentido. Únicamente diseña informes en HTML, que son exportables a CSV. Al igual que pasaba con el resto de herramientas, los usuarios de la plataforma pueden usar ODS tanto para consultar como para filtrar los datos que va a recuperar. Esto es una característica importante, ya que el usuario usa las mismas interfaces para todos los análisis.
Marketplace	PENTAHO dispone de una tienda virtual donde los desarrolladores pueden publicar nuevos servicios.	En esta propuesta, esa opción no se ha barajado. Sin embargo, el hecho de disponer de los datos siguiendo los principios de *Linked Data*, permite que cualquier desarrollador pueda construir aplicaciones de análisis y explotación de los datos a medida. También, en el caso de que lo que se necesite sea ampliar el almacén central de datos, los cuestionarios semánticos son una buena alternativa para añadir información que posteriormente podrá ser evaluada y analizada en el mismo formato.
Diseñador de agregación	Este servicio permite agregar diferentes tablas del almacén de datos para facilitar el análisis.	ODS cumple perfectamente con esa función sin la necesidad de añadir ningún servicio adicional. El hecho de no usar bases de datos relacionales evita los problemas de rendimiento cuando se unen dos tablas grandes, por lo que el mero hecho de usar RDF para almacenar los datos evita este problema.

Tabla 6.2: Comparativa PENTAHO - IN Semántica

Servicio ofrecido	PENTAHO	IN Semántica
Editor de cubos OLAP	Esta solución se basa en el motor de cubos OLAP de PENTAHO, llamado Mondrian. Este diseñador permite la definición de los cubos y sus diferentes dimensiones.	En la solución propuesta, los usuarios pueden definir perfiles semánticos de cada recurso para facilitar su análisis.
Editor de metadatos	Esta herramienta permite que los usuarios puedan añadir metadatos para definir los modelos físicos de datos del almacén.	La propia estructura de las ontologías permite que se añadan tantos metadatos como sean necesarios.
Big data	PENTAHO dispone de una serie de tecnologías que facilitan el análisis de Big Data. Concretamente el uso de bases de datos cargadas en memoria.	No se ha tenido la oportunidad de evaluar el marco de trabajo en entornos donde se requisiese analizar grandes volúmenes de datos. Por ese motivo, no se ha contemplado esta característica.

Además de esos servicios propios de la Inteligencia de Negocio, PENTAHO dispone de servicios de monitorización, sistemas de seguridad y de permisos, de los que la solución propuesta carece. Sin embargo, PENTAHO no tiene mecanismos para la evaluación de los activos de conocimiento, no permite la extensión del modelo de datos a partir de cuestionarios u otros modelos, y no tiene módulos de planificación ni de análisis del impacto. Otra característica fundamental es que PENTAHO está diseñado para trabajar con fuentes de datos estructuradas, mientras que el enfoque propuesto está preparado para incorporar todo tipo de conocimiento, incluyendo contenidos multimedia.

6.7 Expresiones de interés

Durante 2014 y 2015 se ha tenido la oportunidad de enseñar las soluciones de esta propuesta a diferentes entidades que, en algunos casos, han mostrando su interés en poder usarlo. A continuación se explican las diferentes entidades que apuestan por usar los resultados de esta tesis para dar respuesta a algunas de sus necesidades:

- **Banco Mare Nostrum.** Se interesó en 2014 en usar la plataforma para la planificación semántica. En el contexto de fusión que han sufrido diversas entidades financieras españolas, la formación del personal en nuevos sistemas de gestión y procesos de trabajo se ha convertido un reto importante para estas organizaciones. Estaban interesados en un proyecto parecido al del hospital, donde pudieran evaluar el desempeño de sus trabajadores y, sobre éste, generar planes de formación específicos. El proyecto termina en marzo de 2016, cuando se entregará la plataforma terminada para ponerla en producción.

- **Instituto de Neurociencias de Alicante.** El Instituto de Neurociencias de Alicante dispone de un animalario libre de patógenos. Un cambio en la legislación europea en el protocolo de seguimiento de la experimentación animal hace que se deban registrar los procedimientos que se realizan sobre cada animal. Además, toda la información recogida sobre la severidad de esos procedimientos debe poder exportarse en un formato homogéneo que se envía periódicamente a una comisión ética europea. Debido a la variabilidad que pueden encontrar a la hora de elaborar esos experimentos y recoger los datos, identificaron el CRD semántico como una solución ideal. Actualmente se está terminando el desarrollo para ponerlo en marcha a principios de 2016.

- **Red regional de Biobancos de Murcia.** Gracias al proyecto NELA, la red regional de Biobancos de Murcia pudo ver la potencia del CRD semántico. Actualmente se encuentra en fase de análisis para adaptarlo a su funcionalidad, y que sirva para anotar y analizar los datos de las diferentes muestras que se almacenan.

Además de estas colaboraciones con entidades externas, diversos grupos de investigación están interesados en usar estas soluciones, tanto de análisis epidemiológico, como el CRD y el anotador multimedia. Se han recibido expresiones de interés de cinco grupos de investigación biosanitaria de la Región de Murcia y de otro grupo del Hospital de Elche en Alicante. Los responsables de formación continuada de la Comunidad Autónoma de la Región

de Murcia también se han mostrado interesados en pilotar la plataforma de planificación de la formación.

Capítulo 7

Conclusiones

7.1 Verificación de las hipótesis

En este trabajo se presenta una solución integral para la Inteligencia de Negocio usando tecnologías semánticas. La hipótesis principal de este trabajo ha consistido en el uso de Web Semántica para la construcción de un modelo integral, que permita integrar fuentes de datos heterogéneas, su explotación y que emplee técnicas de inferencia para generar nuevo conocimiento que facilite la toma de decisiones. Esta hipótesis se divide a su vez en sub-hipótesis que quedan demostradas a través de las respuestas a las cuestiones que se han planteado y que se detallan a continuación:

- **Las tecnologías de la Web Semántica permiten integrar la información proveniente de fuentes heterogéneas en un almacén de datos semántico. La integración permitirá asociar datos estructurados y no estructurados, y permitirá identificar criterios de evaluación de los activos de conocimiento implicados.** Comprobar esta hipótesis requiere contestar a las siguientes preguntas:

    1. **¿Cuáles son las metodologías existentes para integrar diferentes orígenes de datos en una fuente común?** En la sección 2.4 se han expuesto las diferentes metodologías para la integración de la información, siendo las más comunes la integración basada en almacén centralizado de datos y la basada en mediadores. En el primer modelo se usan procesos de ETL para extraer, transformar y cargar los datos en esa base de datos centralizada. En el segundo se usan vistas específicas sobre las bases de datos de cada uno de los sistemas operacionales y se define un motor de consultas que permita extraer la información de cada una de esas

vistas, por lo que estas consultas tendrán que adaptarse a cada sistema operacional.

2. **¿Cómo se pueden integrar y explotar contenidos que no estén estructurados?** En la sección 2.1.2 se analizan las limitaciones de las soluciones actuales de IN para explotar datos provenientes de aplicaciones sociales o Web 2.0. Estas limitaciones se basan en que las plataformas actuales del mercado están orientadas a la explotación de datos estructurados, por lo que el resto de conocimiento generado por las organizaciones no puede ser evaluado a no ser que se normalice y clasifique manualmente. Para intentar automatizar este tipo de contenidos se comenta la posibilidad de usar técnicas de minería de textos o incluso de usar terminologías u ontologías.
   En la sección 2.2.2.5 se exponen diferentes procesos para añadir información semántica a contenidos que no tienen por qué estar estructurados. En esta sección se definen los tres principales modelos de anotación: manual, automática y semiautomática. También se definen modelos de etiquetado semántico de contenidos.
   En la sección 4.1.3 se exponen los diferentes métodos que se han implementado para poder integrar y explotar este tipo de contenidos. Para la anotación automática y semiautomática se emplean técnicas de procesamiento del lenguaje natural que permiten anotar los contenidos automáticamente. Este modelo tiene tres problemas: (1) en muchos casos no se consigue anotar el 100 % del contenido, (2) puede generar falsos positivos y (3) no se puede usar para contenidos no textuales como imágenes o vídeos. El modelo de anotación manual propuesto se basa en la definición manual de anotaciones por expertos. Para agilizar este modelo emplea recomendadores basados en modelos de similitud semántica que permitan que un usuario pueda reusar todas o parte de las anotaciones de otro recurso con cierto nivel de semejanza. Por último, se ha propuesto un modelo de etiquetado semántico guiado por la ontología que permite la generación de consultas SPARQL al almacén semántico para anotar con todos los resultados devueltos. Este modelo permite generar un gran número de anotaciones con una sola consulta y también es útil para reutilizar las anotaciones de otros recursos.

3. **¿Cómo se pueden clasificar los activos de conocimiento para que puedan ser medidos y evaluados?** En la sección 4.1.4 se expone un innovador método de anotación que añade in-

formación semántica a los recursos para que éstos puedan ser evaluados. En este modelo se distinguen dos modelos de anotación. En el primero se genera el perfil tipo de un recurso, es decir, cuáles son los criterios de evaluación y los niveles mínimos deseables para que ese recurso desempeñe su función correctamente. En el segundo se realiza una evaluación sobre esos recursos. Posteriormente se pueden analizar las diferencias de cada recurso con el perfil esperado obteniendo como resultado el listado de debilidades y fortalezas de la organización.

4. **¿Qué ventajas tienen las tecnologías de la Web Semántica para integrar diferentes tipos de datos que a su vez son heterogéneos entre sí?** La Web Semántica, a través de las ontologías, permite representar cualquier dominio, ofreciendo un vocabulario común de trabajo. Posteriormente, a partir de los modelos de integración, se pueden mapear esos modelos heterogéneos a este vocabulario común haciendo uso de dos enfoques: GAV y LAV. Con GAV se usa una ontología global y se definen las correspondencias entre ésta y los sistemas operacionales. Con LAV se define una ontología por cada sistema operacional y se harán correspondencias entre las propias ontologías. También existe un enfoque híbrido en el que se usa LAV para cada sistema operacional, pero basándose en una ontología que proporcione un vocabulario común. En la sección 2.4.9.1 se han expuesto diferentes casos de uso de integración de información usando Web Semántica. En SocialBROKER (ver sección 5.1) se ha definido un mecanismo de integración que incorpora datos de diferentes índices bursátiles, de los comentarios de una red social y de carteras financieras. Usando un enfoque GAV se ha logrado integrar con éxito estos orígenes de datos, generando servicios de valor añadido como el cálculo de las plusvalías de las carteras o la recomendación de iniciativas financieras.

- **Es posible usar un modelo semántico para añadir nueva información a recursos de negocio.** Comprobar esta hipótesis requiere contestar a las siguientes preguntas:

  1. **¿Es posible incorporar nuevo conocimiento, no recogido en los orígenes de datos, a las soluciones de Inteligencia de Negocio?** Algunas herramientas de IN disponibles en el mercado permiten añadir nueva información a sus soluciones a partir de marcos de desarrollo y servicios de integración. La principal

desventaja de estos modelos es que la organización necesita de un equipo experto en TIC para poder incorporar esos nuevos datos.

2. **¿Es posible usar la Web Semántica para gestionar, integrar y explotar esa información en una solución de Inteligencia de Negocio?** En la sección 4.2 se define un modelo para la generación de cuestionarios semánticos. Este modelo permite definir cómo, cuándo y bajo qué condiciones se va a ampliar la información de un determinado recurso. Este modelo ha sido implementado y evaluado en diversas soluciones. En SocialBROKER (ver sección 5.1) se han usado para la definición de carteras financieras de los usuarios de la red social. En el proyecto NELA (ver sección 5.5) se han usado para definir un CRD semántico que permite recoger en cada una de las visitas de los pacientes reclutados diferentes cuestionarios definidos por los propios gestores del proyecto.

- **Es posible desarrollar un buscador semántico guiado por la ontología que permita navegar por todo el conocimiento del almacén de datos.** Comprobar esta hipótesis requiere contestar a la siguiente pregunta:

  1. **¿Cómo las tecnologías semánticas pueden ayudar a desarrollar un buscador guiado por la/s ontología/s del dominio que se usan como modelo de integración de datos?** En la sección 4.1.1.1 se ha expuesto ODS, un generador de consultas SPARQL guiado por las ontologías del dominio que están ubicadas en el almacén semántico. Gracias a ésto, las mismas ontologías que se han usado para integrar, anotar y clasificar la información de los sistemas operacionales, sirven para consultar y extraer los datos. Este servicio, además de generar consultas SPARQL, permite parametrizar algunas de sus variables para la generación de buscadores avanzados por cada una de esas variables. La interfaz gráfica de ODS permite que cualquier usuario pueda usarlo para formular sus propias consultas sin necesidad de tener conocimientos avanzados en informática en general, ni en Web Semántica en particular.

- **Las representaciones semánticas permiten generar modelos más fácilmente explotables por los usuarios de una solución de Inteligencia de Negocio.** Comprobar esta hipótesis requiere contestar a la siguiente pregunta:

1. **¿Qué ofrece la Web Semántica para generar modelos de datos equivalentes a OLAP?** En la sección 8.2.5 se definen las principales características de OLAP, y cuáles han sido las aproximaciones para representar los datos de este tipo de almacén semánticamente. En la sección 4.4 se definen los perfiles semánticos como un extracto resumido de un recurso. Ese extracto únicamente contiene la información que realmente se desea usar en un análisis concreto. Con este modelo se ha conseguido usar la flexibilidad de OWL a la hora de representar conocimiento, para generar modelos semánticos más cómodos en cuanto a representación y eficiencia, y para realizar análisis concretos sobre conjuntos de recursos con propiedades parecidas. Es decir, usando un enfoque totalmente diferente a OLAP se ha conseguido tener una herramienta que ofrece servicios de análisis equivalente. La principal ventaja de este enfoque es que no es necesario desnormalizar los datos para obtener un analisis eficiente de la información, con el coste que ello conlleva en espacio y rendimiento, sino que simplemente cambiando cómo se representan los datos se puede tener ese mismo análisis. Otra de las ventajas es que mientras que en OLAP es necesario la definición de modelos de datos complejos que deben ser definidos por expertos en TIC, en esta propuesta, cualquier usuario puede generar esos perfiles (guiándose en la propia ontología con ODS) de forma autónoma y personalizada. Por último, también se destaca que se ha evaluado el rendimiento de este modelo en diferentes casos de uso como SocialBROKER (ver sección 5.1), SECARE (ver sección 5.3) y SECOLON (ver sección 5.4), con buenos resultados al usar almacenes de datos NoSQL.

- **Los buscadores semánticos y las redes sociales de evaluación de activos empresariales evalúan el conocimiento generado por una solución de Inteligencia de Negocio.** Comprobar esta hipótesis requiere contestar a las siguientes preguntas:

  1. **¿Qué función puede tener la integración de herramientas de evaluación en una solución de Inteligencia de Negocio?** En la sección 2.1 se describe la colaboración entre los diferentes usuarios en una plataforma de IN como una de sus características esenciales. En la sección 2.6 se han definido diferentes modelos que permiten evaluar el conocimiento de las organizaciones y que éstas no suelen estar integradas en soluciones de IN comerciales. El impacto que ha tenido la IN en las organizaciones se suele

medir en los resultados de la empresa [26], pero gracias a integrar este tipo de herramientas se podrán contestar a preguntas como ¿cuáles han sido los activos intangibles que han participado en esos resultados? ¿cuál es la evaluación del desempeño de los recursos humanos directamente relacionados con esos resultados? Además, si los propios datos de la evaluación se incorporan como una fuente más a la solución de IN, se puede analizar la evolución histórica de la organización o conocer el impacto que ha tenido un determinado plan estratégico.

2. **¿Cómo se puede usar la Web Semántica para implantar metodologías de evaluación del conocimiento de forma genérica?** La flexibilidad que ofrece la representación ontológica a la hora de modelar un dominio también es útil para evaluar los recursos de ese dominio y analizar los resultados de esa evaluación. En la sección 4.3 se han definido diferentes propuestas para realizar esa evaluación usando tecnologías semánticas. Todas ellas se basan en que los recursos se anotan en base a unos criterios de evaluación, cuya representación semántica dará información sobre el perfil tipo que debe tener un determinado recurso y cuál es su evaluación real en un momento determinado. El modelo de evaluación basado en indicadores se ha usado en SocialBROKER (véase sección 5.1) para medir la bondad de las carteras financieras de los inversores, a partir de los datos recogidos de diferentes índices bursátiles. El modelo de evaluación 360º se ha empleado para evaluar el desempeño de recursos humanos en un hospital (véase sección 5.2). En ese mismo caso de uso también se ha usado el modelo de evaluación basado en cuestionarios para la detección de necesidades formativas por parte de los empleados del hospital. La principal ventaja de usar Web Semántica para evaluar los activos conocimiento, además de la representación de la información, consiste en poder enlazar esa evaluación con información externa, proporcionando retroalimentación sobre los aspectos a mejorar. También facilita que los resultados de la evaluación se puedan comparar entre organizaciones diferentes.

