BizkaiLab

Área prioritaria / Lehenetsitako arloa: AP9/9. LA

Bizkaia Aktiba: coyuntura y competitividad / Bizkaia Aktiba: egoera eta lehiakortasuna

Iniciativa / Ekimena: Abagunea

Acción - proyecto / Ekintza - proiektua: BIDEI. Fuentes de datos enlazadas de Bizkaia / BIDEI. Bizkaiko datu iturri elkarlotuak

Responsable / Arduraduna: Diego López de Ipiña

Equipo / Lan taldea: Diego López de Ipiña, Unai Aguilera, Mikel Emaldi, Jon Lázaro, David Buján, Ana Belén Lago, Ainhoa Alonso, Joseba Abaitua

Código Proyecto / Proiektu Kodea: 5761

Estado del arte en confianza y calidad de fuentes de datos enlazadas

Proyecto BIDEI

Diciembre de 2011

Contenido

Objetivos del documento

Este documento pretende servir como referencia del estado del arte actual en aspectos relacionados con la calidad y confianza en fuentes de datos y particularmente de datos enlazados. Además, de la información propia proporcionada por una fuente de datos enlazados existe también un conjunto de metadatos relacionados que pueden ser utilizados por los usuarios para determinar si esta es adecuada para sus necesidades.

Se resumen, en este documento, aspectos relacionados con los modelos de calidad existentes para el cálculo de la calidad de la información proporcionada por una fuente de datos enlazados, así como un estudio de los diferentes modelos para la representación de la calidad de las fuentes que pueden utilizarse. Se recogen también distintas tecnologías relacionadas con la calidad de las fuentes de datos desde diferentes puntos de vista: representación de la información, entrelazado y ciclo de vida de los datos. Además, se incluye también cuál es el estado del arte en aquellos temas relacionados con la confianza en las fuentes de datos enlazadas: aspectos de seguridad en la información, reputación de los proveedores, procedencia de los datos y licencias de datos.

Modelos de calidad en datos enlazados

Técnicas de evaluación de la representación de la información

En este apartado se analiza el estado del arte actual en el ámbito de la evaluación de la representación de la información. En concreto, se tratan las técnicas para asegurar la integridad de los datos en la Web Semántica y se aborda la problemática de los enlaces rotos.

Integridad de los datos en la web semántica

A pesar de no existir grandes soluciones para gestionar la integridad de los datos publicados en la Web Semántica, es una tarea que ha sido tratada por proyectos de gran envergadura [18], denotando la preocupación de la comunidad respecto a esta temática.

Restricciones de integridad

Una de las técnicas para asegurar la integridad de los datos en la Web Semántica ha sido llevada a cabo bajo la inspiración de las restricciones de integridad aplicadas a las bases de datos. Las bases de datos deben ser en todo momento una representación del mundo real, y no deben contener valores irreales [19]. Algunas alternativas para asegurar esta condición son el *data cleaning* [19], de carácter semi-automático, o las restricciones de integridad [20].

Para implementar estas restricciones de integridad se ha empleado el lenguaje OWL (Web Ontology Language) [21]. A partir del planteamiento teórico descrito en [22], en [23] se describe y se implementa parcialmente un sistema que provee:

- Un sistema para representar las restricciones que una instancia de un recurso semántico debe cumplir.
- Métodos de ayuda a la toma de decisión para evaluar el cumplimiento de las restricciones de integridad.

- Recomendaciones para solucionar los casos en los que las restricciones de integridad sean incumplidas.

En definitiva, utilizando el sistema planteado en [23] se podrían evaluar dichas restricciones y reparar los valores que las incumplan.

Enlaces rotos

La creciente iniciativa "Linked Open Data Cloud" (LOD Cloud) [24] ha colaborado en que muchos proveedores de datos semánticos interconecten sus *datasets* con otros *datasets* ubicados en dicha nube. Como se puede observar en la Tabla 1 en el segundo trimestre de 2011 existían hasta un total de 29 *datasets* con más de un millón de enlaces hacia otros *datasets* de la LOD Cloud.

Enlaces salientes	**Número de *datasets***
Menos de 100	30 (10.17 %)
100 a 1.000	990 (30.51 %)
1.000 a 10.000	558 (19.66 %)
10.000 to 100.000	445 (15.25 %)
100.000 to 1.000.000	443 (14.58 %)
Más de 1.000.000	229 (9.83 %)
	237

Tabla 2. Número de enlaces externos de los *datasets* de LOD Cloud. Extraído de http://www4.wiwiss.fu-berlin.de/lodcloud/state

Muchos de estos enlaces acaban siendo movidos o borrados, ya que muchas veces su construcción no sigue las directrices sugeridas por el W3C [25]. Entre su versión 3.2 y 3.3, DBPedia [26], uno de los *datasets* más populares de la LOD Cloud, movió más de 10.000 enlaces y eliminó más de 30.000 [27], debido a la alta volatilidad de los artículos publicados en Wikipedia.

La temática de los enlaces rotos se ha tratado en ámbitos no necesariamente relacionados con la web de datos, como puede verse en [28] [29] [30] [31] [32].

Según [27] los enlaces rotos pueden clasificarse en dos categorías: "enlaces estructuralmente rotos", y “enlaces semánticamente rotos”. Los primeros se refieren a representaciones de recursos que no pueden recuperarse nunca más, y son fáciles de detectar de manera automática. Los “enlaces semánticamente rotos”, en cambio, se refieren a enlaces en los que la interpretación humana de lo que debe representar ese enlace difiere de la representación real.

De la misma manera, [27] define tres tipos de causas por las cuales se crean enlaces rotos:

- Tipo A: enlaces rotos porque se cambia la localización del **recurso origen**.
- Tipo B: enlaces rotos porque se cambia la localización del **recurso destino**.
- Tipo C: enlaces rotos porque se eliminan el **recurso origen** o el **recurso destino**.