- **Los servicios semánticos permiten al usuario generar informes, hacer búsquedas semánticas avanzadas, generación de cuadros de mando semánticos personalizados, análisis del impacto de la toma de decisiones, sistemas de planificación y recomendadores.** Comprobar esta hipótesis requiere contestar a las siguientes preguntas:

1. **¿Cómo se puede usar el buscador semántico para generar servicios más avanzados?** En las secciones 4.5.1, 4.5.2, 4.5.3 se describen tres herramientas que usan ODS para generar servicios más avanzados. Con ODS se pueden generar buscadores parametrizados con las condiciones por las que interese filtrar. Esa característica se ha usado en casos de uso como SECARE (ver sección 5.3) y SECOLON (ver sección 5.4), facilitando que el personal que gestiona bases de datos epidemiológicas pueda configurar sus propios buscadores.
   Con ODS también se pueden generar alertas sobre la información. Esta característica ha sido validada en diferentes casos de uso como SocialBROKER, en el que se usa para enviar alertas a otros usuarios sobre inversiones o contenidos que podían ser de su interés. En el caso de uso del CRD del proyecto NELA (ver sección 5.5) se usa ODS para configurar alertas que informen sobre el estado de reclutamiento del proyecto.
   Por último, ODS facilita la generación de cuadros de mando semánticos. Estos cuadros se generan marcando alguno de los filtros como que se quieren agrupar. Esta herramienta permite la representación gráfica de los datos, y en formato de tabla. Desde esas representaciones se puede interactuar para navegar a los datos que generan un determinado indicador. Esta característica ha sido utilizada en todos los casos de uso del capítulo de validación (ver sección 5) ya que es una herramienta transversal a cualquier solución de IN Semántica.

2. **¿Cómo se pueden integrar algoritmos de recomendación y predictivos en entornos semánticos?** En la sección 4.5.4 se han descrito las dos alternativas para realizar recomendaciones o predicciones sobre ciertas variables del dominio. En la sección 4.5.4.1 se define una función de similitud que permite recomendar recursos a partir de su similitud semántica y de su grado en la evaluación. Este recomendador ha sido validado en el caso de uso del proyecto EUCOMM Tools (ver sección 5.6) con grandes tasas de acierto a la hora de recomendar que dos recursos son parecidos y poder reutilizar sus anotaciones. Este modelo también ha sido validado en el caso de uso SocialBROKER (ver sección 5.1) a la hora de hacer recomendaciones de iniciativas financieras o contenidos. En la sección 4.5.4.2 se propone el uso de algoritmos de redes bayesianas para predecir una determinada variable. Se propone ese algoritmo porque su representación se asemeja a

una ontología a excepción de que en una red bayesiana no se pueden tener ciclos. Para superar esa limitación, cada que vez que se encuentra un ciclo se genera una nueva red teniendo en cuenta que los pasos hasta llegar a ese punto son una nueva distribución de probabilidad. Esto facilita la generación de un árbol de redes bayesianas que se irá ramificando conforme se vayan encontrando variables que puedan tener algún ciclo. Este modelo se ha implementado y evaluado en los casos de uso SECARE ver sección (5.3) y SECOLON (ver sección 5.4).

3. **¿Cómo se pueden integrar los activos de conocimiento con la evaluación de los mismos para analizar el impacto de una determinada decisión estratégica?** Para responder a esta pregunta hay que ver dos de las propuestas del modelo de explotación de la solución de IN semántica: sistema de planificación (ver sección 4.5.5) y análisis del impacto (ver sección 4.5.6). El sistema de planificación se apoya en la evaluación del conocimiento para detectar fortalezas y debilidades. Además explota la anotación semántica de recursos que son clasificados en oportunidades comerciales, oportunidades de mejora y amenazas. Estos recursos también se anotan en base a los criterios de evaluación pudiendo enlazar fácilmente cuáles de ellos pueden ser de utilidad para que la organización mejore. Es importante que los recursos se anoten también en dos variables de coste: temporal y económico, lo que permitirá restringir la generación de un plan en el tiempo y con una financiación limitada. Estos dos parámetros serán decididos por el usuario en el momento de generar el plan estratégico. En el caso de uso de planificación de la formación (ver sección 5.2) se ha validado la generación de planes personales de formación continuada para personal de un hospital.
   El modelo de análisis del impacto se apoyará en los resultados de la evaluación anterior, de los nuevos indicadores resultado del plan estratégico usado y de una nueva evaluación, permitiendo conocer el impacto identificando si el plan estratégico ha conseguido mejorar los indicadores y los activos de conocimiento de la organización.

7.2 Contribuciones

En esta trabajo se presenta una solución global para inteligencia de negocio usando tecnologías semánticas. Las soluciones expuestas se han aplicado con

éxito en varios escenarios de validación:

- **Redes sociales semánticas.** En este caso de uso se ha expuesto SocialBROKER, un entorno para el análisis y la explotación de contenidos y carteras financieras. Los contenidos son anotados a través de técnicas de anotación automática y de etiquetado semántico. Las carteras financieras han sido modeladas a través de los cuestionarios semánticos. Se han desarrollado diversas soluciones de recomendación que tienen en cuenta diferentes métricas de evaluación de cada contenido.

- **Plataforma para la planificación semántica.** Esta solución se usa a día de hoy en un hospital de la Región de Murcia. La plataforma usa el modelo de evaluación para establecer cuál es el desempeño de cada empleado. Además, permite la anotación manual de programas de formación continuada. Cruzando los datos de ambas fuentes es capaz de recomendar cuál debería de ser la formación para cada persona de la organización. Gracias al modelado semántico, los usuarios pueden configurar análisis personalizados de evaluación que comparen los recursos con su perfil tipo, o entre ellos. Esta plataforma está activa desde 2013.

- **Plataforma para el análisis epidemiológico.** Para esta solución se ha tenido la oportunidad de evaluar dos casos de uso. SECARE permite realizar análisis personalizados en un registro local de cáncer, y SECOLON permite, además, predecir los niveles de riesgo de los pacientes dentro de un programa de cribado de cáncer colorrectal. En ambos casos se han usado diferentes aproximaciones de redes bayesianas para generar modelos que ayuden a planificar la carga futura de los diferentes servicios implicados.

- **Cuaderno de recogida de datos.** Se ha usado la solución de CRD semántico para realizar el seguimiento de un proyecto concreto como es NELA. El proyecto ha empezado en 2015 y aún se está evaluando, aunque ya gestiona los datos actualizados de todo el reclutamiento de los pacientes.

- **IN semántica en contenidos multimedia.** El anotador manual de contenidos ha permitido la puesta en marcha de una plataforma para la anotación y el análisis de la expresión génica de ciertas sustancias sobre ratones modificados genéticamente. Esta plataforma se usa en el marco de un proyecto europeo denominado EUCOMM Tools. Las anotaciones se realizan sobre imágenes recuperadas de un microscopio confocal. Esta plataforma está activa desde principios de 2014 y es

consultada por todos los miembros del proyecto. Gracias a los cuadros de mando y a los perfiles semánticos, los investigadores del proyecto han podido analizar y evaluar sus hipótesis.

Las principales aportaciones que se pueden extraer de esta tesis son:

- Diseño de un modelo de evaluación de activos de conocimiento basado en tecnologías semánticas. La definición de criterios de evaluación y el modelo de anotación semántica permite la generación de perfiles tipo que cada recurso debería cumplir. Además, permiten evaluar los recursos a partir de tres tipos diferentes de metodologías:
  - Evaluación 360 grados. Usada para la evaluación de activos de conocimiento. Basada en que personas diferentes que están relacionadas con el recurso lo evalúen en una convocatoria concreta.
  - Evaluación basada en indicadores y rangos. Usada para calcular indicadores a partir de los datos objetivos de la empresa. Los indicadores se clasifican en rangos para conocer su impacto. Los indicadores se relacionan con los criterios de evaluación que son necesarios para conseguir datos positivos.
  - Evaluación basada en cuestionarios. Usada para evaluar conocimientos a través de exámenes, hacer test de ánimo laboral, test de comportamento social, etc. Esta herramienta puede ayudar a evaluar tanto el conocimiento tangible como el intangible.

  Gracias a los modelos de evaluación del conocimiento, la plataforma permite generar planes estratégicos a partir de las fortalezas y debilidades encontradas. Además permite recomendar cuáles son las oportunidades comerciales, de mejora o las amenazas que pueden tener un impacto en los resultados de la organización. Si los planes o parte de ellos se ponen en marcha, en una posterior evaluación se puede analizar el impacto que ha tenido el plan recomendado en los indicadores de la organización, conviertiéndose en un nuevo indicador que pueda ser útil a la hora de priorizar un recurso u otro.

- Diseño de un modelo de generación de perfiles semánticos de recursos. El enfoque propuesto, equivalente a OLAP en los sistemas de IN tradicionales, consiste en la definición de una representación semántica reducida de un recurso o perfil semántico. Gracias a estos perfiles los usuarios pueden configurar análisis de una forma más sencilla y eficiente, agilizando el proceso de explotación de los datos.

- Diseño e implementación de un generador gráfico de consultas SPARQL. ODS permite generar consultas SPARQL desde un entorno gráfico e intuitivo al usuario, guiándose en los conceptos y propiedades ontológicas. Además de generar consultas, con ODS también se puede filtrar cuál es la información que se quiere recuperar, pudiendo usar funciones de agregación (suma, media, máximo, mínimo, etc.) en cualquier consulta. También permite ordenar los resultados de la respuesta. El modelo de generación de consultas en ODS puede serializarse para posteriormente reutilizarse para volver a hacer una búsqueda o calcular un indicador. Esta serialización también permite que el usuario pueda generar formularios de filtro avanzado, marcando los campos que se requieran como parametrizables.

- Diseño de un modelo para la generación de cuestionarios semánticos. En muchos ámbitos de negocio, una misma línea de negocio puede tener una gran heterogeneidad en su representación, como por ejemplo los cuadernos de recogida de datos. El modelo de generación de cuestionarios semánticos permite la definición de conceptos, relaciones y propiedades que pueden completar la información de un determinado recurso. Además, permite definir las fases o procesos en los que se rellenará una información u otra. La información recogida en estos cuestionarios formará a pasar parte del almacén de datos semántico directamente, sin necesidad de ningún mecanismo de transformación.

- Diseño de un modelo para la recomendación basado en funciones de similitud. Este modelo se basa en una extensión de la función de similitud basada en nodos propuesta por [88]. La extensión prioriza aquellos recursos mejor valorados a través del modelo de evaluación del conocimiento de los recursos. Este modelo ha sido aplicado para anotar imágenes microscópicas de ratones transgénicos, obteniendo que más de un 60 % de las anotaciones se han generado a partir de este recomendador.

- Diseño de un modelo para la recomendación basado en redes bayesianas. El modelo de recomendación se basa en los perfiles semánticos. Como se ha comentado los perfiles semánticos son una representación ontológica reducida de un recurso. Una ontología se puede representar como un grafo dirigido cíclico. Una red bayesiana es un algoritmo que permite predecir variables usando el teorema de Bayes. Las redes bayesianas se representan como un grafo dirigico acíclico. Este modelo de recomendación se basa en usar grupos de perfiles semánticos de recursos que tengan criterios parecidos, generando las diferentes probabilidades condicionadas y usando Bayes para predecir el valor de la variable que

desconocemos. En el caso de que los perfiles semánticos tengan ciclos se genera un árbol de redes que se irá ramificando cada vez que el usuario añada una variable que pueda generar un ciclo. Este modelo ha permitido predecir con un gran porcentaje de acierto los niveles de riesgo de tener cáncer de pacientes en un programa de cribado colorrectal.

- Cuadros de mando semánticos. Sobre ODS se ha desarrollado una herramienta que permite representar los datos agregados de forma gráfica y con una tabla de los datos. Gracias al modelo de persistencia de consultas, los resultados gráficos son navegables, es decir, haciendo clic sobre el gráfico o sobre los datos se puede ver el listado de los recursos que lo generan. Esta herramienta permite generar gráficos en formato pastel, de barras y rádar. En los dos últimos se pueden comparar los resultados de dos o más consultas SPARQL, que permitan tener una vista comparativa entre los recursos.

- Puesta en producción de tres soluciones semánticas que están operativas actualmente. Una de las principales aportaciones del trabajo es la posibilidad de evaluar en un entorno real, con datos reales, cómo se comporta la propuesta de IN semántica en diferentes entornos. Los tres sistemas están operativos y siguen usándose continuamente para: (1) evaluar el desempeño de los trabajadores de un hospital y generar planes de formación personalizados, (2) clasificación y explotación de una cohorte de ratones transgénicos a partir de la anotación de imágenes de expresión génica obtenidas de microscopio digital, y (3) un cuaderno de recogida de datos para el reclutamiento y explotación de un proyecto de investigación biomédica donde se reclutan mujeres embarazadas y a sus bebés.

7.3 Conclusiones generales

La IN es una herramienta que ha tenido un gran impacto en las organizaciones. Existen múltiples soluciones en el mercado de IN, pero la ausencia de modelos semánticos formales dificulta los análisis personalizados, el enlace con otros recursos externos y la comparativa entre diferentes entidades. Las soluciones actuales no tienen mecanismos para explotar contenidos no estructurados provenientes de soluciones tan comunes hoy en día como las redes sociales y las plataformas Web 2.0. Otra carencia común es la ausencia de herramientas que permitan evaluar los activos de conocimiento, y que esa evaluación retroalimente la plataforma para la futura toma de decisiones.

En este trabajo se han definido una serie de metodologías e implementado un conjunto de herramientas que permiten integrar, entregar información y analizar los datos usando tecnologías semánticas. Esas tecnologías permiten que los usuarios, sin ayuda de personal experto en TIC, puedan configurar buscadores avanzados, generar cuadros de mando personalizados, realizar planes estratégicos y evaluar su impacto, y establecer análisis complejos a partir de perfiles semánticos. Además, gracias al uso de tecnologías semánticas se han sentado las bases para que se pueda enlazar con fuentes de datos externas y realizar análisis comparativos con otras organizaciones.

Las soluciones desarrolladas se han evaluado en diferentes dominios: económico y financiero, epidemiológico, evaluación del desempeño, investigación clínica e investigación biosanitaria, con resultados satisfactorios y con tres de ellos desplegados en entornos reales, con datos reales. Este trabajo es un ejemplo de cómo las ontologías pueden guiar todo el ciclo de vida de una herramienta de IN, desde la integración de los datos, hasta la explotación y generación de conocimiento.

7.4 Vías futuras

A continuación se van a detallar las vías futuras de esta tesis. Dichas vías se dividen en los tres niveles de la IN: integración de datos, entrega de información y análisis.

- **Integración de datos.** Una de las limitaciones principales del modelo de integración propuesto está en la necesidad de transformar los datos de las fuentes en información semántica. Aunque se pueden usar algunas de las herramientas de IN definidas, en el caso de no extraer los datos, no se podrían usar los modelos de evaluación de activos de conocimiento ni los cuestionarios semánticos. Para paliar esta limitación, como trabajo futuro se propone el desarrollo de un modelo de interoperabilidad entre los modelos semánticos puros y los modelos que realizan un envoltorio semántico sobre los modelos relacionales tradicionales (*D2RQ*, *Linked Data Views*, etc.). Esto permitirá que se puedan usar el cien por cien de las herramientas de IN Semántica propuestas en esta tesis sin necesidad de hacer una extración, transformación y carga de los datos de las fuentes reales de la organización. La principal ventaja de este modelo mixto consistirá en poder hacer un análisis en tiempo real de la actividad de la entidad.

- **Entrega de información.** La herramienta clave de este nivel es ODS

(ver apartado 4.1.1.1). Como línea futura se plantean los siguientes trabajos:

  - Ampliar el modelo de generación de consultas de ODS para que permita agregar los resultados de las consultas por más de una variable. Al realizar esa ampliación, la generación de cuadros de mando se verá enriquecida al poder definir representaciones gráficas que comparen diferentes líneas de negocio en una sola consulta.
  - Extender ODS con la posibilidad de añadir subconsultas dentro de la consulta principal para ampliar las capacidades de generación de filtros a los usuarios.
  - Usar el modelo de serialización de consultas SPARQL para implementar un registro de consultas que pueda ser analizado y explotado usando las mismas herramientas definidas para la IN Semántica. Gracias a esta herramienta, los administradores de la plataforma podrán saber cuáles son los conceptos más empleados para analizar el rendimiento de la organización.

- **Análisis.** Las principales líneas futuras en este nivel son: (1) la inclusión de más algoritmos de minería de datos que puedan ser seleccionados y parametrizados por los usuarios, (2) desarrollar más casos de validación para medir el impacto de las decisiones estratégicas de las organizaciones y (3) mejorar la herramienta de visualización interactiva para que, además de navegar a los buscadores de las fuentes de datos que generan una determinada representación, también se pueda navegar entre diferentes cuadros de mando.