Además, define las posibles soluciones al problema:

- Ignorar el problema: aunque sea difícil considerarla como una solución, es la más frecuente, relegando la problemática a capas superiores de la aplicación consumidora.
- Links embebidos: solucionan los problemas de tipo A, ya que el link hacia el recurso destino se encuentra embebido en el recurso origen. Este es el modelo utilizado en la web de documentos.
- Referencias relativas: las referencias relativas pueden prevenir el surgimiento de enlaces rotos en algunas situaciones (cuando se relocaliza un *dataset* en su totalidad, tipo A) pero no previenen de los enlaces rotos de tipo C.
- Indirección: consiste en introducir una capa de redirección, como PURL [33] o DOI [34].
- Colleciones estáticas: consiste en almacenar los enlaces en una colección estática. En este enfoque no se permite la modificación o eliminación de los recursos.
- Actualizaciones regulares: consiste en la actualización regular de los enlaces, actualizando los enlaces rotos a una nueva ubicación "predecible" de alguna manera.
- Redundancia: consiste en mantener un repositorio de diferentes enlaces que apunten al mismo recurso junto con un servicio que provea al usuario de estos enlaces.
- Enlaces dinámicos: con este método se intentan generar los enlaces en tiempo de ejecución, sin la necesidad de almacenarlos. Al no ser una solución trivial, no es válida para muchos problemas de la vida real.
- Notificación: los recursos informan a los clientes que los enlazan cuando se realiza un cambio.
- Detectar y corregir: consiste en contrastar el recurso hallado en un enlace con un valor esperado. En caso de que difieran, se podrían utilizar técnicas heurísticas para solucionar el enlace roto.
- Búsqueda y edición manual: es la técnica más común dentro de la Web.

La herramienta DsNotify [35] [27] [36] presenta una alternativa a las técnicas mencionadas, combinando las técnicas de "detectar y corregir", "notificación" e "indirección", aplicadas a un conjunto de *datasets* bajo su supervisión. Su funcionamiento se basa en la monitorización

constante de los *datasets* supervisados, intentando establecer correlaciones entre los nuevos *datasets* y los eliminados.

El *framework* de enlazado Silk [37] [38] [39] [40] proporciona un conjunto de herramientas para descubrir y mantener enlaces entre recursos de la Web de Datos. Silk está compuesto por tres componentes:

- El motor de búsqueda de enlaces, basado en la similitud entre recursos o identificadores compartidos.
- Herramienta para evaluar los enlaces generados.
- Protocolo para hacer frente a los posibles cambios producidos en los enlaces entre *datasets*.

Silk utiliza el lenguaje declarativo Silk-LSL (Link Specification Language) con el que el usuario puede establecer el tipo de enlaces RDF que deben ser descubiertos entre los diferentes *datasets* y las condiciones que deben cumplir. Para la problemática de la integridad de los enlaces entre los diferentes *datasets*, Silk utiliza el denominado *Web of Data - Link Maintenance Protocol (WOD-LMP)* [41] [37]. Este protocolo se basa en el patrón publicación-subscripción [42] para propagar los cambios que un *dataset* pueda tener entre sus suscriptores.

Calidad en entrelazado de fuentes de datos

Uno de los principios fundamentales de datos enlazados es que las fuentes de datos contengan enlaces que apunten a información contenida en otras fuentes de datos. El modelo de datos RDF proporciona los mecanismos necesarios para permitir conectar información de diferentes orígenes de datos en un único grafo global que posteriormente podrá ser procesado para obtener la información de las diferentes fuentes de datos.

Además del entrelazado de los propios datos también es necesario que los vocabularios utilizados por las distintas fuentes de datos sean relacionados con aquellos utilizados por otras. La situación ideal en el entrelazado de fuentes de datos es la utilización de vocabularios comunes que se encuentren ampliamente aceptados en el dominio de datos correspondiente. Sin embargo, esto no es siempre posible ya que los proveedores de fuentes de datos enlazadas podrán necesitar la utilización de vocabularios mucho más específicos que los disponibles. En estas situaciones las diferentes fuentes de datos deberán relacionar los vocabularios específicos utilizados entre sí, permitiendo que se pueda navegar por dichos vocabularios y los datos correspondientes.

En el trabajo realizado en [43] se recogen una serie de indicadores, que han sido obtenidos de los criterios clásicos de calidad para las bases de datos y aplicados a las fuentes de datos enlazados. En concreto, los indicadores recogidos en la siguiente tabla pueden aplicarse al ámbito del entrelazado entre fuentes de datos enlazados.

Categoría	Indicadores
Uniformidad en la representación	uso de un formato establecido
	indicar el content-type siempre que sea posible
	uso de vocabularios aceptados
	referencias URIs aceptadas
	número de enlaces entre los vocabularios utilizados
Cantidad de datos	número de triples
	número de enlaces internos
	número de enlaces externos (datos y vocabularios)
	ámbito de la información
	nivel de detalle de la información

Tabla 3. Indicadores para la medida de la calidad en el entrelazado de fuentes de datos

En la tabla anterior, se ha incluido también como indicador de la calidad el número de enlaces existentes entre los vocabularios utilizados y el número de estos. Una fuente de datos que proporcione un mayor número de relaciones entre su vocabulario y otros vocabularios procedentes de otras fuentes proporcionará será apreciada con una mayor calidad por parte de los usuarios. Esto es debido a que les resultará más fácil procesar y descubrir relaciones entre la información disponible.

Aunque los indicadores anteriores han sido identificados en la literatura sobre entrelazado de fuentes de datos no se ha encontrado ninguna referencia al valor o rango de valores para que una fuente de datos sea considerada de mejor calidad que otra. Algunos indicadores son obvios en cuanto a sus implicaciones, por ejemplo, un mayor uso de referencias a fuentes externas o de vocabularios podría implicar una mayor riqueza en la información contenida en la fuente de datos y, por lo tanto, una mayor calidad percibida en el uso por parte de los

usuarios. A pesar de que la inexistencia de una ponderación adecuada de estos indicadores reduce la automatización del proceso de selección de las fuentes, los usuarios pueden verse beneficiados de la existencia de dicha información a la hora de seleccionar las fuentes de datos o realizar comparaciones entre ellas.

La inclusión de metadatos que para la descripción de las fuentes de datos posibilita a los usuarios seleccionar aquellos que mejor se adecuen a las necesidades de la tarea que quieren llevar a cabo. La información sobre aspectos relacionados con el entrelazado entre fuentes de datos podrá ser utilizada para calcular otras medidas de calidad relacionadas. Los indicadores de calidad podrán posteriormente ser utilizados por los usuarios para filtrar las fuentes de datos según su idoneidad. Para ello se podrán utilizar propuestas como las presentadas en [44], donde se propone un framework para el filtrado de fuentes de datos en función de políticas de calidad.