Queda también pendiente como trabajo futuro la mejora de las soluciones de IN semánticas propuestas en esta tesis y descritas en la sección 4.6. A continuación se indican las principales líneas futuras asociadas a cada una de estas soluciones.

- **Red Social Semántica.** Para esta solución, el trabajo futuro se basaría en hacer una mayor validación de la plataforma, añadiendo a SocialBROKER (ver sección 5.1) otras fuentes de conocimiento como noticias o información de otras redes sociales. Con la adición de estas fuentes se conseguirá una plataforma colaborativa e interoperable con otros sistemas.

- **Plataforma para la Planificación.** La validación de esta solución se ha realizado en un entorno real, concretamente en un hospital. Como trabajo futuro se plantean las siguientes líneas:

  - Evaluar más recursos, además del personal del hospital. Por ejemplo los equipos sanitarios, las guías clínicas o los modelos organizativos, que permitan obtener una valoración global de los activos del hospital.
  - Generar cuadros de indicadores sobre los datos reales de funcionamiento del hospital.
  - Generar planes estratégicos que, además de incluir itinerarios formativos, también recomienden la mejora de los procesos o la renovación de ciertos equipos.
  - Hacer un análisis del impacto real que un plan estratégico concreto ha tenido, tanto en la evaluación del conocimiento, como en los resultados reales.

- **Plataforma para el Análisis Epidemiológico.** En el ámbito de esta solución se plantea como trabajo futuro:

  - Integración con los sistemas de información de Historia Clínica Electrónica (HCE) permitiendo que ambos sistemas (HCE y registro epidemiológico) puedan transferirse información automáticamente.
  - Incorporación de más algoritmos predictivos para que el clínico pueda elegir el que mejor resultado dé dependiendo del análisis que se pretenda realizar.

- **Cuaderno de Recogida de Datos (CRD).** El CRD Semántico sólo ha podido validarse en un estudio clínico (ver sección 5.5). Como trabajo futuro se plantea que se pueda ampliar su uso en más estudios de investigación clínica. Además, también se considera necesario la mejora de las capacidades de integración del CRD. Esta mejora consiste en poder definir relaciones entre los campos de los cuestionarios del CRD y otras fuentes de información externa para que puedan ser rellenados automáticamente. Dentro de este ámbito se plantea la integración con plataformas de apoyo a la investigación biosanitaria como biobancos, animalarios y servicios de genómica o proteómica.

- **IN Semántica en Contenidos Multimedia.** Esta solución ha sido validada con el proyecto EUCOMM Tools (ver sección 5.6). Como trabajo futuro se ha planteado el uso de modelos ontológicos más formales en los que se puedan reutilizar otras ontologías como la *Gene Ontology* [211]. Esta estandarización permitirá que los datos puedan enlazarse

fácilmente con otros repositorios semánticos usados en el ámbito de la investigación en ciencias de la salud y la vida.

7.5 Publicaciones y contribuciones en congresos

7.5.1 Publicaciones JCR

- Jesualdo Tomás Fernández-Breis, José Alberto Maldonado, Mar Marcos, María del Carmen Legaz-García, David Moner, Joaquín Torres-Sospedra, Ángel Esteban-Gil, Begoña Martínez-Salvador, Montserrat Robles. Leveraging Electronic Healthcare Record Standards and Semantic Web Technologies for the Identification of Patient Cohorts. Journal of the American Medical Informatics Association (JAMIA) (factor de impacto 2013: 3,932).

- Ángel Esteban-Gil, Francisco García-Sánchez, Rafael Valencia-García, Jesualdo Tomás Fernández-Breis. SocialBROKER: A collaborative social space for gathering semantically-enhanced financial information, Expert Systems with Applications 39(12) 9715-22 (factor de impacto 2012: 1,854).

7.5.2 Congresos

- Ángel Esteban-Gil, Jesualdo Tomás Fernández-Breis and Clara Miranda-López. Performance evaluation of human resources in a hospital, SWAT4LS 2015, Cambridge, United Kingdom, poster.

- Ángel Esteban-Gil and Jesualdo Tomás Fernández-Breis, Semantically-enhanced Electronic Case Report Forms, International Semantic Web Conference 2015, Industry Track.

- Ángel Esteban-Gil, Jesualdo Tomás Fernández-Breis and Martin Boeker. Analysis and visualization of disease courses in a semantic enabled cancer registry, SWAT4LS 2014, Berlin, Germany.

- Jesualdo Tomás Fernández-Breis, Francisco Frutos-Morales, Ángel Esteban Gil, Dagoberto Castellanos-Nieves, Rafael Valencia-García, Francisco García-Sánchez, and María del Mar Sánchez-Vera. Recommendation of personalized learning contents supported by semantic web tech-

nologies, Communications in Computer and Information Science series 278, pp. 540-545. Springer, Heidelberg.

- Ángel Esteban-Gil, Francisco García-Sánchez, Rafael Valencia-García, Jesualdo Tomás Fernández-Breis. A Social-Empowered Platform for Gathering Semantic Information Communications in Computer and Information Science series 278, pp. 534-539. Springer, Heidelberg.

- Francisco Frutos-Morales, Mª del Mar Sánchez-Vera, Dagoberto Castellanos-Nieves, Ángel Esteban-Gil, Carlos Cruz-Corona, Mª Paz Prendes-Espinosa, Jesualdo Tomás Fernández- Breis. An extension of the Oe-LE platform for generating semantic feedback for students and teachers.Procedia - Social and Behavioral Sciences, Volume 2, Issue 2, 2010, Pages 527-531.

- Ángel Esteban-Gil, Jesualdo Tomás Fernández-Breis, Dagoberto Castellanos-Nieves, Rafael Valencia-García, Francisco García-Sánchez. Semantic enrichment of SCORM metadata for efficient management of educative contents (2009) Procedia - Social and Behavioral Sciences, Volume 1, Issue 1, Pages 927-932.

Bloque IV

English

Chapter 8

Summary

8.1 Introduction

Business Intelligence (BI) is defined as a set of methodologies and tools that allow to transform the data in information and the information in knowledge [1]. From the Information Technology (IT) point of view, BI consists in debugging and integrating data from several information systems and transforming them in one structured data source for exploiting through different tools like advanced searchers, reporting services, dashboards and predictor models. The BI technological architecture has three levels: integration, information delivery and analysis. On the technical side, the Semantic Web proposes to give a meaning to multiple contents [10]. These contents can be processed by computers, which may generate new knowledge automatically and ensure data consistency. The Semantic Web has been identified as a new technology that can be useful in BI solutions, in the integration layer [11; 12; 13], to evaluate the information quality [14], in the exploitation layer [15; 16; 17], and of course, in the sharing layer for comparing business results between companies [18; 19].

Nowadays, the approaches that integrate semantic web with BI propose the use of the semantic technologies as a complementary technology to improve some skills but, to the best of my knowledge, there is no complete solution based in semantic technologies for BI. This thesis proposes a functional solution for BI platforms based on semantic tools and technologies. The lifecycle of the platform is supported by ontologies, which enable the integration, transformation and exploitation of the data using the same vocabulary. The final result is a software solution that enables the generation and sharing of knowledge thanks to the advantages of semantic technologies. Furthermore, semantic models for (1) adding structured information

to the semantic repository, and (2) for evaluating the knowledge assets and analysing the impact of the strategic actions in the company are results of this work.

8.2 State of art

8.2.1 Business Intelligence

The main objective of BI is to know what happens in the organisation and why it happens, what could happen in the future (based on historical data), and which action should be taken. [2] defines thirteen fundamental features of BI solutions. These features are classified in three levels: integration, information delivery and analysis. Next, we describe each of them:

- **Integration**
  - **BI infrastructure.** This service integrates all the tools of the BI solution, such as users, security, corporate image, searchers, etc. It is the entry point to the stored information.
  - **Metadata management.** Metadata management is a fundamental feature because it enables the use of a common language for transforming and exploiting the data.
  - **Development tools.** BI solutions have to be able to incorporate new applications and services. BI solutions often offer component wizards for building new applications without coding.
  - **Collaboration.** BI solutions must offer discussion forums about the data, annotations, indicators definition among others for the users.
- **Information delivery**
  - **Reporting.** This is another essential feature of BI solutions, enabling (1) performing searches over the data with several criteria, (2) representing the results of the search in a illustrative way for the user: and (3) generating periodical reports. The alerts are a special report type, including information about which activities are wrong or exceed the time required.
  - **Dashboards.** They offer a graphical representation of the data.
  - **Ad hoc queries.** They enable users to perform queries over the data warehouse without requiring IT expertise.

  - **Microsoft Office integration.** Microsoft Excel is often used as a middleware for exploiting, and even editing the content of the data warehouse.
  - **Search-based BI.** This service allows performing advanced searches over the data warehouse.

- **Analysis**
  - **OLAP.** OLAP provides fast and efficient data analysis. The storage system for OLAP is optimized for querying, not for other maintenance operations. There are several technological OLAP architectures like relational databases, multidimensional databases, storage in-memory and NoSQL databases.
  - **Interactive visualisation.** The user is able to navigate any representation provided for the data. For example, if the user has defined a dashboard, she has to be able to navigate the list of data that represents a piece of the chart.
  - **Predictive modelling and data mining.** These services use advanced maths and artificial intelligence techniques for generating predictive models or new knowledge from the data warehouse.
  - **Scorecards.** This service links dashboard with real company indicators.

Nowadays, BI solutions have their cores centred in delivering information, although the analysis and improvement of the integration features have become a hot topic for BI solutions in large companies [3]. BI is recognised as a tool with a real positive impact in the companies profits [7]. There is a general agreement about the use of BI for improving the organisation strategies [22]. In the last years, the fast progress of BI has generated a high number of technological platforms, with high heterogeneity levels.

The BI evolution can be described as follows:

- **BI 1.0.** BI 1.0 is centred in integrating, transforming and exploiting structured data of the information systems of the company [24; 25; 7].
- **BI 2.0.** BI 2.0 incorporates the information generated by new 2.0 applications like social networks, wikis, forums, etc. In this case, it needs natural language processing and text mining for classifying and exploiting the contents in open text [32; 4].
- **BI 3.0.** BI 3.0 is related to the Internet of Things [26]. Nowadays there are no solutions at this level, but they are included in [35] as technologies that could change the BI market.

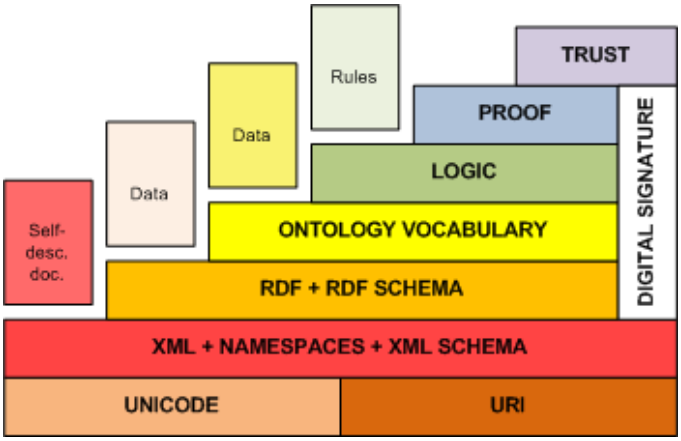


Figure 8.1: The Semantic Web Stack

8.2.2 Semantic Web

The Semantic Web can be seen as the next-generation web in which information is given well-defined meaning, better enabling computers and people to work in cooperation [10]. Ontologies [53] constitute the standard knowledge representation mechanism for the Semantic Web. Technologies like OWL [63] for the ontology construction enable a formal representation of the domain. Tim Berners-Lee proposed a layered architecture for implementing the Semantic Web (see figure 8.1)

The URIs (Uniform Resource Identifier)[47] allow uniquely to identify a resource, and they are expressed using Unicode. XML [48] technologies, which include the NameSpaces and the XML Schema [49], are very used in the Web. This layer is included in the Semantic Web architecture for serializing the semantic information. In this way the semantic information can be transferred using existing technologies. RDF (Resource Description Framework) [50] allows to describe the resources of the Semantic Web. A resource is any thing that can have other elements linked. RDF stores the information as triples that are composed by subject, predicate and object. The subject is the source of the relationship, the predicate expresses the relation, and the object is the destination of the relation. The information stored using RDF can be defined as a directed labelled graph, where the subject and object are the nodes and the predicate is the edge. RDF Schema (RDFS) [51] is an extension of the RDF language that allows to define the resources as classes, organise the classes in hierarchies, define the relationships between classes as properties as well as their domains and ranges.

However, these languages are insufficient to precisely define the knowl-

edge in many domains. Ontologies [53] constitute the standard knowledge representation mechanism for the Semantic Web. Technologies like OWL [63] for the implementation of ontologies enable a formal representation of the domain. Ontologies have been used for representing knowledge in several domains as biomedicine [54; 55], organisational memories [56; 57], knowledge management [58], bioinformatics [59], finances [60] or e-Learning [61]. Ontologies allow to share and reuse the knowledge, easing the implementation process of expert systems.

The logic layer allows to define rules for reasoning. There, reasoners are able to infer new knowledge over the original semantic model. The last layers have not yet been developed. In general, these layers intend to establish safety standards to evaluate the resources and the reliability with the help of the digital signature.

8.2.2.1 Methods for semantic transformation

The methods for transformation and semantic representation of information follow similar approaches. The extraction process and the information transformation are based on the definition of mappings between the entry data source and the output semantic model.

There are several tools for semantic transformation:

- **D2RQ.** It allows to query data stored in relational databases using SPARQL thanks to the generation of virtual RDF graphs [104]. This tool is totally automatic.
- **Triplify.** It allows to publish as Linked Data [93] the content of relational databases [105]. The transformation process is partially automatic.
- **Linked Data Views (Virtuoso).** OpenLink Virtuoso [106] is a database management system that works with several persistence models (relational, XML, object-relational, virtual and RDF). Linked Data Views [107] allow to query with SPARQL any persistence model stored in Virtuoso. It is an automatic process whenever all the information is stored in Virtuoso.
- **XS2OWL.** This tool allows to transform XML schemas in OWL [108]. Thanks to this, we can use SPARQL for querying XML databases. The transformation process is totally automatic.
- **RDB2OWL.** It is an approach for transforming the data stored in relational databases in RDF or OWL[109]. The user has to define

the mappings between the entries and the outputs. Given that the mappings are manual, transforming large ontologies can be tedious.

- **Karma.** It allows to link the source model to ontologies for generating a semantic representation of the data source [110]. This process is partially automatic whenever we have annotated the source model.
- **Populous**. It is an assistant for building ontologies [111]. With Populous we can use patterns to guide the ontology design. Populous is able to import data from CSV automatically.
- **OGO System.** This approach proposes a tool for integrating different relational databases in one semantic repository based in one ontology [112].
- **SWIT.** It is a semantic transformation engine capable of generating RDF and OWL repositories from both relational and XML-based databases [113]. Besides transforming the data, SWIT prevents the generation of logically inconsistent data with the support of DL reasoners by not transforming inconsistent source content. The transformation method has three main steps: (1) definition of the mapping rules between the fields of the database and the ontology; (2) generation of the OWL data; and (3) importing the OWL data into the semantic data store.

8.2.2.2 BI and Semantic Web

BI and the Semantic Web have been integrated in some recent works [114]. [12] proposes the use of ontologies for extraction, transformation and loading processes, [11] proposes a formal modelling of the business domain, and [13] validates BI data using reasoners. The MUSING (MUlti-industry, Semantic-based next generation business INtelliGence) European project integrates Semantic Web and natural language processing technologies for building BI solutions. Their main results are in the scope of information integration [11; 116; 117] and standardisation of business information [116].

8.2.3 Web 2.0

Web 2.0 and Social Networks are of paramount importance for the research on collaborative systems. The term "Web 2.0" is commonly associated with a cluster of technologies and design patterns that assist in developing Web applications that facilitate interactive information sharing, interoperability

and collaboration on the World Wide Web [32]. Social Networks can be regarded as one of the multiple services provided using Web 2.0 technologies. However, it has become clear that the social networking construct is critical to the success of Web 2.0 applications [84]. A social network is a collection of individuals (usually human beings) or organisations linked together by a set of relations [119]. Entities in social networks are known as "nodes" and the relations or connections between them as "ties". The value of social applications benefits from the network effect, which states that the value of a service to a user arises from the number of people using the service [84]. Social Networks and Web 2.0 tools have been successfully exploited in various contexts [120]. Web 2.0 platforms are essential for BI according to [2], because they allow to generate discussion forums, sharing and annotating information, and knowledge evaluation.

Before the rise of the Semantic Web, social networks were represented through a matrix and the properties of these networks therefore studied as a subset of graph theory [121]. Currently, there are other approaches for representing social networks, such as the Friend of a Friend (FOAF) specification [64] or the Semantically-Interlinked Online Communities (SIOCs) initiative [65]. The combination of Semantic Web technologies with Web 2.0 application design patterns has given rise to the social-semantic Web, also referred to as Web 3.0 [122].

8.2.4 Information integration

The heterogeneity of the data in the information systems of the organisation is a common problem for BI solutions. To solve it, the creation of integration methodologies is required [129; 130]. The integration process represents more than 80% of the work needed to implement a BI solution [6]. There are two types of integration processes:

- **Horizontal integration.** The data integrated are complementary although they are found in different data models.
- **Vertical integration.** In this case, the approach integrates data semantically equivalent, but stored in different systems with different data models.

Data integration is defined as the process of combining data from different sources. This process generates a unified global model for exploiting services [131].

The most used integration approaches can be classified by the source data: unstructured data, structured data o linked resources. In the first case,

the integration approaches propose codification systems based on keywords or vocabularies. If source data are structured, then there are two possible options: (1) solutions based on data warehouse or centralized data [129], and (2) systems based on views that allow to obtain a on-demand views of data [133]. The last approach considers data as linked resources, easing the navigation between the data [135]. This integration approach is not useful for exploiting the information.

The semantic technologies are considered an ideal environment for data transformation and integration [142]. The ontologies are a basic element to define formal models that delete the data heterogeneity.

8.2.5 On-Line Analytical Processing (OLAP)

OLAP is one of the most common tools in the BI solutions [5]. The main function of OLAP consists in expediting the queries of large amounts of data [148]. In the last years, several approaches have been developed for integrating Semantic Web with OLAP. [13] proposes the mapping between the data sources and OWL or RDF ontologies, for using SPARQL for filling the data warehouse. [15] proposes the generation of OLAP structures from semantic models automatically. [17] proposes the use of OLAP for structured data and semantic repositories for unstructured data like documents, images, etc.

One of the most important works for combining OLAP and the Semantic Web is QB (The RDF Data Cube Vocabulary) proposed by W3C [18]. QB can be used as a common language for defining OLAP cubes for sharing the information in a common semantic model, but we cannot use it for exploitation. [19] proposes a new vocabulary QB4OLAP that allows to perform semantic queries to OLAP using SPARQL.

8.2.6 Knowledge evaluation

The impact of BI in the organisations is measured by the company results [26]. BI is also useful for measuring the information quality of the companies [8]. BI is defined as the set of methodologies and tools that allows to transform the data in information and the information in knowledge, but the evaluation of such knowledge and the evaluation of the impact of a strategic action have not been successfully addressed so far. According to [5], the users of a BI platform must to be able to evaluate several knowledge representations. This essential skill is not provided by the current commercial solutions [26].

The evaluation of knowledge assets is difficult due to their intangibility [9]. [156] identifies the human capital as one of the most importance assets of a company. For evaluating human resources there is a method called 360 degrees [151], which consists in defining features for every organisation job. Each skill is defined with metrics, and periodical evaluations are performed using such skills and metrics. The name of the model comes from the fact that the evaluations are done by your manager, your colleagues, colleagues from other departments of the company, even by customers and, finally, also by the worker.

[9] defines several methods for evaluating knowledge assets, discussing about their advantages and disadvantages. These methods are: Skandia Navigator [155], IC-Index [158], Technology Broker [160], Intangible Asset Monitor [162], MVA and EVA [159], and Citation-weighted Patents [155].

8.3 Aims ot the thesis

BI requires the integration of several data sources. In many cases, data are distributed in heterogeneous systems or in formats where the classification of images or videos is nearly impossible. Furthermore, the lack of formal models for exploiting organisations is a problem with personalised analyses, and when linking to external resources or comparing companies is required.

The Semantic Web has been recognised as a core technological space in which modelling and exploiting knowledge of BI environments is feasible. Projects like EU MUSING have been pioneers in the use of the Semantic Web in BI solutions, although it mainly deals with unstructured information.

This thesis proposes methodologies and technologies for building BI solutions using semantic technologies that meet most of proposed essential requirements for this type of platforms. In the scope of BI 2.0, this proposal integrates information of social networks. The solution proposed is able to generate semantic questionnaires that add new knowledge and integrate a methodology for measuring and evaluating the knowledge. The following services offered by the solutions should be pointed out: (1) to perform personalised analysis of every company' activities, (2) to evaluate the knowledge in any organisation, (3) to recommend the actions to be carried out and (4) to analyse the impact of these actions in the organisation.

Hence, the main aim of this thesis is to research and develop methodologies and tools based on semantic technologies that can be integrated in a complete BI solution. The achievement of this goal requires performing the next tasks:

- Design and implementation of several mechanisms for the semantic classification of unstructured or multimedia contents.
- Design and implementation of an annotation model based on evaluation criteria that can be used for measuring and evaluating the knowledge.
- Design and implementation of a model that allows to generate semantic questionnaires that can be used to add new knowledge to a business resource.
- Design and implementation of a graphical designer of SPARQL queries driven by the ontology.
- Design and implementation of several models for evaluating the knowledge assets.
- Design and implementation of semantic profiles that represent a subset of the properties of a business resource with which the user can perform a personalized analysis.
- Design and implementation of semantic services for exploiting the data: advanced searchers, alert management, semantic dashboards, system for planning help, recommendations and impact analyses.
- Validation of the semantic BI solution in several domains as finances, medicine, and support to biomedical research and performance evaluation.

8.3.1 Research hypothesis

The main hypothesis is that semantic web technologies are the appropriate technological solution for building a complete BI solution. This hypothesis can be divided in the next sub-hypotheses:

- **Semantic web technologies permit to integrate heterogeneous data sources in a semantic repository. This integration allows to classify structured and unstructured data and allows to identify evaluation criteria of the generated knowledge.**
- **Semantic models are useful for generating new, structured information in the semantic repository.**
- **Semantic search helps to exploit the data in BI solutions.**

- **Semantic representations are useful to generate reduced, efficient, exploitation-oriented datasets.**
- **Semantic Web technologies and Semantic Social Networks are helpful to evaluate the knowledge assets of the company.**
- **Semantic Web technologies permit the development of advanced data exploitation services, such as: reporting, advanced searchers, semantic dashboards, planning, impact analysis and recommend systems.**

8.3.2 Methodology

The methodological steps followed in this thesis are described next.

- Analysis of the state of art:
  - Business Intelligence.
  - Semantic Web.
  - Web 2.0.
  - Information Integration.
  - On-Line Analytical Processing (OLAP).
  - Knowledge Evaluation.
- Methodological definition of the approach:
  - Study of the integration model for building a semantic repository.
  - Development of an integration process for adding new semantic information.
  - Development of a graphical tool for defining SPARQL queries.
  - Development of a generic model for knowledge evaluation.
  - Development of a method for defining reduced representations of data for efficient analysis.
  - Study and selection of solutions for exploiting the information generated in previous phases.
- Coding of the different proposed methods using semantic technologies.
- Validation of the semantic BI platform in several domains, like financial, performance evaluation, and biomedical research.

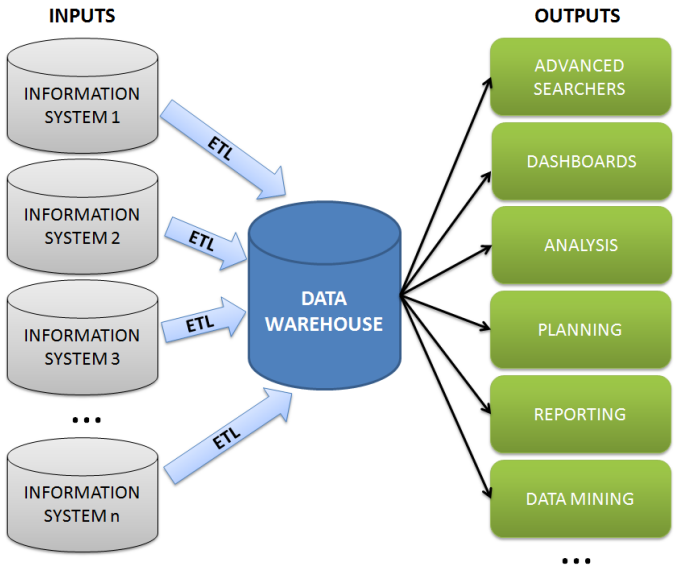


Figure 8.2: BI schema

8.4 Results

This section explains the set of methodologies and tools proposed in this thesis that can be used to create a semantic BI solution. Following the architecture described in [5], the approach has three service layers: integration, information delivery and analysis.

Figures 8.2 and 8.3 show the differences between a traditional model and our semantic approach, which deals with any type of data source, including natural language or multimedia content. Furthermore, our solution allows to extend the domain using semantic questionnaires.

8.4.1 Integration Information Model

8.4.1.1 Semantic repository

The semantic repository is in a central part of the architecture. The repository has two types of data sources: (1) an OWL files server with the formal representation of the domains, and (2) an RDF repository which stores the data. The ontologies guide all the layers of the solution: integration, information delivery and exploitation.

ODS (Ontology Driven-Searcher) is the service for information delivery.

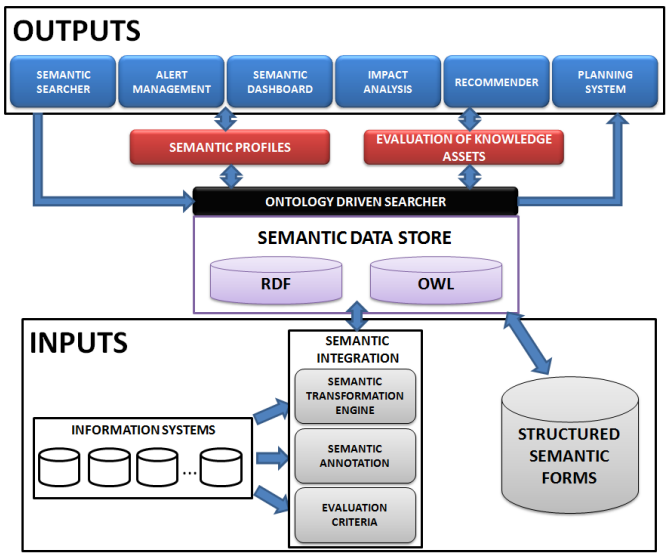


Figure 8.3: Semantic BI architecture

This tool is an editor of SPARQL queries supported by OWL models. The tool uses the underlying domain ontologies to show the necessary information to visually define SPARQL queries.

8.4.1.2 Transformation methodology

Our transformation methodology includes the next phases:

- Search and development of ontologies that model the domain of the organisation. In this phase, automatic transformation tools can be used.
- Integration of all the ontologies into a single, main one.
- Election of the tool or tools for the semantic transformation, prioritizing those tools in which the transformation process is guided by domain ontologies.
- Definition of the mappings between the source data schema and final semantic model. This phase is not necessary if syntactic transformation models are used.
- Information extraction. In this phase, we could use tools like D2RQ that allows to perform SPARQL queries over the source data. This kind of tool permits transforming data originally in RDF format.

- Generation of OWL or RDF information. It is better to use OWL format to guarantee the logical consistency of the dataset.
- Load the data in the RDF repository.

8.4.1.3 Semantic annotation

In this work, four types of annotations are addressed:

- **Automatic annotation.** It employs natural language processing techniques to extract knowledge entities from texts in natural language. Users can therefore introduce free-text content, which is more intuitive and usable for humans. Then, this content is processed and the semantic information gathered is stored in the knowledge base. The automatic annotation tool is inspired by the methodology proposed in [166] and uses the GATE framework [167]. The downside of this approach is that the success rate (i.e. the number of knowledge entities extracted from the text versus the actual number of knowledge entities available) is below one-hundred percent. Consequently, the platform is not taking advantage of all semantic information that can be gathered.
- **Manual annotation.** The user adds semantic information manually. For annotating: (1) the user selects the resource, (2) the user uses ODS for retrieving the ontology terms object of the annotation process, and (3) the annotation is stored in the semantic RDF repository.
- **Semi-automatic annotation.** In this case, it uses the automatic annotation, but only the annotations validated are stored in the RDF repository.
- **Semantic tagging.** It is defined as one or more SPARQL queries that return semantic elements for annotating a concrete resource. The result of this process is the creation of a set of tuples $\langle x, y, \langle z \rangle \rangle$, where x is the annotated resource, y is the concrete *owl:objectProperty* used for annotating and z is the query result.

8.4.1.4 Evaluation criteria annotation

Conceptually speaking, the evaluation criteria are defined as a reference point for measuring the suitability or the performance of a resource (i.e., person, activity, material, etc.) in a concrete moment in an organisation. The semantic representation of an evaluation criterion is an *owl:Class*. Table 8.4 describes

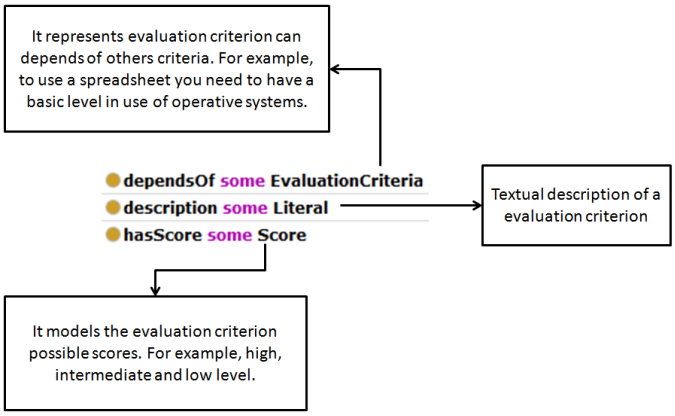


Figure 8.4: Semantic properties of an evaluation criterion

the semantic properties of this class. Furthermore, we can use *owl:subClassOf* for expressing that an evaluation criterion is a specialisation of another one. For example, spreadsheet could be a subclass of office solutions.

The methodology permits defining a criteria catalogue for evaluating one or more organisations. This model will be used for evaluating and planning the needs of the companies. It will also be useful for making multi-companies compartive analyses. This model of semantic annotation allows to know the functions and the minimum score of a business resource. The semantic representation of the resource annotation is the class *DesirableLevel*, whose relationships can be seen in figure 8.5.

8.4.2 Semantic Reports

Semantic reports are useful for adding information of a business resource. The data added are directly generated in a semantic format. We have defined a simple semantic model that can be adapted and extended to concrete scenarios. This semantic model has the next elements:

- *Report.* This class represents the information of a business resource. It has properties and relationships defined by the data managers for every use case.
- *Stage.* This class represents the phase of the resource lifecycle. Each stage is linked with one or more *Report*.

Figure 8.6 shows the phases of semantic reporting, which include:

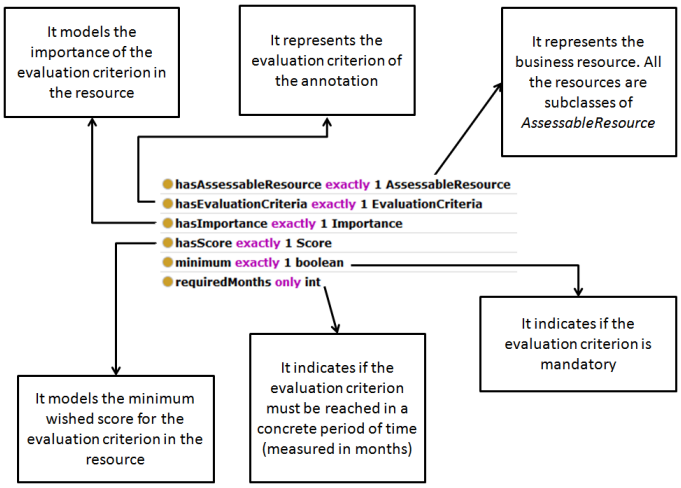


Figure 8.5: The annotation model for evaluation criteria

- Questionnaire. They are useful to extend the information of a concrete business resource.
- State machine. State machines are responsible for establishing the rules to fill the reports.
- Semantic translation. The previous elements generate an ontology that is useful as scheme to manage and exploit the new data generated.
- Semantic running engine. It generates web forms for adding and updating the information for each semantic report. Logically, it follows the rules defined in the report fields and in the state machines.

8.4.3 Knowledge Evaluation Model

This service proposes a Web 2.0 platform for evaluating the knowledge assets from different perspectives. Figure 8.7 shows the architecture of the Knowledge Evaluation model. The key aspect is the evaluation criteria annotation (described in 8.4.1.4). This process is useful to define evaluation criteria maps of the organisation resources. For example, we can define competency maps for the job positions of the organisations. Thanks to this model, the organisation can evaluate and compare the resources.