Relaciones entre datasets

Uno de los soluciones más extendidas para la descripción de conjuntos de datos es VoID [45]. Este vocabulario permite a los usuarios conocer información sobre el propio dataset. Entre la información proporcionada se incluye, no solamente el tema sobre el que proporciona información el dataset, sino también sobre otros aspectos estadísticos del mismo y referentes a su estructura.Entre la información proporcionada por el vocabulario VoID se encuentra qué vocabularios son utilizados en el dataset, estadísticas sobre su tamaño y ejemplos de recursos típicos incluidos en el mismo. Además, permite aplicar estas descripciones a diferentes subconjuntos dentro de un mismo dataset que pueden tener diferentes características.

El vocabulario VoID contiene propiedades sobre la estructura interna del dataset como puede ser el número de tripletas, entidades, clases, propiedades, número de sujetos y objetos distintos, que permiten conocer la cantidad de información proporcionada por el mismo y la calidad referente a sus relaciones internas. Además, existen aspectos específicos para la descripción de los aspectos relacionados con el entrelazado de fuentes de datos.Los linksets son colecciones de tripletas RDF que relacionan información (sujetos y objetos) de diferentes datasets. Sobre cada linkset puede conocerse qué dos datasets relaciona y qué tipo de enlaces existen entre ellos.

Además, los dataset y linkset descritos hacen uso de vocabularios que también pueden ser descritos por VoID permitiendo posteriormente el listado de dichos vocabularios por los usuarios.

Relaciones entre vocabularios

Como se ha comentado anteriormente, la calidad de una fuente de datos está relacionada con el uso de vocabularios y la cantidad de relaciones que existan entre estos y otros vocabularios externos. La existencia de conexiones entre vocabularios permite a los usuarios relacionar la información contenida en distintas fuentes de datos y poder obtener, tras la aplicación de distintos procesos, información que no se encontraba explícitamente recogida. El establecimiento de relaciones entre vocabularios permite determinar cuando dos o más fuentes de información hacen referencia a una misma entidad, reduciendo así el número de elementos duplicados y aumentando a su vez la información conocida sobre una determinada

entidad mediante la fusión de datos. La utilización de relaciones explícitas reduce el uso de herramientas de descubrimiento de relaciones entre fuentes de datos (p.e. Silk [40]).

De acuerdo a lo recogido en [18], entregable realizado como parte del proyecto europeo FP7 LOD2 [46], la fusión de datos mejora la calidad de tres formas distintas: Completitud: se puede obtener más información sobre una entidad añadiendo más fuentes de datos. Cada nueva fuente de datos podrá incluir otros objetos, atributos, relaciones, etc. Brevedad: al eliminar información redundante se reduce la extensión de la información necesaria para describir un recurso. Consistencia: requiere resolver los conflictos que puedan surgir entre diferentes fuentes de datos. La utilización de relaciones entre los vocabularios ayuda a determinar y resolver dichos conflictos.

La utilización de las construcciones *owl:equivalentClass*, *rdfs:subClassOf* y *rdfs:subPropertyOf* definidas en OWL y RDF, respectivamente, permiten establecer relaciones entre los diferentes esquemas de datos utilizados. El número de relaciones diferentes entre distintos vocabularios puede ser utilizado como un indicador de la calidad de la información. Al aumentar el número de relaciones explícitas entre vocabularios aumentará también el grado de uniformidad y consistencia de la misma, así como la cantidad de información disponible.

En el mismo entregable del proyecto LOD2, si bien estas soluciones son adecuadas, no permiten expresar relaciones de una mayor granularidad que pueden existir entre los vocabularios. Por ejemplo, aquellas relaciones necesarias para expresar que diferentes recursos hacen referencia a uno mismo o transformaciones de datos entre vocabularios (p.e. conversiones de unidades o tratamiento de strings). Estas relaciones son estudiadas en otros trabajos como [47] y [48] donde se proponen soluciones a dichos problemas.

Ciclo de vida de los datos

A la hora de publicar y hacer uso de datos enlazados es necesario tener en cuenta también aquellos aspectos relacionados con su ciclo de vida. Los datos publicados por un proveedor pasan por diferentes estados desde que el momento en que se publican por primera vez hasta eliminación final por parte del proveedor. Estos cambios pueden incluir correcciones, actualizaciones, modificaciones de vocabularios, migraciones entre datasets, etc.

El conocimiento, por parte de los usuarios, de esta información permite calcular y conocer aspectos relacionados con la calidad y confiabilidad de la misma. Determinar cuándo cierta información fue actualizada por ultima vez o qué cambios se realizaron sobre ella permite que los usuarios tengan una mayor conocimiento sobre la información que están utilizando y puedan decidir qué fuentes utilizar.

Además de todo esto, es recomendable que los datos proporcionen información, en forma de metadatos, sobre su temporalidad. La temporalidad de un conjunto de datos hace referencia a su periodo de validez. Dependiendo del dominio de aplicación y del origen de los datos proporcionados la validez de los mismos podrá ser de varios años o de pocos minutos. En dominios en los que la información no cambia con mucha frecuencia (p.e información geográfica/toponímica) los datos podrán tener una validez que se extenderá a lo largo de grandes periodos de tiempo. Por el contrario, pueden existir dominios en los que la

información de las fuentes de datos sean de naturaleza muy dinámica (p.e fuentes de datos móviles) y, por lo tanto, tendrán periodos de validez mucho más cortos.

Conocer la temporalidad de los datos utilizados permite a los usuarios utilizar aquellas que mejor se adecuen a sus requisitos y necesidades minimizando el uso que pueda hacerse de información que se encuentre desfasada o invalidada.

Evolución y actualización de los datos

En [49] se ha identificado el conjunto de requisitos necesarios para comunicar a los usuarios los cambios que se producen en un conjunto de datos a lo largo de su vida. A partir de estos requisitos los autores definen los siguientes principios que deben ser tenidos en cuenta cuando se trata la evolución y actualización de los datos: descripción, granularidad, comunicación, descubrimiento y escalabilidad. Los 3 primeros están relacionados directamente con la calidad y confiabilidad de los datos publicados. Los aspectos de descubrimiento y escalabilidad hacen referencia, respectivamente, a cómo los usuarios descubren los cambios que se han producido y a la posibilidad de que las solución se pueda a aplicar a los volúmenes de datos y usuarios característicos de la web.

Descripción

Hace referencia al hecho de cómo el usuario descubre que los datos de un dataset han cambiado y cómo obtener dichos cambios. Estos metadatos deben estar recogidos en un formato que pueda ser procesado fácilmente (p.e RDF). Existen diferentes vocabularios para capturar y representar el dinamismo de los datasets.