Figure 8.8 shows an excerpt of the evaluation criteria ontology. The main classes of the ontology are described next:

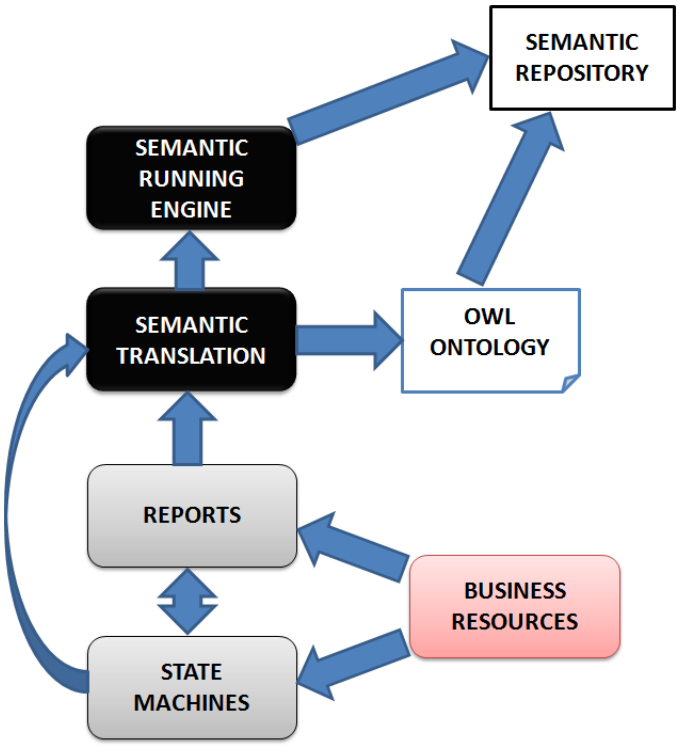


Figure 8.6: Semantic Report Methodology

- *EvaluationCriteria.* This class represents the minimum evaluation unit. This concept allows the definition of relations and hierarchies between criteria.
- *Score.* It represents the scores that a criterion can receive. For example, the knowledge criterion called office software could be assigned the values low level, intermediate level or high level.
- *AssessableResource.* This class models the hierarchy root of the items that can be evaluated. We allow to associate evaluation criteria over any organisation resource as human capital, infrastructures, processes, equipment, etc.
- *Importance.* This class represents the importance of the evaluation criteria through a numeric value. It is defined as an OWL class because it is important to describe the meaning of every value.
- *DesirableLevel.* This class permits to establish relations between the assessable resources and the evaluation criteria. *DesirableLevel* stores

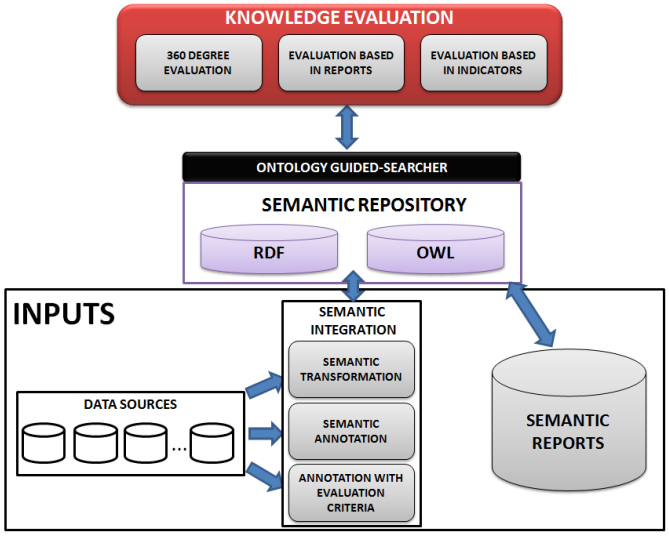


Figure 8.7: Knowledge Evaluation

information about the desirable *Score*, the *Importance*, or whether the criterion is mandatory or is reachable in a period. This concept allows to build an evaluation criteria map of any assessable resource.

- *Evaluator.* This class models a person that has evaluated resources in an evaluation call.
- *EvaluationCall.* This class represents an evaluation call. The call establishes an evaluation period, the resources to evaluate and the evaluators of every resource.
- *ResourceEvaluation.* This class represents the rating that an evaluator gives an evaluation criterion on an appeal in the context of a call for evaluation.
- *ResourceScore.* This class calculates the mean score of the different evaluations received by a resource for each criterion.

In this thesis three evaluations models have been described:

- **360 degree model.** This model is useful for measuring the intangible resources of an organisation. For example, the workers' knowledge, the processes efficiency, leadership, workers' attitude, etc. It is also useful for evaluating tangible resources that cannot be evaluated by activity indicators of the company.

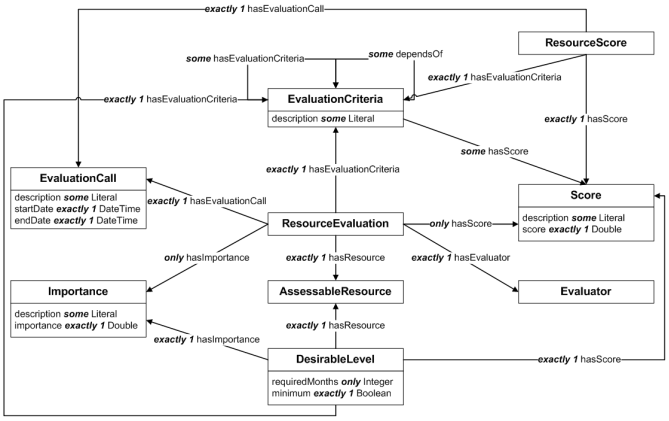


Figure 8.8: Evaluation criteria ontology

- **Evaluation based in reports.** In this model we use the semantic reports for defining tests and for evaluating the knowledge level or the social behaviour of a human resource. This model is complementary of the 360 degree. For example, with semantic reports we can evaluate the level of knowledge of a worker, and with 360 degree we can evaluate whether the worker applies this knowledge in her work position.

- **Evaluation based on indicators.** This model is useful to evaluate objectively the performance of an organisation activity. In this case, the information is retrieved from the semantic repository. For extracting the data the users can use SPARQL queries defined with ODS. This means that every user can define its own indicators and share or compare with others defined by other users.

The previous models can be used both separately and jointly, using indicators for evaluating the tangible resources, 360 degree for evaluating intangible assets and semantic reposts for measuring the knowledge level or social behaviour of the human resources.

8.4.4 Semantic profile of a Business Resource

A semantic profile is the subset of semantic information of a business resource that is interesting for a particular analysis. The profile of the resource i is calculated as:

$$PS(r) = S(d) \cup S(PS(ir)) \tag{8.1}$$

where $N(d)$ represents a subset of the selected *owl:datatypeProperty* and $SP(o)$ represents a recurrent call of the individuals linked through *owl:objectProperty* to i. ODS can be used for building the semantic profile. First, the user selects in ODS for the appropriate fields. Then, she uses ODS for filtering the results. ODS allows the use of aggregate functions like count, average, min, max, etc. to generate the values that will be finally returned.

Conceptually speaking, the semantic profile is defined as a set of relations and properties that some resources have. Semantic profiles permit to identify groups of resources that share some properties and are therefore useful for comparing and studying such groups. Ontologies are of special interest for creating profiles because they allow for aggregation and selection of individuals from a conceptual perspective. Our approach can also generate the semantic profile of a group of resources by applying one or more criteria.

8.4.5 Exploitation Model

This proposal includes a set of methods for exploiting the information stored in the semantic repository:

- **Semantic searcher.** This tool uses ODS for defining queries over the semantic data model.
- **Alert management.** This tool allows to generate alerts over the semantic data. It uses ODS for defining the alerts as queries and comparing the results with thresholds when these have been defined. For example, if the calculation of an indicator returns a value lower than 50 the user may receive a high critical alert over the activity performance related with the indicator.
- **Semantic dashboard.** This tool permits users to formulate incremental, user-defined queries with a graphical user interface based on ODS. The query results can be displayed in several customisable ways, allowing for the generation of on-demand dashboards.
- **Recommendation module.** This tool uses two approaches for recommending personalised actions to concrete users based on its needs. The first approach uses similarity functions for generating groups of resources with similar properties. The classification of a resource is proposed by this method when this has a good evaluation. The second approach uses bayesian networks to predict the value of a variable or a resource using groups of resources filtered by concrete criteria.

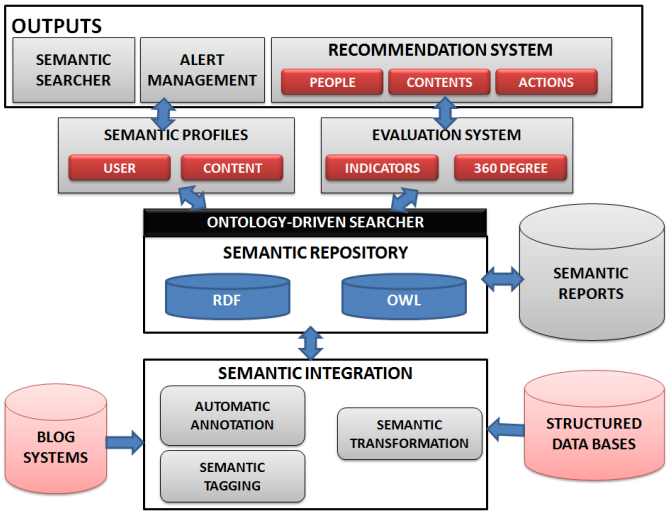


Figure 8.9: Architecture of the semantic social network

- **Planning system.** This method has two profiles. The first profile allows to detect the strengths and weaknesses of a company, thanks to the Knowledge Evaluation model. The second profile uses these results and the annotation of threats and opportunities for designing a plan adapted to the special requirements of each organisation unit (i.e, departments, services, etc.).

- **Impact analysis.** Impact analysis uses three data sources to measure the impact of a strategic plan. The first source of data is a knowledge evaluation before the execution of the plan, the second is the own plan, and the last one is the results of a further knowledge evaluation call. With this information, the system is able to offer what actions included in the strategic plan have been positive/negative for the company. This information is useful for the Planning System.

8.4.6 Semantic Business Intelligence Solutions

8.4.6.1 Semantic social network

In this section we present a semantic platform that permits the exploitation of the generated knowledge by users of a social network or corporate intranet.

Figure 8.9 shows that the semantic repository is the core of the system. The data sources are: (1) systems based on blogs and structured data bases, and (2) the definition of semantic reports. In addition to this, a set of BI

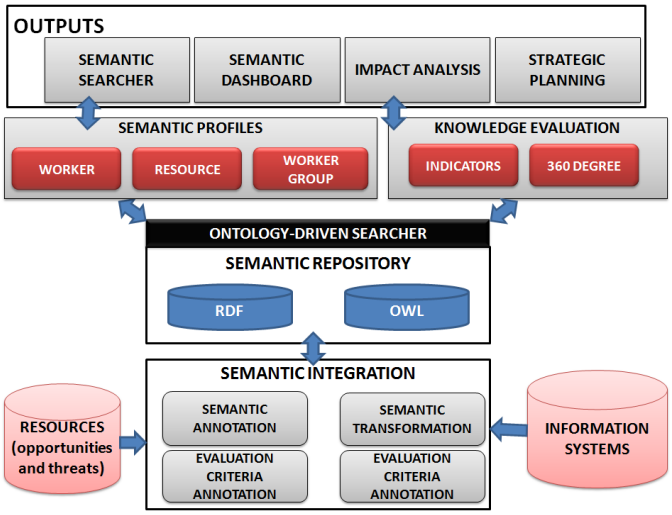


Figure 8.10: Architecture of the strategic planning platform

services are integrated for improving the knowledge management. Among these services, we highlight the content and user recommendation system, and the knowledge evaluation system.

8.4.6.2 Semantic platform for strategic planning

Strategic planning is a key process for companies. BI solutions offer a snapshot of the current situation of the company. For this reason, they are useful to help to the definition of a strategic plan. SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis is one of the most used, just before the performance of plans [174], which is the approach followed in this thesis.

Several semantic BI tools have been integrated to identify the strengths and weakness of the organisation knowledge. The platform is able to classify the opportunities and threats, for both internal and external. Furthermore, the platform can measure the impact of the strategic plan in the organisation. Figure 8.10 shows that semantic repository is the core of the system. The source data are: (1) data of the information systems for calculating the activity indicators, and (2) the semantic annotation of opportunities and threats. The main services of this platform are the automatic plan definition and the impact analysis. Furthermore, other services like semantic searchers or customisable dashboards are available.

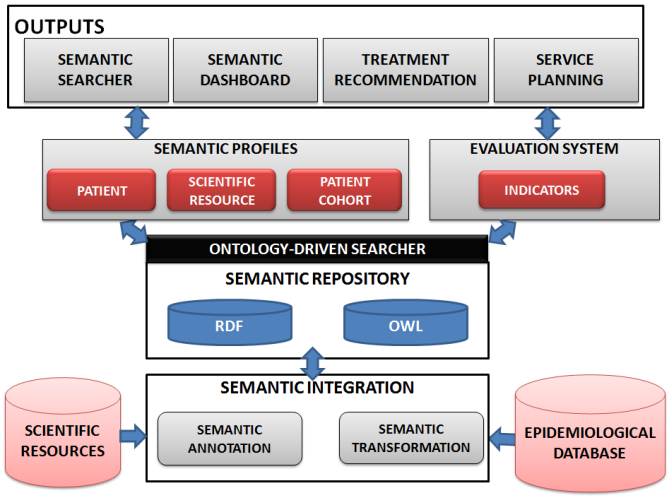


Figure 8.11: Architecture of the epidemiological analysis platform

8.4.6.3 Semantic platform for epidemiological analysis

In the last 20 years, epidemiology has become a basic tool for the disease research and treatment [173]. Nowadays, there are several technological solutions able to manage and analyse the information of patients with a determined diagnostic [175]. However, the lack of formal semantic models is a problem when personalised analyses or external data links are required. To solve this problem, a semantic model for the exploitation of epidemiological databases has been developed [176]. We also use the semantic BI tools for analysing the information in a dynamic and agile way.

Figure 8.11 shows the complete architecture of the system. The source data are: (1) the Electronic Health Record (EHR) that will be semantically transformed, and (2) the scientific resources related with the domain.

In this platform, we are able to recommend treatments of a patient or to predict service burden in the future using bayesian networks.

8.4.6.4 Semantic case report form

Biomedical researchers need software solutions that allow work with very heterogeneous and changing information, depending of the concrete project. A case report form (CRF) is a set of questionnaires used to save the captured data of each patient recruited in a biomedical project.

Many technological solutions are available to manage data for CRF nowadays. In these solutions we can find two approaches: (1) use of relational

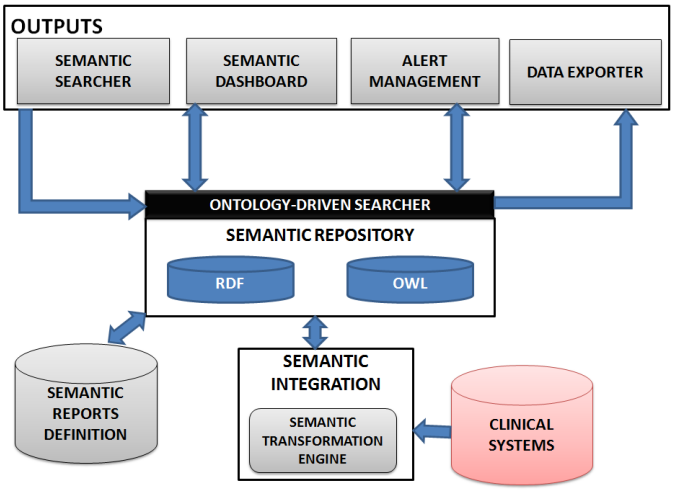


Figure 8.12: CRF architecture

databases to store the data; and (2) use of non-relational databases to store the data. The main disadvantage of the first approach is the little flexibility of the relational model to modify their structure, without changing the software. The main disadvantage of the second approach is the lack of a well-defined model for exploitation, generation of alerts, or quality assurance of the data. In this context, the use of semantic web technologies allows us to store biomedical data in a flexible data model and to exploit this information thanks to the semantic model that describes the data. Furthermore, these technologies permit the reuse of biomedical ontologies and the semantic interoperability of health resources are required.

Figure 8.12 shows the approach of the architecture. In this case we use the semantic reports to define the CRF. The platform can optionally use the semantic transformation engine to integrate EHR data. Finally, the platform offers the next services: (1) semantic searcher, (2) customisable dashboards and (3) alert management for recruitment processes.

8.4.6.5 Semantic platform for multimedia content analysis

The classification of multimedia contents is a complex task for the companies. Nowadays, this type of content is classified using keywords. This process is useful to simplify searches. However, the lack of formal models is a problem when the users need to compare several multimedia contents or want to link other contents. The Semantic Web is considered as a useful tool for classifying and exploiting this type of content.

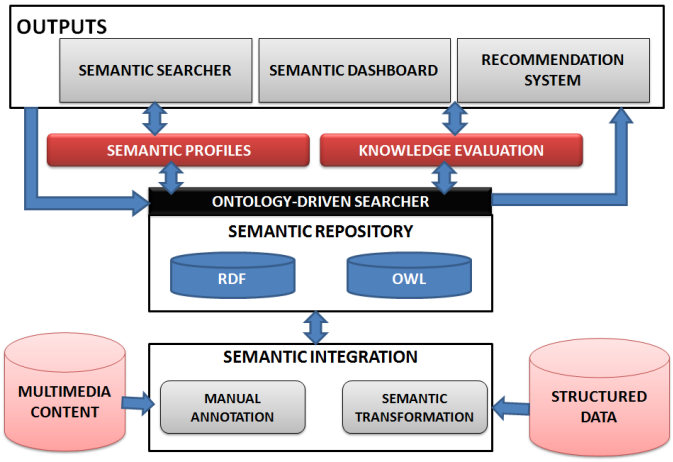


Figure 8.13: Architecture for multimedia content analysis

In the biohealth domain, it is very common to use images for several tasks like radiology, microscopy, etc. Clinical terminologies are usually employed for classifying these images. In this case, we propose a platform that semantically classifies biomedical images. Thanks to this process, the users can use the BI tools for exploiting the multimedia resources.

Figure 8.13 shows the architecture of this platform. In this case, the inputs are: (1) the multimedia content manually annotated, and (2) structured data for doing the annotations. These are the services for the users: (1) semantic searcher, (2) customisable dashboard and (3) a recommendation system to reuse the annotations between images.

8.5 Validation

8.5.1 SocialBROKER: Semantic social network in financial domain

SocialBROKER is a prototype implementation of the proposed platform for semantic social networks customized for the finances domain. It allows for creating collaborative groups that define financial-related initiatives and proposals. The integration of Semantic Web concepts and Social Networks enable our system to automatically create semantically-empowered relationships between the investors registered within the platform based on their interactions. This further allows the system to analyse the social behaviour of the network members and to generate a customised profile for each of

them.

For this use case scenario, we have built a financial ontology based on the existing ontologies. The results of this platform have been evaluated using four parameters: resource usage, performance levels, comparison with other finances services such as Google Finances or Yahoo Finances, and the comparison with other social networks like IMDB.

8.5.2 Semantic planning of training

This platform has been developed in collaboration with a local hospital, using the proposed Semantic platform for strategic planning. Based on the requirement analysis made, an ontology was developed that models the semantics of competencies needed for 83 hospital services. We have used this model and SWIT for transforming and storing the organisational data from the human resources management system in a semantic data store. We have implemented a Semantic Web platform that permits the next tasks: management of evaluation calls, management of competency catalogues and evaluations based on the 360º methodology. The users may formulate incremental, customised, queries with a graphical user interface based in ODS. The query results can be displayed in several customisable ways, allowing the generation of dashboards on demand.

The platform is active since 2013. Since then, the managers of the platform have defined 21 evaluation calls and it has been used to evaluate 377 workers. Nowadays, there are 1,520 users of the platform including evaluators and managers. The mean of the desirable levels of the different clinical departments is 4.33 (out of 5) and the mean of the last evaluation was 3.95, obtained at the end of 2014. Nowadays, there are 24 evaluation calls open for evaluating the replacement staff summer holidays in 12 clinical departments. Thanks to the performance evaluation process, plans of training have been generated automatically.

8.5.3 SECARE: Semantic exploitation of a local cancer registry

The approach described in the semantic platform for epidemiological analysis has been applied in an institutional cancer registry scenario. Based on the requirement analysis [189; 190], an ontology was developed that models the semantics of an institutional cancer registry. We have used this model and SWIT for transforming and storing simulated data from cancer registries in a semantic data store. We have implemented a Semantic Web platform that

permits users to formulate incremental, user-defined queries with a graphical user interface based in ODS. The query results can be displayed in several customisable ways, allowing the generation of on-demand dashboards. The complex timelines of the disease of individuals and aggregated patients can be clearly represented.

The main result of this study is the development of a Semantic Web platform that facilitates the analysis and visualisation of data from cancer registries including (1) the representation of the disease course of a patient, (2) the representation of the aggregated disease courses of a group of patients, and (3) the definition of customisable dashboards for patient selection and visualisation of the data.

8.5.4 SECOLON: Semantic exploitation of a local colorectal cancer screening

Early detection of colorectal cancer has become an essential tool to combat this disease [200]. The prevention program running in the Region of Murcia aims to track people between 50 and 69 years old and to diagnose the level of risk for contracting this disease. In this case, we have collaborated with the people who coordinate the program for developing a Web platform that exploits this record. To do this, we have developed an ontology that models the process of screening for this type of cancer. In addition, we have used different methods of semantic IN our proposal to build a platform to exploit the program data and recommend what level of risk will have a patient in the future.

The development of this prototype is based on the solution proposed in the semantic platform for epidemiological analysis. SECOLON is a semantic platform that exploits a record of the screening of a patient from two perspectives: (1) reporting and semantic control custom boxes, and (2) recommendation on the levels of risk that a patient will have in the future. In this case we have had the opportunity to work with real data from 322,839 patients recruited since 2006, facilitated by the Colorectal Cancer Prevention Program of the Ministry of Health of the Region of Murcia.

8.5.5 Semantic CRF for NELA Project

The approach described in the Semantic Case Report Form has been applied in the NELA Project. The aim of the project NELA [181] is to investigate the impact of nutrition during pregnancy and early life on the health of children in the Region of Murcia. The platform is completely functional since

January 2015. Currently there are three data managers using the platform. Furthermore, we have 98 patients that have filled its own questionnaire. We have 102 users registered in the platform.

The data managers of NELA have used ODS, and they have stored filters for the generation of three semantic dashboards: (1) graphical representation of the patients by age; (2) graphical representation of patients with invasive diagnostic test during pregnancy; and (3) graphical representation of the patients by consumption of drugs (tobacco, alcohol, cocaine, etc.).

8.5.6 Semantic annotator for EUCOMM Tools Project

The EUCOMM project (tools for functional annotation of the mouse genome) [182] is a member of the European project IKMC (International Knockout Mouse Consortium). Its main objective is to generate a database of mutations in protein-coding genes. The annotation process consists in using images to describe whether a particular gene is not expressed in a given mouse or with specific conditions. A very important fact for the annotation is in which anatomical location the expression occurs. For annotating the images, the anatomy of the brain and the body of the mouse have been described.

The development of this platform is based on the solution for multimedia content described as the Semantic platform for multimedia content analysis. The platform was launched in late 2013 and the annotation process will be completed by the end of 2015. The main results are:

- 137 mouses recorded.
- 964 images associated with these mouses.
- 110,570 semantic annotations.
- 110,570 entries, of which 66,144 were generated reusing our recommendation engine.
- 36 sets of images have been generated, and their semantic similarity is above 90%.

8.6 Discussion and future work

[2] defines thirteen fundamental features for any BI solution. These features can be classified in three levels: integration, information delivery and analysis. Next, we describe how our solutions addresses each of them:

- **Integration**
  - **BI infrastructure.** The semantic alternative is the RDF repository, which centralises the data of the organisation. In this case, the advantage of using Semantic Web technologies is that the repository can be exploited by software agents.
  - **Metadata management.** This is supported by the OWL repository. The ontologies guide all the lifecycle of the solution, from integration to exploitation and analysis.
  - **Development tools.** The data are stored following linked data principles. This eases the development of other exploitation tools. Furthermore, the customisable capabilities of our solution allows to generate advanced searches and to perform new queries for specific analyses without requiring IT expertise. In addition to this, the semantic reports permit adding new semantic information of any resource.
  - **Collaboration.** The knowledge evaluation model is an example of collaboration in a BI solution. Concretely, the 360 degree evaluation permits the collaborative evaluation of a resource. The development of a semantic social network also allows the exploitation of the typical unstructured contents generated in this type of platforms.
- **Information delivery**
  - **Reporting.** We have developed two types of reports. The first consists in generating customisable data tables with ODS. The second one consists in defining alerts using again ODS.
  - **Dashboards.** ODS also allows the generation of customisable dashboards, thanks to its capabilities of query aggregation.
  - **Ad hoc query.** The semantic alternative is ODS. As we have seen, ODS allows to navigate the platform data in a personalized way.
  - **Microsoft Office integration.** This is not addressed in this thesis, because the Semantic Web technologies do not contribute to improve this task.
  - **Search-based BI.** ODS allows to persist the queries and to identify the search parameters. This capability generates customisable advanced searches over any business resource.

- **Analysis**
  - **OLAP.** The semantic profiles are the alternative to OLAP. Ontologies are of special interest for creating profiles, because they allow for aggregation and selection of individuals from a conceptual perspective.
  - **Interactive visualisation.** The reports and the dashboards permit to navigate the concrete data collection and their visual representation.
  - **Predictive modeling and data mining.** The recommendation system is our solution for this feature. Our approach implements two methods based on: (1) similarity functions and (2) bayesian networks.
  - **Scorecards.** Thanks to the definition of semantic dashboards and the evaluation model, the users can see easily the evolution of the company and the current state. They can even see why the company has such performance indicators.

We have evaluated the platform in five domains: finances, epidemiology, training planning, clinical research and biomedical research.

The advantages of use semantic web technologies in BI are:

- The ontologies guided all the lifecycle of the BI solution, so the same vocabulary is used in the several phases.
- The way how SPARQL queries are built is more intuitive for human beings than relational queries.
- The semantic representation of the queries allows its serialisation and reuse in other scopes.
- Ontologies are of special interest for creating semantic profiles because they allow for aggregation and selection of individuals from a conceptual perspective.
- The use of Linked Data allows the exploitation, sharing and comparison of the information between different companies.
- The use of semantic reports allows the construction of new information models to extend the data sources.

The main limitations of this work are:

- The precision of automatic annotation is not one hundred percent, what may generate incorrect annotations, that is, false positives. Furthermore, this process is not useful for multimedia content as images or videos.
- The semantic reports do not take advantage of the OWL expressiveness for defining rules. It only allows the definition of ranges and cardinalities.
- The semantic profiles may generate redundant information in the repository. The platform administrators should decide if the semantic profile is commonly used in various analysis or not for maintaining such redundant data.
- We need more evaluation of the semantic social network, because we only used a simulated data source.
- The performance evaluation of human resources in a hospital has only been applied to temporary workers. The training planning has only been evaluated with a small set of learning resources without contents.
- The service planning in the epidemiological platform only works with recruitment patients. It does not take into account the inclusion of new patients.
- The ontologies of SECARE, SECOLON and EUCOMM are preliminary versions. These ontologies need to be reviewed and extended, although they have served to demonstrate that the semantic exploitation of this type of clinical data is possible in a robust and scalable way.
- The recommendation system only uses two algorithms.

As future work we propose:

- Extension of SocialBROKER with new data sources as financial news or economic laws.
- Integration of the Semantic CRF with the information of Electronic Health Records.
- Integration of the epidemiological data with learning resources.
- Evaluation of more resources of the hospital for improving the planning.
- Extension of the number of algorithms for data mining and predictions.

8.7 Hypothesis verification

The main hypothesis in this work consisted in using semantic web technologies for building a complete BI solution. Such hypothesis was divided in sub-hypotheses, which can be demonstrated through the answers to the next questions:

- **Semantic web technologies permit to integrate heterogeneous data sources in a semantic repository. This integration allows to classify structured and unstructured data and to identify criteria for the evaluation of the generated knowledge.** In section 8.2.4 several integration methodologies have been described. In section 8.2.2.1 we propose different tools for semantic transformation. To classify semantically-structured content, a transformation methodology has been defined in section 8.4.1.2. To annotate semantically-unstructured data, several annotation models were defined in section section 8.4.1.3: automatic, semi-automatic, manual and semantic tagging. An innovative model of annotation was proposed in 8.4.1.4. This process is able to classify knowledge assets with evaluation criteria. In SocialBROKER (view section 8.5.1) we have demonstrated the effective use of the automatic annotation, the evaluation of knowledge assets and the transformation of structured data. In EUCOMM-Tools project we validated our manual annotation system.

- **Semantic models are useful for generating new, structured information in the semantic repository.** In section 8.4.2 we have defined the "Semantic Reports" model. This model is useful to add new semantic information to any business resource. This process is aligned with the semantic repository, allowing the use of the analysis tool for exploiting the information added. In the use case described in section 8.5.5 we have demonstrated the use of the "Semantic Reports" in a biomedical project named NELA [181].

- **Semantic search helps to exploit the data in BI solutions.** ODS (Ontology-driven searcher) is a tool that allows the performance of SPARQL queries in a graphical way. This tool permits the generation of advanced searchers and semantic dashboards (thanks to aggregation capabilities). All the generated queries can be serialized in the semantic repository. Furthermore, the queries can be saved and parametrized, allowing for its reuse in further analyses.

- **Semantic representations are useful to generate reduced, efficient, exploitation-oriented datasets.** In section 8.2.5 we have defined OLAP as a model for the analysis and exploitation of the data warehouse. In section 8.4.4 we have defined the "Semantic Profile", which is our alternative to OLAP. The "Semantic Profile" is defined as a set of relations and properties shared by some resources. This model has been validated in several use cases: SECARE (view 8.5.3), SECOLON (view 8.5.4) and SocialBROKER (view 8.5.1).

- **Semantic Web technologies and Semantic Social Networks are helpful to evaluate the knowledge assets of the company.** In section 8.2.6 we have described the state of art of the evaluation of knowledge assets. In section 8.4.3 we have defined a method for integrating a model of evaluation of knowledge assets in our semantic BI solution. This method has been validated in the use case SocialBROKER (view 8.5.1) and Planning of Training (view 8.5.2).

- **Semantic Web technologies permit the development of advanced data exploitation services, such as: reporting, advanced searchers, semantic dashboards, planning, impact analysis and recommend systems.** In section 8.4.5 we have described several tools for exploiting the data of the semantic repository. The proposed tools have been validated in the several use cases. We highlight the evaluation of the recommendation model in the use cases SECARE and SECOLON to predict the applied therapy for a patient with a particular diagnosis.

8.8 Contributions

This thesis has provided a solution for business intelligence using semantic technologies. The solutions we have discussed have been successfully implemented and validated in a number of use cases:

- **Semantic social networks.** SocialBROKER is a prototype of a semantic social network in finances domain.

- **Semantic platform for strategic planning.** This solution is used to evaluate the performance of human resources in a real hospital. Training plans adapted to the specific requirements of each worker are automatically generated. This platform is active since 2013.

- **Semantic platform for epidemiological analysis.** SECARE and SECOLON are two examples of the advantages of semantic BI applied to epidemiological data analysis.
- **Semantic case report form**. The recruitment process of the NELA project is completely managed by our Semantic CRF approach.
- **Semantic platform for multimedia content analysis**. The project EUCOMMTools has used our BI solution to classify and analyse a set of images that describe the gene expression of mouses in several stages.

The main contributions that can be extracted from this thesis are:

- Design of an evaluation model based on semantic knowledge technologies. The definition of the evaluation criteria and the semantic annotation model allows the generation of profiles that should be held by each resource. Furthermore, it allows to evaluate the resources from three different methodologies:
  - 360 degree evaluation model.
  - Model based on indicators.
  - Model based on reports.
- Design of a semantic model for generating resource profiles. Our approach, equivalent to OLAP systems in traditional BI, consists in defining a semantic reduced representation of a resource. Using these profiles, users can easily configure more efficient analyses, speeding up the process of data exploitation.
- Design and implementation of a SPARQL queries generator driven by ontologies. ODS can generate SPARQL queries from an intuitive, graphical user environment, guided by the ontology concepts and properties. In addition to generating the queries, ODS can also filter what information you want to recover, sort the results, and use aggregate functions (count, sum, average, maximum, minimum, etc.) under normal or aggregate queries. The query generation model defined in ODS can be later reused for other searches or for calculating indicators. This serialisation also allows the user to generate forms of advanced filter, checking the fields that are required as parameters.
- Design of a model for the generation of semantic reports. The semantic model for reporting allows the definition of concepts, relationships and properties that can complete the information for a particular resource.

- Design of a model for the recommendation based on similarity functions. This model has been applied to annotate microscopic images of transgenic mouses, obtaining that more than 60% of the annotations have been reused thanks to the recommender.
- Design of a model for the recommendation based on Bayesian networks. This model has predicted, with a high success rate, the level risk for colon cancer in patients in a screening program.
- Customisable semantic dashboards.
- Three out of the six evaluation use cases are running in real environments. The platform for training planning is running since 2013 in a real hospital. The recruitment process of the NELA project is being managed by our semantic CRF. Finally, the identification of gene expression in mouses is managed by the semantic annotator for the EUCOMM Tools project.

8.9 General conclusions

BI is a tool that has a large impact on organisations. There are many IN solutions in the market, but the lack of formal semantic models is a problem when custom analyses, links with external resources and the comparison between different entities are required. In addition, current solutions do not have mechanisms to exploit unstructured content from currently common solutions such as social networks and Web 2.0 platforms. Another common shortcoming is the lack of tools to evaluate the knowledge generated, and the use of such evaluation results for future decision-making.