En [50] se puede encontrar otro ejemplo de ontología para describir cambios en los datasets. Los autores definen un conjunto de eventos que pueden aplicarse sobre un recurso: creación, eliminación, actualización y mover. Los eventos de creación y eliminación ocurren, como su nombre indica, cuando un recurso se hace disponible a través de una URI o cuando deja de estar disponible, respectivamente. La actualización hace referencia a el cambio en la representación del recursos (se añaden o se eliminan datos relacionados con el mismo). El evento de mover hace referencia a la posibilidad de que un recurso que antes era accedido a través de una URI sea posteriormente accedido a través de otra. Consiste en una eliminación seguida de una creación con la condición de que los recursos relacionados sean muy similares. La similitud entre dos recursos, con el fin de determinar si se ha producido un movimiento, se determina por distintos mecanismos: sintácticos, heurísticos, reglas, razonamiento, etc.

The Dataset Dynamics Vocabulary (Dady [51]) es un vocabulario definido mediante una ontología para ser utilizado conjuntamente con VoID [45]. Permite la representación de la frecuencia de las actualizaciones del dataset así como incluir una referencia a la URI de actualización. Linking Open Descriptions of Events (LODE) [52] define un modelo para la representación de eventos. DSNotify Eventset [36] es otro vocabulario para describir cambios en un dataset que permite ordenar por tiempo los eventos de cambio que se han producido. Al igual que el vocabulario anterior se usa en conjunción con VoID.

Otra solución para describir los cambios en un dataset es Talis Changeset [53], resultado del proyecto Talis, que también define un conjunto de términos para describir los cambios que se han producido. Su principal aportación es que introduce el concepto de delta para indicar las

diferencias entre dos datasets. Cada delta incluye las eliminación y las posteriores inserciones necesarias para transformar una versión del dataset en otra.

Triplify [54] es un sistema que proporciona información contenida en bases de dato SQL como datos enlazados. Proporciona un mecanismos para la actualización de colecciones de datos. Estas actualizaciones están descritas mediante un vocabulario propio.

Los usuarios pueden estar interesados no solamente en aquellos cambios que se producen sobre los datos propiamente dichos, sino también sobre las ontologías que los describen. Así, OWL2 Change ontology [55] permite describir qué cambios se han realizado entre diferentes versiones de una ontología OWL2. Otra solución similar es CHAO [56] que representa los cambios que se pueden llevar a cabo sobre una ontología, tanto básicos como más complejos. Esta orientada a su uso con la herramienta de creación de ontologías Protége.

Granularidad

Los cambios producidos sobre un conjunto de datos pueden ocurrir a distintos niveles. Los autores en [49] han definido 4 niveles de granularidad:

- Dataset: indicar que el dataset en su conjunto ha cambiado.
- Recurso: notificar cambios en un recurso concreto.
- Sentencia: cambios a nivel de tripleta dentro del RDF.
- Grafo: para cambios a nivel de grafos que son conjuntos de tripletas RDF.

Comunicación

Existen diferentes estrategias mediante las cuales los proveedores de datos pueden notificar los cambios ocurridos a sus usuarios. Existen enfoques pull (p.e Atom [57], OAI-PMH [58] y RDFSync [59]) en las que los usuarios consultan activamente a los proveedores de datos para obtener los cambios que han sucedido en un periodo de tiempo. Por otro lado, pueden ser los propios proveedores los que notifiquen a los usuarios mediante la aplicación de un enfoque push (p.e PingTheSemanticWeb [60], Semantic PingBack [61], WebHooks [62], SPARUL [63], WOD-LWP [37], sparqlPuSH [64]).

Temporalidad del contenido

La información proporcionada por un proveedor de datos enlazados puede tener una utilidad únicamente dentro de un periodo de validez determinado. En [49] se hace un estudio de la frecuencia con la que la información puede cambiar dentro de un dataset. Obviamente, esto es un aspecto que dependerá del dominio que está siendo modelado por el conjunto de datos enlazados, sin embargo, todos los datos publicados son susceptibles de modificación y cambio durante su vida y, por lo tanto, tener una temporalidad asociada.

Existen ontologías para representar la relación entre conceptos temporales como, por ejemplo, en [65]. Esta ontología introduce los conceptos de Time Entity, Time Instant y Time Interval. Estos conceptos son utilizados describir las sentencias RDF con información temporal que puede indicar la validez de los datos.

Un estudio de la temporalidad del contenido se realiza en [66], donde se propone el concepto de Linked Timelines, que define no solamente la validez temporal de las sentencias RDF individuales contenidas en el dataset, sino también relaciones temporales entre ellas. Extiende

la OWL Time ontology [67] para incluir nuevas propiedades que permitan introducir esta meta información en los datasets.

La temporalidad de los contenidos está relacionada directamente con el ciclo de vida de los datos. Cuando un usuario detecta que los datos utilizados han quedado obsoletos, porque su validez ha expirado, deberá comprobar la existencia de datos actualizados en el proveedor, siempre que no se produzca una notificación automática (solución push) por parte de este.

Modelos de confianza para datos enlazados

Seguridad en la Web Semántica

La Web de Datos habilita que sus usuarios puedan acceder a datos de otros usuarios almacenados en cualquier lugar del mundo, normalmente de forma libre y abierta. En este tipo de comunicaciones es importante tener en cuenta aspectos de seguridad de la información, como la confidencialidad e integridad de la transferencia de los propios datos o el control de acceso a información, especialmente cuando se trabaja sobre contenido sensible. De hecho, en [68] se remarca que una de los mayores desafíos de los datos linkados es la privacidad y seguridad de los datos publicados bajo sus premisas.

Seguridad en el control de acceso a datos semánticos

Los sistemas de control de acceso tradicionales se basan en sistemas de bases de datos centralizados que contienen toda la información de autenticación y autorización necesaria para el control de acceso. Esta arquitectura resulta insuficiente para el ámbito de la web semántica, debido a su naturaleza descentralizada en la que es complicado establecer una autoridad central capaz de controlar el acceso.

Así pues, se han desarrollado modelos de control de acceso aplicables a datos semánticos definidos en torno a medatados a nivel de documento, estableciendo políticas de seguridad por cada documento. Las políticas de seguridad, en el ámbito de la web, se refieren a una serie de reglas que definen los requisitos de seguridad para el acceso y modificación de un documento o dataset [69]. Estas reglas especifican una serie de credenciales a cumplir para para tener acceso al recurso; estas credenciales pueden ser o bien autenticaciones basadas en usuario y contraseña, o bien una serie de características que el agente que quiere acceder al recurso debe cumplir. Se han creado varias ontologías y lenguajes para representar estas políticas en la web semántica (haciendo posible razonar sobre estos metadatos):

- **Web Access Control Vocabulary (ACL):** ACL [70] es un vocabulario para representar listas de control de acceso sobre Web Access Control (WAC) [71], un sistema descentralizado para el establecimiento de permisos de acceso a recursos, organizado en torno a usuarios y grupos identificados por URIs HTTP o WebIDs [72]). En este sistema, los grupos y usuarios, hospedados en cualquier host, están identificados por la URI de una clase de usuarios, sobre la cual se pueden realizar búsquedas para recoger todos los usuarios pertenecientes a dicha clase. De esta forma, se puede filtrar el acceso a un recurso a usuarios tanto internos como externos al host en el que se hospeda el recurso. La ontología especifica la clase *AgentClass*, que representa grupos de usuarios (representados a su vez mediante *foaf:Agent* [73]); además, es capaz de

representar modos de acceso a recursos (*acl:Read, acl:Write, acl:Append, acl:Control*). De esta forma, se pueden formar tripletas en forma de listas de acceso que establecen permisos basados en perfiles FOAF.

- **Privacy Preference Ontology (PPO):** PPO [74] es una ontología ligera que trabaja sobre WAC y que permite crear preferencias de seguridad complejas, basadas en una serie de atributos que un usuario debe satisfacer para acceder a un recurso. Esta ontología define una clase principal denominada *PrivacyPreference*, y una serie de propiedades para especificar el recurso a proteger, los requisitos a cumplir y los diferentes privilegios de acceso. Los requisitos que el solicitante debe cumplir para acceder al recurso pueden ser: pertenecer a una clase de una ontología, tener una cierta propiedad o valor, o relacionarse con otros recursos específicos.

Además, WebID Incubator Group [75], grupo de investigación del W3C, ha desarrollado FOAF+SSL [76] [77], un protocolo de autenticación segura sobre el que se pueden aplicar este tipo de ontologías de control de acceso. Se trata de un sistema de autenticación de una sola conexión que hace uso de la capa SSL [78] construida virtualmente sobre todo navegador web que implementa HTTPS. Está basado en WebID [72], un sistema de identificación para la web basada en URIs [79]. FOAF+SSL hace uso de una arquitectura basada en los estándares de las infraestructuras de clave pública (PKI), usando certificados X.509 que contienen WebIDs de perfiles FOAF.

Haciendo uso de estas ontologías, lenguajes y protocolos de autenticación, se han creado diversos modelos de seguridad para la web semántica, como Policy-enabled Linked Data Server (PeLDS) [80], sistema de almacenamiento de tripletas RDF con seguridad basada en reglas SWRL [81] y protocolos de autenticación FOAF+SSL.

Seguridad en la modificación y consulta de datos semánticos

Los lenguajes de consulta tradicionales como SQL, XPath o LDAP han adolecido de problemas de seguridad basadas en inserciones de usuario no controladas, en las que es posible concatenar consultas programadas con información directamente introducida por el usuario. En [82] se analiza esta misma problemática aplicada a lenguajes de consulta semánticos como SPARQL [83] y SPARUL [63], identificando problemas como la inyección SPARQL para la obtención de datos cuyo acceso pueda ser restringido, o inyección SPARUL para la modificación de datos semánticos sin permiso.

En [84] se analiza esta misma problemática, y se proponen soluciones como el control de la entrada de usuario, aportando una serie de parches para solucionar de forma automática estos problemas en frameworks como Jena o Pellet.

Procedencia de los datos

Uno de los aspectos principales a la hora de evaluar la autenticidad de un dataset, y por tanto, su confiabilidad, es la procedencia de los datos que lo componen, como se estudia en [85]. El W3C Provenance Incubator Group [86] define la procedencia de un recurso de la Web de Datos [68] como “un registro que describe personas, entidades y procesos involucrados en la producción y lanzamiento, o que de otro modo hayan tenido influencia sobre dicho recurso” [87]. Por tanto, se considera la información de procedencia de un dataset como un

registro de su historia, desde su creación, incluyendo información acerca de sus origenes, y sus diferentes accesos y modificaciones.

Tipos de procedencia

En el trabajo de [88] se distinguen dos tipos de procedencia: why-provenance (procedencia del porqué) y where-provenance (procedencia del dónde) La primera representa los orígenes que estuvieron envueltos en la creación de la información, mientras que la segunda representa la localización exacta de dónde se ha extraído dicha información. En [89] se distingue también otro tipo de procedencia, how-provenance (procedencia del cómo), que se refiere a cómo participaron los orígenes en la creación de la información.

Modelos para la representación de la procedencia

La procedencia de los datos ha sido tradicionalmente investigada en otros ámbitos, para los que se han propuesto diferentes modelos, como para bases de datos [90], biología [91], en el ámbito científico [92] [93].

En el ámbito de la procedencia en la web, en el trabajo realizado en [87] se concluyen una serie de recomendaciones relativas a la representación de información acerca de la procedencia de datos, que se resumen en los siguientes puntos:

- Cada recurso del cual se quiere identificar la procedencia debería estar referenciada con una URI, además de identificar al autor de los datos y el proceso llevado a cabo para su creación.
- Sería recomendable que los recursos estuvieran identificados por métodos de identificación de origen, como la licencia o la firma digital.
- La información acerca de la procedencia debería ser accesible para su uso y verificación.
- Sería recomendable la inclusión de un histórico de versiones del recurso, con información temporal acerca de la creación, modificación y acceso de cada versión.