In this thesis we have defined a set of methodologies and implemented a set of tools to integrate, deliver information and analyse data using semantic technologies. These technologies allow users, without the help of ICT experts, to configure advanced searches, build custom dashboards, make strategic plans, evaluate their impact, and establish complex analyses from semantic profiles. Furthermore, thanks to the use of semantic technologies, we have established the bases to link to external data sources and comparative analysis with other organisations.

The developed solutions have been evaluated in different domains: finances, epidemiology, performance evaluation, clinical research and biomedical research, with satisfactory results and with three of them deployed in real environments. This work is an example of how ontologies can guide the entire

lifecycle of a BI tool, from data integration, to exploitation and knowledge generation.

Bibliografía

[1] Luhn HP (1958) A business intelligence system. IBM Journal of Research and Development 2: 314–319.

[2] Sallam RL, Richardson J, Hagerty J, Hostmann B (2011) Magic quadrant for business intelligence platforms. Gartner Group, Stamford, CT .

[3] Power DJ, Sharda R (2012) Business Intelligence and Analytics. Wiley Encyclopedia of Management .

[4] Doan A, Ramakrishnan R, Halevy AY (2011) Crowdsourcing systems on the world-wide web. Communications of the ACM 54: 86–96.

[5] Bitterer A, Schlegel K, Hostmann B, Gassman B, Beyer MA, et al. (2007) Hype Cycle for Business Intelligence and Performance Management, 2007. Gartner Research, Stamford, CT .

[6] Inmon WH (2005) Building the data warehouse. John wiley & sons.

[7] Watson HJ, Wixom BH (2007) The current state of business intelligence. Computer 40: 96–99.

[8] Popovic A, Coelho PS, Jaklic J (2009) The impact of business intelligence system maturity on information quality. Information research 14.

[9] Bontis N (2001) Assessing knowledge assets: a review of the models used to measure intellectual capital. International journal of management reviews 3: 41–60.

[10] Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Scientific American 284: 34–43.

[11] Saggion H, Funk A, Maynard D, Bontcheva K (2007) Ontology-based information extraction for business intelligence. Springer.

[12] Maynard D, Yankova M, Kourakis A, Kokossis A (2005) Ontology-based information extraction for market monitoring and technology watch. In: ESWC Workshop End User Apects of the Semantic Web, Heraklion, Crete.

[13] Niemi T, Toivonen S, Niinimaki M, Nummenmaa J (2007) Ontologies with Semantic Web/grid in data integration for OLAP. International Journal on Semantic Web and Information Systems (IJSWIS) 3: 25–49.

[14] Nebot V, Berlanga R, Pérez JM, Aramburu MJ, Pedersen TB (2009) Multidimensional integrated ontologies: a framework for designing semantic data warehouses. In: Journal on Data Semantics XIII, Springer. pp. 1–36.

[15] Niinimaki M, Niemi T (2009) An ETL process for OLAP using RDF/OWL ontologies. In: Journal on Data Semantics XIII, Springer. pp. 97–119.

[16] Declerck T, Krieger HU, Saggion H, Spies M (2008) Ontology-Driven Human Language Technology for Semantic-Based Business Intelligence. In: ECAI. pp. 841–842.

[17] Skoutas D, Simitsis A (2006) Designing ETL processes using semantic web technologies. In: Proceedings of the 9th ACM international workshop on Data warehousing and OLAP. ACM, pp. 67–74.

[18] Cyganiak R, Reynolds D, Tennison J (2014) The rdf data cube vocabulary (2014). `http://www.w3.org/TR/vocab-data-cube/`.

[19] Etcheverry L, Vaisman A (2012) QB4olap: a new vocabulary for OLAP cubes on the semantic web. Proceedings of COLD .

[20] Moss LT, Atre S (2003) Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-support Applications. Addison-Wesley Professional.

[21] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, et al. (2008) Top 10 algorithms in data mining. Knowledge and Information Systems 14: 1–37.

[22] Williams S, Williams N (2010) The profit impact of business intelligence. Morgan Kaufmann.

[23] Panian Z (2009) Expected progress in the field of business intelligence. In: Proceedings of the 8th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases. World Scientific and Engineering Academy and Society (WSEAS), pp. 170–175.

[24] Chaudhuri S, Dayal U, Narasayya V (2011) An overview of business intelligence technology. Communications of the ACM 54: 88–98.

[25] Turban E, Sharda R, Aronson JE, King D (2008) Business intelligence: A managerial approach. Pearson Prentice Hall Upper Saddle River.

[26] Chen H, Chiang RH, Storey VC (2012) Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly 36: 1165–1188.

[27] Imhoff C, Galemmo N, Geiger JG (2003) Mastering data warehouse design: relational and dimensional techniques. John Wiley & Sons.

[28] Chan JO (2015) Optimizing Data Warehousing Startegies. Communications of the IIMA 5: 1.

[29] Walker DM (2006) Overview architecture for enterprise data warehouses. Data Management & Warehousing, Ver 1.

[30] Moniruzzaman ABM, Hossain SA (2013) Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. arXiv preprint arXiv:13070191 .

[31] Bonnet L, Laurent A, Sala M, Laurent B, Sicard N (2011) Reduce, you say: What nosql can do for data aggregation and bi in large repositories. In: Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on. IEEE, pp. 483–488.

[32] O'reilly T (2007) What is Web 2.0: Design patterns and business models for the next generation of software. Communications & strategies : 17.

[33] Lusch RF, Liu Y, Chen Y (2010) The phase transition of markets and organizations: The new intelligence and entrepreneurial frontier. IEEE COMPUTER SOC 10662 LOS VAQUEROS CIRCLE, PO BOX 3014, LOS ALAMITOS, CA 90720-1314 USA.

[34] Giles M (2011) Beyond the PC. The Economist .

[35] Chen H (2011) Editorial: Design science, grand challenges, and societal impacts. ACM Transactions on Management Information Systems (TMIS) 2: 1.

[36] Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2: 1–135.

[37] Tuzhilin A (2010) Towards the Next Generation of Recommender Systems. In: Proceedings of the 1st International Conference on E-Business Intelligence (ICEBI2010),. Atlantis Press.

[38] Anderson C (2013) The long tail. Wereldbibliotheek.

[39] Wang MCGA, Chen XZH, Mao DZW (2011) Intelligence and Security Informatics .

[40] Chen H (2009) AI, E-government, and Politics 2.0. Intelligent Systems, IEEE 24: 64–86.

[41] Yang H, Callan J (2009) OntoCop: Constructing Ontologies for Public Comments .

[42] Brumfiel G (2011) High-energy physics: Down the petabyte highway. Nature News 469: 282–283.

[43] Hanauer DA, Rhodes DR, Chinnaiyan AM (2009) Exploring clinical associations using "-omics" based enrichment analyses. PLoS One 4: e5203.

[44] Hanauer DA, Zheng K, Ramakrishnan N, Keller BJ (2011) Opportunities and challenges in association and episode discovery from electronic health records. IEEE Intelligent Systems 26: 83–87.

[45] Lin Y, Brown RA, Yang HJ, Li S, Lu H, et al. (2011) Data mining large-scale electronic health records for clinical support. IEEE Intelligent Systems 26: 87–90.

[46] Wactlar H, Pavel M, Barkis W (2011) Can computer science save healthcare? IEEE Intelligent Systems 26: 79–83.

[47] Berners-Lee T. Uniform Resource Identifiers (URI): Generic Syntax. `http://tools.ietf.org/html/rfc2396`. Último acceso: Mayo 2015.

[48] W3C. Extensible Markup Language (XML). `http://www.w3.org/XML/`. Último acceso: Mayo 2015.

[49] W3C. XML Schema. http://www.w3.org/XML/Schema. Último acceso: Mayo 2015.

[50] RDF Working Group. Resource Description Framework (RDF). http://www.w3.org/RDF/. Último acceso: Mayo 2015.

[51] W3C. Resource Description Framework Schema (RDFS) 1.1. http://www.w3.org/TR/rdf-schema/. Último acceso: Mayo 2015.

[52] Brewster C, O'Hara K (2004) Knowledge representation with ontologies: the present and future. IEEE Intelligent Systems 19: 72–81.

[53] Gruber TR (1993) A translation approach to portable ontology specifications. Knowledge Acquisition 5: 199–220.

[54] Shahar Y, Musen MA (1996) Knowledge-based temporal abstraction in clinical domains. Artificial intelligence in medicine 8: 267–298.

[55] Schulz S, Romacker M, Faggioli G, Hahn U (1999) From knowledge import to knowledge finishing: automatic acquisition and semi-automatic refinement of medical knowledge. In: Proceedings of the Banff Knowledge Acquisition for Knowledge-Based Systems Workshop. Citeseer.

[56] Dieng R, Corby O, Giboin A, Ribiere M (1999) Methods and tools for corporate knowledge management. International journal of human-computer studies 51: 567–598.

[57] Schwartz DG (1999) When email meets organizational memories: addressing threats to communication in a learning organization. International journal of human-computer studies 51: 599–614.

[58] Fernández-Breis JT (2003) Un entorno de integración de ontologías para el desarrollo de sistemas de gestión del conocimiento. Tesis Doctoral, Universidad de Murcia. https://digitum.um.es/xmlui/handle/10201/185.

[59] Stevens R, Wroe C, Lord P, Goble C (2004) Ontologies in Bioinformatics. In: Staab PDS, Studer PDR, editors, Handbook on Ontologies, Springer Berlin Heidelberg, International Handbooks on Information Systems. pp. 635–657.

[60] Zhang Z, Zhang C, San Ong S (2000) Building an ontology for financial investment. In: Intelligent Data Engineering and Automated Learning?IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents, Springer. pp. 308–313.

[61] Brase J, Nejdl W (2004) Ontologies and Metadata for eLearning. In: Handbook on ontologies, Springer. pp. 555–573.

[62] Guarino N (1998) Formal Ontology and Information Systems. IOS Press, pp. 3–15.

[63] W3C. OWL Web Ontology Language. `http://www.w3.org/TR/owl-features/`. Último acceso: Mayo 2015.

[64] Brickley D, Miller L (2012) FOAF vocabulary specification 0.98. Namespace document 9.

[65] Breslin JG, Decker S, Harth A, Bojars U (2006) SIOC: an approach to connect web-based communities. International Journal of Web Based Communities 2: 133–142.

[66] Miles A, Bechhofer S (2009) SKOS simple knowledge organization system reference. W3C recommendation 18: W3C.

[67] W3C. OWL 2 Web Ontology Language. `http://www.w3.org/TR/owl2-overview/`. Último acceso: Mayo 2015.

[68] Baader F, Brand S, Lutz C (2005) Pushing the EL envelope. In: Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI 2005). Morgan-Kaufmann Publishers, pp. 364–369.

[69] Lenat DB, Guha RV (1989) Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1st edition.

[70] Uschold M, King M (1995) Towards a Methodology for Building Ontologies. In: Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95.

[71] Sure Y, Staab S, Studer R (2009) Ontology Engineering Methodology. In: Handbook on Ontologies, Berlin Heidelberg: Springer Berlin Heidelberg.

[72] Gruninger M, Fox MS (1996) The Logic of Enterprise Modelling. In: Bernus P, Nemes L, editors, Modelling and Methodologies for Enterprise Integration, Springer US, IFIP - The International Federation for Information Processing. pp. 140–157.

[73] Lopez M, Gomez-Perez A, Sierra J, Sierra A (1999) Building a chemical ontology using Methontology and the Ontology Design Environment. IEEE Intelligent Systems and their Applications 14: 37–46.

[74] Pinto HS, Tempich C, Staab S (2009) Ontology Engineering and Evolution in a Distributed World Using DILIGENT. In: Staab S, Studer R, editors, Handbook on Ontologies, Springer Berlin Heidelberg, International Handbooks on Information Systems. pp. 153–176.

[75] Simperl E (2009) Reusing ontologies on the Semantic Web: A feasibility study. Data & Knowledge Engineering 68: 905–925.

[76] Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25: 1251–1255.

[77] Maedche A, Staab S (2004) Ontology learning. In: Handbook on ontologies, Springer. pp. 173–190.

[78] Buitelaar P, Cimiano P, Magnini B (2005) Ontology learning from text: methods, evaluation and applications, volume 123. IOS press.

[79] Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y (2007) Pellet: A practical OWL-DL reasoner. Web Semantics: Science, Services and Agents on the World Wide Web 5: 51–53.

[80] Horrocks I (1998) The FaCT system. In: Proceedings of the International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX '98), volume 1397 in Lecture Notes in Artificial Intelligence. Springer, pp. 307–312.

[81] Shearer R, Motik B, Horrocks I (2008) HermiT: A Highly-Efficient OWL Reasoner. In: OWLED. volume 432.

[82] Kiryakov A, Popov B, Terziev I, Manov D, Ognyanoff D (2004) Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web 2: 49–79.

[83] Oren E, Möller K, Scerri S, Handschuh S, Sintek M (2006) What are semantic annotations? Technical Report, Digitial Enterprise Research Institute (DERI), Galway.

[84] Hendler J, Golbeck J (2008) Metcalfe's law, Web 2.0, and the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 6: 14–20.

[85] Passant A, Laublet P (2008) Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In: LDOW.

[86] Pesquita C, Faria D, Falcao AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS computational biology 5: e1000443.

[87] Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19: 17–30.

[88] Resnik P (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11: 95–130.

[89] Hliaoutakis A, Varelas G, Voutsakis E, Petrakis EG, Milios E (2006) Information Retrieval by Semantic Similarity:. International Journal on Semantic Web and Information Systems 2: 55–73.

[90] Rodriguez M, Egenhofer M (2003) Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering 15: 442–456.

[91] W3C. SPARQL Query Language for RDF. `http://www.w3.org/TR/rdf-sparql-query/`. Último acceso: Mayo 2015.

[92] W3C Member Submission. SPIN: SPARQL Inferencing Notation. `http://spinrdf.org/`. Último acceso: Septiembre 2015.

[93] Berners-Lee T. Linked Data - Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`.

[94] Heath T, Bizer C (2011) Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology 1: 1–136.

[95] Bizer C, Heath T, Berners-Lee T (2009) Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5: 1–22.

[96] Linked Data - Connect Distributed Data across the Web. `http://linkeddata.org/`. Último acceso: Mayo 2015.

[97] W3C. The Linking Open Data cloud diagram . `http://lod-cloud.net/`. Último acceso: Mayo 2015.

[98] Berners-Lee T. Relational Databases on the Semantic Web. http://www.w3.org/DesignIssues/RDB-RDF.html. Último acceso: Mayo 2015.

[99] Spanos DE, Stavrou P, Mitrou N (2012) Bringing Relational Databases into the Semantic Web: A Survey. Semantic Web 3: 169–209.

[100] W3C. RDB2RDF Working Group. http://www.w3.org/2001/sw/rdb2rdf/. Último acceso: Mayo 2015.

[101] W3C. A Direct Mapping of Relational Data to RDF. http://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/. Último acceso: Mayo 2015.

[102] W3C. R2RML: RDB to RDF Mapping Language. http://www.w3.org/TR/2012/REC-r2rml-20120927/. Último acceso: Mayo 2015.

[103] Freie Universität. The D2RQ Mapping Language - The D2RQ Platform. http://d2rq.org/d2rq-language. Último acceso: Mayo 2015.

[104] Bizer C, Seaborne A (2004) D2RQ - treating non-RDF databases as virtual RDF graphs. In: Proceedings of the 3rd International Semantic Web Conference (ISWC2004).

[105] Auer S, Dietzold S, Lehmann J, Hellmann S, Aumueller D (2009) Triplify: Light-weight Linked Data Publication from Relational Databases. In: Proceedings of the 18th International Conference on World Wide Web. New York, NY, USA: ACM, WWW '09, pp. 621–630. doi: 10.1145/1526709.1526793.

[106] Erling O, Mikhailov I (2009) RDF Support in the Virtuoso DBMS. In: Pellegrini T, Auer S, Tochtermann K, Schaffert S, editors, Networked Knowledge - Networked Media, Springer Berlin Heidelberg, number 221 in Studies in Computational Intelligence. pp. 7–24.

[107] OpenLink. Virtuoso Open-Source: Mapping Relational Data to RDF in Virtuoso. http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSSQLRDF. Último acceso: Mayo 2015.

[108] Tsinaraki C, Christodoulakis S (2007) XS2owl: A Formal Model and a System for Enabling XML Schema Applications to Interoperate with OWL-DL Domain Knowledge and Semantic Web Tools. In: Thanos C, Borri F, Candela L, editors, Digital Libraries: Research and Development, Springer Berlin Heidelberg, number 4877 in Lecture Notes in Computer Science. pp. 124–136.

[109] Bumans G, Cerans K (2010) RDB2owl: A Practical Approach for Transforming RDB Data into RDF/OWL. In: Proceedings of the 6th International Conference on Semantic Systems. New York, NY, USA: ACM, I-SEMANTICS '10, pp. 25:1–25:3. doi:10.1145/1839707.1839739.