Existen algunos modelos de datos de representación de procedencia en la web que implementan total o parcialmente estas recomendaciones:

- **Open Provenance Model (OPM):** Es un modelo de ámbito general, entendido como un núcleo ampliable e independiente de dominio. OPM define la procedencia como un grafo causal, donde los nodos pueden ser clases de tipo *artifact* (estado de los datos en un momento concreto), *process* (acciones llevadas a cabo para crear, modificar o acceder a los artifacts) o *agents* (controladores de los procesos). Las aristas del grafo definen las relaciones causales entre los nodos, también definidas: *used*, *wasControlledBy*, *wasTriggeredBy*, *generated*, *wasDerivedFrom*. Por otra parte, OPM define también otros conceptos como *accounts* (subgrafos parciales del grafo general, utilizados para representar diferentes estados del mismo grafo en diferentes momentos), o *roles* (permiten describir en profundidad algunas de las relaciones causales definidas). [94]
- **PROV-DM:** Se trata del modelo de datos que está preparando el W3C Provenance Working Group [95] para ser adoptado como oficial, y que está trabajando sobre la base de OPM. Tiene como objetivo proveer de un modelo de datos que sirva para

representar entidades, personas y procesos envueltos en la producción o bien de un dato, o de cualquier objeto. Es un modelo independiente de dominio, pero con una serie de puntos de extensión definidos para poder extenderlo para dominios específicos. El modelo se basa en las relaciones entre tres elementos principales: *Entity* (se trata del objeto o el dato en sí, *Process Execution* (representa una actividad que tiene un efecto sobre las entidades) y *Agent* (representa la persona que lanza y controla las actividades). [96]

Además, existen varios vocabularios y ontologías que sirven para representar la procedencia como metadatos a nivel de documento o dataset:

- Open Provenance Model Vocabulary (OPMV) y Open Provenance Model Ontology (OPMO), que implementan las bases del modelo OPM.
- Provenir Ontology: Usado principalmente para representar la procedencia en el ámbito científico. Define en tres clases base para representar los componentes de la procedencia: *data* (entidad para representar el producto original y sus derivados), *process* (entidad para representar los procesos que afectan a un producto), *agent* (entidad para representar quién ha llevado a cabo un proceso). [97]
- Provenance Vocabulary: Desarrollado para describir la procedencia de los datos enlazados. Está definido como una ontología OWL modular, con un núcleo simple y de ámbito general extendible por módulos. Éste modelo define dos dimensiones de procedencia: la creación de datos, y su acceso. También define conceptos como actores, procesos y artefactos aplicados a ambas dimensiones. [98] [99]
- Dublin Core (DC): DC provee un vocabulario de términos que sirven para representar los metadatos básicos de un recurso. En términos de procedencia, contiene términos básicos para referirse a quién creo un recurso, cuándo, basándose en qué, etc. [100]

Cabe comentar que, a expensas de la finalización de PROV-DM (que está previsto que derive en una ontología oficial de W3C), ninguno de los modelos de datos ni ontologías específicos para la representación de la procedencia ha sido adoptada como estándar ni recomendación oficial, pese a que OPM es el modelo normalmente utilizado como referencia. Así pues, la ontología más extendida para representación de la procedencia es DC, pese a no ser una ontología específica para ello y no alcanzar la completitud de OPM o PROV-DM (no contempla relaciones más allá de la creación ni ahonda en las diferentes entidades que pueden existir en relaciones de procedencia).

Procedencia como aspecto para la calidad y confianza

El estudio de la procedencia es la base para establecer criterios de **verificabilidad** a la hora de calcular la calidad y confianza. La verificabilidad se refiere a dar la posibilidad de comprobar la exactitud y grado de corrección de una información, para lo cual es necesario conocer los orígenes de los datos a verificar.

Para ello, en el trabajo realizado en [43] se especifican varios indicadores para el cálculo de la verificabilidad en la web semántica:

- Inclusión de información básica de procedencia.
- Uso de vocabularios dedicados para la representación de la procedencia.

- Uso de firmas digitales. [101] [102]

El primero se refiere a al menos la inclusión de datos básicos como autor, editor y colaboradores de la información, así como las referencias utilizadas para su creación. Para ello, es posible utilizar ontologías básicas como DC. El segundo se refiere a la utilización de modelos de representación de la procedencia más avanzados, como OPMV, Provenance Vocabulary o PROV-DM. Estos modelos y ontologías han sido explicadas en el apartado 4.2.2.

Por otro lado, en [103] se presenta un modelo de evaluación de la calidad para la procedencia en la web basado en la **temporalidad**, que se refiere a lo actualizada que está una información teniendo en cuenta las acciones llevadas a cabo hasta llegar a la última versión [104]. Este modelo trabaja sobre un modelo de grafos de procedencia que representan las relaciones entre *actores* (creadores, editores...) y *artefactos* (la propia información), así como las acciones llevadas a cabo para crear, modificar o acceder a dichos artefactos. El modelo establece una serie de cálculos para obtener el ratio de calidad de cada relación teniendo en cuenta su temporalidad.

Licencias de publicación de los datos

La problemática acerca del licenciamiento de los datos es una cuestión que a menudo crea controversia, ya que al contrario que en el caso de las patentes y el copyright, hasta hace unos años no existía un licenciamiento específico a la hora de publicar datos. En el año 2004, en [106] ya se debatía sobre de las cuestiones legales acerca de la publicación de bases de datos en EEUU. Al igual que en el mundo del software, a la hora de publicar datos existe una gran polarización entre el creciente movimiento del Open Access [107] (cada vez más extendido entre las instituciones académicas) y la posición tradicional.

En este apartado se describen las soluciones existentes para licenciar los datos abiertos, de la misma manera que se describen algunas de las diferentes licencias existentes en campos de alguna manera relacionados, como el software libre o las obras multimedia.

Licencias en el campo del software libre

Entre los movimientos Open Data y Open Source (software libre) existen numerosas similitudes en torno a la filosofía subyacente, por lo que el conocimiento acerca de las licencias de software libre puede servir para la compresión de la filosofía del movimiento Open Data. El software libre es aquel software cuyo código está disponible públicamente, aunque solamente esto no hace que el software sea libre. Para que el software pueda ser clasificado como libre, debe permitir las cuatro libertades del software libre [108]:

1. La libertad para utilizar el software con cualquier propósito.
2. La libertad para analizar su funcionamiento y modificarlo según las necesidades de cada uno. Para esto es necesaria la difusión del código fuente.
3. La libertad de distribuir copias del software.
4. La liberdad de distribuir copias del software modificadas por uno mismo.

Para regular la publicación de software libre existen diferentes licencias. Las cuatro licencias más utilizadas según Black Duck [109] son: GPL (GNU General Public License) [110], MIT License [111], Artistic License [112] y LGPL [113].

Licencias en el campo de las obras multimedia

A la hora de licenciar las obras multimedia, las licencias Creative Commons [114] son las más populares. Existen varias modalidades de Creative Commons:

- **Reconocimiento:** la obra puede ser distribuida de cualquier manera y bajo cualquier condición, siempre y cuando se reconozca la autoría de la obra original.
- **Reconocimiento-CompartirIgual:** las obras derivadas a partir de la obra original deben distribuirse bajo la misma licencia que ésta.
- **Reconocimiento-SinObraDerivada:** esta licencia permite la distribución de la obra siempre y cuando no se modifique la misma.
- **Reconocimiento-NoComercial:** esta licencia permite la distribución de la obra siempre y cuando no se haga uso comercial de ella.
- **Reconocimiento-No Comercial-CompartirIgual:** es la combinación resultante de las licencias **Reconocimiento-NoComercial** y **Reconocimiento-CompartirIgual**.
- **Reconocimiento-NoComercial-SinObraDerivada:** es la combinación resultante de las licencias **Reconocimiento-SinObraDerivada** y **Reconocimiento-NoComercial**.

A pesar de que estas licencias Creative Commons han despertado un gran interés dentro de la comunidad Open Data, hay que remarcar que son licencias para licenciar **obras** y no **datos** [115].

Open Data Commons

Open Data Commons [116] son una serie de herramientas legales para la publicación de datos abiertos, siendo las más remarcables las licencias para datos abiertos. Estas licencias se basan en el borrador de la llamada Talis Community License [53], licencia creada por la empresa Talis a partir de sus necesidades como editor de datos abiertos.

La Talis Community License se basa en las leyes de la Comunidad Europea y reconoce al consumidor de datos, entre otros, el derecho a:

- Usar, reproducir, consumir, mostrar, transmitir, sublicenciar, vender y distribuir copias de los datos iniciales, sin necesidad de haber contribuido en ellos.
- Modificar una copia de los datos originales con el objetivo de formar una obra derivada.
- Usar, reproducir, consumir, mostrar, transmitir, sublicenciar, vender y distribuir copias de la obra derivada siempre y cuando se haga bajo los términos de esta licencia e indicando en todo momento la procedencia de los datos originales.
- Dividir la obra en porciones que estén sujetas a diferentes licencias.

Por otra parte, el editor de los datos tiene la obligación de adjuntar una copia de la licencia junto a sus datos.

A partir de este borrador se redactaron tres licencias: Public Domain Dedication and License (PDDL) [117], Attribution License (ODC-By) [118] and Open Database License (ODC-ODbL) [119].

Public Domain Dedication and License (PDDL)

Esta licencia concede al consumidor la libertad de:

- Copiar, distribuir y utilizar la base de datos.

- Crear obras derivadas a partir de los datos originales.
- Modificar y transformar la base de datos.

Attribution License (ODC-By)

Esta licencia concede al consumidor las mismas libertades que la licencia PDDL añadiendo la restricción de que debe reconocer la autoría de los datos originales dentro de los distintos usos u obras derivadas.

Open Database License (ODC-ODbL)

Esta licencia concede al consumidor todas las libertades y restricciones de ODC-By añadiendo las siguientes restricciones:

- El consumidor debe distribuir las obras derivadas bajo la misma licencia que los datos originales.
- La base de datos debe ser abierta, aunque paralelamente, puede publicarse bajo técnicas que restrinjan su uso (como el DRM [120]).

Apertura de datos en España

El 24 de octubre de 2011 se aprobó el Real Decreto 1495/2011 por el que se desarrolla la Ley 37/2007, de 16 de noviembre, sobre la reutilización de la información del sector público, para el ámbito del sector público estatal. Este Decreto establece las siguientes disposiciones [121]:

- Se permite la reutilización de los documentos para fines comerciales y no comerciales, siempre que dicho uso no constituya una actividad administrativa pública.
- La reutilización autorizada incluye, entre otras, la copia, difusión, modificación, adaptación, extracción, reordenación y combinación de la información.
- Se entiende como documento toda información cualquiera que sea su soporte material o electrónico así como su forma de expresión gráfica, sonora o en imagen utilizada, incluyendo, en consecuencia, también los datos en sus niveles más desagregados o “en bruto”.
- Esta autorización conlleva, la cesión gratuita y no exclusiva de los derechos de propiedad intelectual, correspondientes a tales documentos para el desarrollo de las actividades englobadas dentro de la reutilización autorizada.
- Queda prohibida la desnaturalización de la información reutilizada.
- Debe citarse la fuente de los documentos reutilizados, mencionando la fecha de la última actualización de los documentos originales.
- No se podrá indicar, sugerir o insinuar que el organismo titular de los datos participa o patrocina la reutilización realizada con los datos.
- Los usuarios que realizan la reutilización son los que tendrán que responder frente a terceros por daños que pudieran derivarse de ella.
- Los órganos públicos no serán responsables del uso realizado por la reutilización de sus datos.
- Los órganos públicos no garantizan la continuidad en la puesta a disposición de los documentos reutilizables, ni en contenido ni en forma, ni asume responsabilidades por cualquier error u omisión contenido en ellos.

Conclusiones

En este entregable se ha realizado un estudio de los diferentes modelos de evaluación de la calidad que existen para el cálculo de la confianza en las fuentes de datos enlazados. Estos modelos podrán ser utilizados para obtener indicadores que faciliten a los usuarios la selección de las fuentes de datos más adecuadas. El entregable es un estudio preliminar que recoge las distintas estrategias que se han encontrado para la inclusión de modelos de confianza y calidad en los datos enlazados, no se recogen todas las técnicas existentes sino aquellas que se ha considerado que son las más utilizadas en la actualidad.

La mayoría de las tecnologías recogidas aquí para la anotación de la confianza y calidad de las fuentes de datos enlazadas no han sido estandarizadas y, por lo tanto, no han sido acogidas por los proveedores de información. Tras la realización del estado del arte, se recomienda que los datos enlazados publicados desde la plataforma cuenten con meta-información que permita a los usuarios de los mismos conocer cuál es el nivel de calidad de los datos utilizados y determinar si son adecuados para sus fines.

La confianza de las fuentes de datos es un aspecto que se encuentra directamente relacionado con la calidad percibida de las mismas por parte de los usuarios. Se han recogido aspectos relacionados con la seguridad de las fuentes de datos y con el licenciamiento de las mismas, aspectos, ambos, que tienen un impacto directo en la confianza percibida por los usuarios con respecto a las fuentes de datos enlazadas. El ciclo de vida de los datos enlazados permite que los usuarios de conozcan que cambios han sufrido los datos consumidos desde la última vez que fueron consultados. En dominios en los que exista una gran dinamicidad, por ejemplo aplicaciones móviles, el conocimiento de la temporalidad de los datos es fundamental. Los datos publicados por los proveedores podrán tener un tiempo de vida útil que dependerá de su naturaleza. Otro aspecto fundamental para determinar la confianza y calidad de una fuente de datos es conocer su procedencia, así como los diferentes transformaciones que los datos han sufrido desde su proveedor original hasta que son consumidos por los usuarios finales.