[110] Knoblock CA, Szekely P, Ambite JL, Goel A, Gupta S, et al. (2012) Semi-automatically Mapping Structured Sources into the Semantic Web. In: Simperl E, Cimiano P, Polleres A, Corcho O, Presutti V, editors, The Semantic Web: Research and Applications, Springer Berlin Heidelberg, number 7295 in Lecture Notes in Computer Science. pp. 375–390.

[111] Jupp S, Horridge M, Iannone L, Klein J, Owen S, et al. (2010) Populous: A tool for populating ontology templates. arXiv:10121745 [cs] .

[112] Miñarro Giménez JA (2012) Entorno para la gestión semántica de información biomédica en investigación traslacional. Tesis Doctoral, Universidad de Murcia.

[113] Legaz-García MdC, Martínez-Costa C, Miñarro-Giménez JA, Fernández-Breis JT, Schulz S, et al. (2014) Ontology patterns-based transformation of clinical information. Stud Health Technol Inform 205: 1018–22.

[114] Berlanga R, Romero Moral s, Simitsis A, Nebot V, Pedersen TB, et al. (2011) Semantic web technologies for business intelligence .

[115] CORDIS. MUlti-Industry, Semantic-based Next Generation Business INtelliGence. `http://cordis.europa.eu/project/rcn/79377_en.html`. Último acceso: Julio 2015.

[116] Declerck T, Krieger HU (2006) Translating XBRL Into Description Logic. An Approach Using Protege, Sesame & OWL. In: BIS. pp. 455–467.

[117] Spies M (2010) An ontology modelling perspective on business reporting. Information Systems 35: 404–416.

[118] Debreceny R, Gray GL (2001) The production and use of semantically rich accounting reports on the Internet: XML and XBRL. International Journal of Accounting Information Systems 2: 47–74.

[119] Lytras MD, Angel Sicilia M, Sampson D, Downes S (2005) Semantic networks and social networks. The Learning Organization 12: 411–417.

[120] Cuéllar MP, Delgado M, Pegalajar MC (2011) Improving learning management through semantic web and social networks in e-learning environments. Expert Systems with Applications 38: 4181–4189.

[121] Garton L, Haythornthwaite C, Wellman B (1997) Studying online social networks. Journal of Computer-Mediated Communication 3: 0–0.

[122] Mikroyannidis A (2007) Toward a social semantic web. Computer 40: 113–115.

[123] Hendler J, Berners-Lee T (2010) From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. Artificial Intelligence 174: 156–161.

[124] Stankovic M (2008) Modeling online presence. In: J. Breslin, U. Bojars, A. Passant and S. Fernandez, editeurs, Proceedings of the First Social Data on the Web Workshop, Karlsruhe, Germany. Citeseer, volume 405, p. 58.

[125] Gruber T (2008) Collective knowledge systems: Where the social web meets the semantic web. Web semantics: science, services and agents on the World Wide Web 6: 4–13.

[126] Krotzsch M, Vrandecic D, Volkel M (2006) Semantic mediawiki. In: The Semantic Web-ISWC 2006, Springer. pp. 935–942.

[127] Passant A, Bojars U, Breslin JG, Hastrup T, Stankovic M, et al. (2010) An Overview of SMOB 2: Open, Semantic and Distributed Microblogging. In: ICWSM.

[128] Passant A, Laublet P, Breslin JG, Decker S (2009) SemSLATES: Improving enterprise 2.0 information systems using semantic Web technologies. In: Collaborative Computing: Networking, Applications and Worksharing, 2009. CollaborateCom 2009. 5th International Conference on. IEEE, pp. 1–10.

[129] Sujansky W (2001) Heterogeneous Database Integration in Biomedicine. Journal of Biomedical Informatics 34: 285–298.

[130] Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P (2007) Data integration and genomic medicine. Journal of Biomedical Informatics 40: 5–16.

[131] Calvanese D, Giacomo GD (2005) Data Integration: A Logic-Based Perspective. AI Magazine 26: 59–70.

[132] Anand N (2014) ETL and its impact on Business Intelligence. International Journal of Scientific and Research Publications : 508.

[133] Stein LD (2003) Integrating biological databases. Nature Reviews Genetics 4: 337–345.

[134] Lenzerini M (2002) Data Integration: A Theoretical Perspective. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. New York, NY, USA: ACM, PODS '02, pp. 233–246. doi:10.1145/543613.543644.

[135] Hernandez T, Kambhampati S (2004) Integration of Biological Sources: Current Systems and Challenges Ahead. Sigmod Record 33: 51–60.

[136] Sahoo SS, Bodenreider O, Rutter JL, Skinner KJ, Sheth AP (2008) An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence. Journal of biomedical informatics 41: 752–765.

[137] Fielding RT (2000) Architectural styles and the design of network-based software architectures. Ph.D. thesis, University of California, Irvine.

[138] Cann AJ (2006) Really Simple Syndication. Teaching Bioscience Enhancing Learning Series ?Effective Use of Technology in the Teaching of Bioscience .

[139] Box D, Ehnebuske D, Kakivaya G, Layman A, Mendelsohn N, et al. (2000) Simple object access protocol (SOAP) 1.1.

[140] Mellor SJ, Balcer M, Foreword By-Jacoboson I (2002) Executable UML: A foundation for model-driven architectures. Addison-Wesley Longman Publishing Co., Inc.

[141] Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, et al. (2006) HL7 clinical document architecture, release 2. Journal of the American Medical Informatics Association 13: 30–39.

[142] Feigenbaum L, Herman I, Hongsermeier T, Neumann E, Stephens S (2007) The Semantic Web in Action. Scientific American 297: 64–71.

[143] Wache H, Vögele T, Visser U, Stuckenschmidt H, Schuster G, et al. (2001) Ontology-Based Integration of Information - A Survey of Existing Approaches. In: Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing. pp. 108–117.

[144] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J (2008) Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics 41: 706–716.

[145] Chen H, Yu T, Chen JY (2013) Semantic Web meets Integrative Biology: a survey. Briefings in Bioinformatics 14: 109–125.

[146] Cheung KH, Yip KY, Smith A, deKnikker R, Masiar A, et al. (2005) YeastHub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics 21: i85–i96.

[147] Miñarro-Giménez JA, Egaña Aranguren M, Martínez Béjar R, Fernández-Breis JT, Madrid M (2011) Semantic integration of information about orthologs and diseases: The OGO system. Journal of Biomedical Informatics 44: 1020–1031.

[148] Salka C (1998) Ending the ROLAP/MOLAP debate: usage based aggregation and flexible HOLAP. In: Data Engineering, 1998. Proceedings., 14th International Conference on. IEEE, p. 180.

[149] MARIUS G, AREF M, BILAL H (2009) Real time on-line analytical processing for business intelligence. UPB Sci Bull, Series C 71.

[150] Giacometti A, Marcel P, Negre E, Soulet A (2009) Query recommendations for OLAP discovery driven analysis. In: Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP. ACM, pp. 81–88.

[151] Craig SB, Hannum K (2006) Research update: 360-degree performance assessment. Consulting Psychology Journal: Practice and Research 58: 117.

[152] Kutcher E, Donovan J, Lorenzet SJ (2009) 360-Degree Feedback. Handbook of Improving Performance in the Workplace: Volumes 1-3 : 221–250.

[153] Hazucha JF, Hezlett SA, Schneider RJ (1993) The impact of 360-degree feedback on management skills development. HUMAN RESOURCE MANAGEMENT-ANN ARBOR- 32: 325–325.

[154] Pearce CL (2004) The future of leadership: Combining vertical and shared leadership to transform knowledge work. The Academy of Management Executive 18: 47–57.

[155] Bontis N (1996) There's a price on your head: managing intellectual capital strategically. Business Quarterly 60: 40–78.

[156] Edvinsson L, Malone MS (1997) Intellectual Capital: Realizing Your Company's True Value by Finding Its Hidden Brainpower .

[157] Huseman RC, Goodman JP (1998) Leading with knowledge: The nature of competition in the 21st century. SAGE publications.

[158] Roos J, Edvinsson L, Roos G (1998) Intellectual capital: navigating in the new business landscape. New York University Press.

[159] Bontis N (1999) Managing organisational knowledge by diagnosing intellectual capital: framing and advancing the state of the field. International Journal of technology management 18: 433–462.

[160] Brooking A (1996) Intellectual capital. Cengage Learning EMEA.

[161] Lynn BE (1999) Culture and intellectual capital management: a key factor in successful ICM implementation. International Journal of Technology Management 18: 590–603.

[162] Sveiby KE (1997) The new organizational wealth: Managing & measuring knowledge-based assets. Berrett-Koehler Publishers.

[163] Bontis N, Girardi J (2000) Teaching knowledge management and intellectual capital lessons: an empirical examination of the TANGO simulation. International Journal of Technology Management 20: 545–555.

[164] Strassmann PA (1998) The value of knowledge capital. American Programmer 11: 3–10.

[165] Schobera D, Choquetb R, Depraeterec K, Endersd F, Daumked P, et al. DebugIT: Ontology-mediated layered Data Integration for real-time Antibiotics Resistance Surveillance .

[166] Ruiz-Martínez JM, Miñarro-Giménez JA, Guillén-Cárceles L, Castellanos-Nieves D, Valencia-García R, et al. (2008) Populating ontologies in the eTourism domain. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 03. IEEE Computer Society, pp. 316–319.

[167] Cunningham H, Maynard D, Bontcheva K, Tablan V (2002) A framework and graphical development environment for robust NLP tools and applications. In: ACL. pp. 168–175.

[168] Dittrich Y, Vaucouleur S, Giff S (2009) ERP customization as software engineering: knowledge sharing and cooperation. Software, IEEE 26: 41–47.

[169] Latella D, Majzik I, Massink M (1999) Towards a formal operational semantics of UML statechart diagrams. In: Formal Methods for Open Object-Based Distributed Systems, Springer. pp. 331–347.

[170] Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning 20: 197–243.

[171] Hill T, Westbrook R (1997) SWOT analysis: it's time for a product recall. Long range planning 30: 46–52.

[172] Gangadharan GR, Swami SN (2004) Business intelligence systems: design and implementation strategies. In: Information Technology Interfaces, 2004. 26th International Conference on. IEEE, pp. 139–144.

[173] Sackett DL, Haynes RB, Tugwell P, others (1985) Clinical epidemiology: a basic science for clinical medicine. Little, Brown and Company.

[174] Jackson SE, Joshi A, Erhardt NL (2003) Recent research on team and organizational diversity: SWOT analysis and implications. Journal of management 29: 801–830.

[175] Melton LJ (1996) History of the Rochester epidemiology project. In: Mayo Clinic Proceedings. Elsevier, volume 71, pp. 266–274.

[176] Ferreira JD, Pesquita C, Couto FM, Silva MJ (2012) Bringing epidemiology into the Semantic Web. In: ICBO. Citeseer.

[177] World Health Organization (WHO). International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). `http://apps.who.int/classifications/icd10/browse`. Último acceso: Septiembre 2015.

[178] European Medicines Agency. EudraCT Public Web Report for July 2015. `https://eudract.ema.europa.eu/docs/statistics/`

EudraCT_Statistics_2015/EudraCT_Public_Report_July_2015.pdf. Último acceso: Septiembre 2015.

[179] European Medicines Agency. European Clinical Trials Database (EudraCT). https://eudract.ema.europa.eu/. Último acceso: Septiembre 2015.

[180] Stamou G, Van Ossenbruggen J, Pan JZ, Schreiber G, Smith JR (2006) Multimedia annotations on the semantic web. MultiMedia, IEEE 13: 86–90.

[181] Grupo de Investigación de Enfermedades Respiratorias Infantiles. Proyecto NELA: Nutrition in Early Life and Asthma. http://www.nela.imib.es/. Último acceso: Septiembre 2015.

[182] International Knockout Mouse Consortium (IKMC). EUCOMM: Tools for Functional Annotation of the Mouse Genome. http://www.mousephenotype.org/about-ikmc/eucommtools. Último acceso: Septiembre 2015.

[183] García-Manotas I, Lupiani E, García-Sánchez F, Valencia-García R (2012) Populating knowledge based decision support systems. Integrated and Strategic Advancements in Decision Making Support Systems .

[184] OpenLink. Virtuoso Open-Source. http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/. Último acceso: Mayo 2015.

[185] The owlapi. http://owlapi.sourceforge.net/. Último acceso: Mayo 2015.

[186] Muir CS, Nectoux J (1977) Role of the cancer registry. National Cancer Institute Monograph 47: 3–6.

[187] Jensenl M, Whelan S (1991) . Planning a cancer registry .

[188] Esteban-Gil A, Fernández-Breis JT, Boeker M (2014) Analysis and Visualization of Disease Courses in a Semantic Enabled Cancer Registry. In: SWAT4LS.

[189] Dorn HF (1949) The Use of Statistics in Cancer Control Programs*. American Journal of Public Health and the Nations Health 39: 602–606.

[190] Neitlich HW, Priest SL, O'Sullivan VJ (1983) Development of a regional computerized cancer registry and its impact on medical care. Journal of medical systems 7: 251–255.

[191] Salem ABM, Alfonse M (2007) Building Web-Based Lung Cancer Ontology. In: Proceedings of 1st National Symposium on e-Health and Bioengineering-EHB. pp. 177–182.

[192] Alfonse M, Aref MM, Salem ABM (1818) Ontology-Based Knowledge Representation for Liver Cancer. In: Proceedings of the International eHealth, Telemedicine and Health ICT Forum for Educational, Networking and Business. Luxembourg, GD of Luxembourg, ISSN. volume 9334, pp. 821–825.

[193] Abidi SR (2007) Ontology-based modeling of breast cancer follow-up clinical practice guideline for providing clinical decision support. In: Computer-Based Medical Systems, 2007. CBMS'07. Twentieth IEEE International Symposium on. IEEE, pp. 542–547.

[194] Yates JW, Chalmer B, McKegney FP, others (1980) Evaluation of patients with advanced cancer using the Karnofsky performance status. Cancer 45: 2220–2224.

[195] American Society of Anesthesiologists. ASA Physical Status Classification System. `https://www.asahq.org/clinical/physicalstatus.htm`. Último acceso: Septiembre 2015.

[196] Sobin LH, Gospodarowicz MK, Wittekind C (2011) TNM classification of malignant tumours. John Wiley & Sons.

[197] Cardillo, E, Tamilin, A, Eccher, C, Serafini, L. ICD-10 Ontology. `https://dkm.fbk.eu/technologies/icd-10-ontology`. Último acceso: Septiembre 2015.

[198] World Health Organization (WHO). International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3). `http://www.who.int/classifications/icd/adaptations/oncology/en/`. Último acceso: July 2015.

[199] Centers for Medicare and Medicaid Services (CMS) and the National Center for Health Statistics (NCHS). The 2014 ICD-10-Procedure Coding System (ICD-10-PCS). `http://www.cms.gov/Medicare/Coding/ICD10/2014-ICD-10-PCS.html`. Último acceso: Septiembre 2015.

[200] Quintero E, Castells A, Bujanda L, Cubiella J, Salas D, et al. (2012) Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. New England Journal of Medicine 366: 697–706.

[201] Martinez-Romero M, Vázquez-Naya JM, R Rabunal J, Pita-Fernández S, Macenlle R, et al. (2010) Artificial intelligence techniques for colorectal cancer drug metabolism: Ontologies and complex networks. Current drug metabolism 11: 347–368.

[202] Astler VB, Coller FA (1954) The prognostic significance of direct extension of carcinoma of the colon and rectum. Annals of surgery 139: 846.

[203] Ross A, Rusnak C, Weinerman B, Kuechler P, Hayashi A, et al. (1999) Recurrence and survival after surgical management of rectal cancer. The American journal of surgery 177: 392–395.

[204] Jiang L, Zhang H (2006) Weightily averaged one-dependence estimators. In: Proceedings of the 9th Biennial Pacific Rim International Conference on Artificial Intelligence, PRICAI 2006. volume 4099 of *LNAI*, pp. 970-974.

[205] Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition 30: 1145–1159.

[206] Huang Z, Huang D, Ni S, Peng Z, Sheng W, et al. (2010) Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer. International journal of cancer 127: 118–126.

[207] Khambata-Ford S, Garrett CR, Meropol NJ, Basik M, Harbison CT, et al. (2007) Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. Journal of Clinical Oncology 25: 3230–3237.

[208] Ng EK, Chong WW, Lam EK, Shin VY, Yu J, et al. (2009) Differential expression of microRNAs in plasma of colorectal cancer patients: a potential marker for colorectal cancer screening. Gut .

[209] Weinmayr G, Weiland SK, Björkstén B, Brunekreef B, Büchele G, et al. (2007) Atopic sensitization and the international variation of asthma symptom prevalence in children. American journal of respiratory and critical care medicine 176: 565–574.

[210] Garcia-Marcos L, Canflanca IM, Garrido JB, Varela ALS, Garcia-Hernandez G, et al. (2007) Relationship of asthma and rhinoconjunctivitis with obesity, exercise and Mediterranean diet in Spanish schoolchildren. Thorax 62: 503–508.

[211] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature genetics 25: 25–29.