La introducción de conjuntos de metadatos que permitan a los usuarios conocer los distintos aspectos recogidos en este documento mejorará el uso de dichas fuentes de datos. En el diseño e implementación de las herramientas de cálculo de confianza y calidad, así como en el diseño y desarrollo de la plataforma deberá tenerse en cuenta los aspectos aquí recogidos con el objeto de que los usuarios puedan seleccionar aquellas fuentes de datos con una la calidad y confianza que mejor se ajuste a sus necesidad y, por otro lado, que de los datos publicados por la propia plataforma. La falta de estándares claros que cubran todas las necesidad en cuanto a calidad y confianza de los datos podrá requerir durante el proyecto la elaboración de soluciones propias que integren los aspectos más relevantes que han sido identificados.

Referencias

1. ↑ 18,0 18,1 LOD2 - Creating Knowledge out of Interlinked Data. FP7-ICT-2009-5, 2011, “Deliverable 4.3.1 Survey Report State of the art on Mapping, Quality Assessment and Data Fusion (http://lod2.eu/Deliverable/D4.3.1.html),” 2011.

2. ↑ [19.0] [19.1] A. Fuxman, E. Fazli, and R. J. Miller, “ConQuer: efficient management of inconsistent databases,” in Proceedings of the 2005 ACM SIGMOD international conference on Management of data, New York, NY, USA, 2005, pp. 155–166.
3. ↑ M. Arenas, L. Bertossi, and J. Chomicki, “Consistent query answers in inconsistent databases,” in Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, New York, NY, USA, 1999, pp. 68–79.
4. ↑ D. L. McGuinness and F. Van Harmelen, “OWL web ontology language overview,” W3C recommendation, vol. 10, pp. 2004-03, 2004.
5. ↑ B. Motik, I. Horrocks, and U. Sattler, “Adding integrity constraints to OWL,” in Proceedings of the Workshop OWLED, 2007.
6. ↑ [23.0] [23.1] J. Tao, “Adding integrity constraints to the semantic web for instance data evaluation,” The Semantic Web–ISWC 2010, pp. 330-337, 2010.
7. ↑ “The Linking Open Data cloud diagram.” [Online]. Available: http://richard.cyganiak.de/2007/10/lod/. [Accessed: 30-Nov-2011].
8. ↑ “Hypertext Style: Cool URIs don’t change.” [Online]. Available: http://www.w3.org/Provider/Style/URI.html.en. [Accessed: 30-Nov-2011].
9. ↑ S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives, “DBpedia: A Nucleus for a Web of Open Data,” IN 6TH INT’L SEMANTIC WEB CONFERENCE, BUSAN, KOREA, p. 11--15, 2007.
10. ↑ [27.0] [27.1] [27.2] [27.3] B. Haslhofer and N. Popitsch, “DSNotify: Detecting and Fixing Broken Links in Linked Data Sets,” Aug. 2009.
11. ↑ H. Ashman, “Electronic document addressing: dealing with change,” ACM Comput. Surv., vol. 32, no. 3, pp. 201–212, Sep. 2000.
12. ↑ I. Caughey, D. B. Ingham, S. J. Caughey, and M. C. Little, “Fixing the ‘Broken-link’ Problem: The W3Objects Approach,” 1996.
13. ↑ A. Morishima, A. Nakamizo, T. Iida, S. Sugimoto, and H. Kitagawa, “Bringing your dead links back to life,” 2009, p. 15.
14. ↑ T. A. Phelps and R. Wilensky, “Robust Hyperlinks Cost Just Five Words Each,” Computer Science Division, University of California at Berkeley, Berkeley, California, Jan. 2000.
15. ↑ J. Martinez-Romo and L. Araujo, “Recommendation System for Automatic Recovery of Broken Web Links,” in Proceedings of the 11th Ibero-American conference on AI: Advances in Artificial Intelligence, Berlin, Heidelberg, 2008, pp. 302–311.
16. ↑ “PURL Home Page.” [Online]. Available: http://purl.oclc.org/docs/index.html. [Accessed: 30-Nov-2011].
17. ↑ “The Digital Object Identifier System.” [Online]. Available: http://www.doi.org/. [Accessed: 30-Nov-2011].
18. ↑ “DSNotify.” [Online]. Available: http://dsnotify.org/. [Accessed: 30-Nov-2011].
19. ↑ [36.0] [36.1] N. Popitsch and B. Haslhofer, “DSNotify – A solution for event detection and link maintenance in dynamic datasets,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 9, no. 3, pp. 266-283, Sep. 2011.
20. ↑ [37.0] [37.1] [37.2] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, “Discovering and Maintaining Links on the Web of Data,” in The Semantic Web - ISWC 2009, vol. 5823, Springer Berlin Heidelberg, 2009, pp. 650-665.
21. ↑ C. Bizer, J. Volz, G. Kobilarov, and M. Gaedke, “Silk - A Link Discovery Framework for the Web of Data,” Apr. 2009.
22. ↑ A. Jentzsch, R. Isele, and C. Bizer, “Silk–Generating RDF Links while publishing or consuming Linked Data,” Poster at the International, 2010.
23. ↑ [40.0] [40.1] R. Isele, A. Jentzsch, and C. Bizer, “Silk Server - Adding missing Links while consuming Linked Data,” Architecture, 2010.

24. ↑ "Link Synchronization Protocol - Specification." [Online]. Available: http://www4.wiwiss.fu-berlin.de/bizer/silk/wodlmp/. [Accessed: 30-Nov-2011].
25. ↑ "Publish–subscribe pattern - Wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/Publish%E2%80%93subscribe_pattern. [Accessed: 30-Nov-2011].
26. ↑ 43,0 43,1 Annika Flemming and Olaf Hartig, "Quality Criteria for Linked Data sources - trdf," 2010. [Online]. Available: http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources. [Accessed: 28-Nov-2011].
27. ↑ C. Bizer and R. Cyganiak, "Quality-driven information filtering using the WIQA policy framework," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, no. 1, pp. 1-10, 2009.
28. ↑ 45,0 45,1 Keith Alexander (Talis), Richard Cyganiak (DERI, National University of Ireland, Galway), Michael Hausenblas (DERI, National University of Ireland, Galway), and Jun Zhao (University of Oxford), "Describing Linked Datasets with the VoID Vocabulary." [Online]. Available: http://www.w3.org/TR/2011/NOTE-void-20110303/. [Accessed: 28-Nov-2011].
29. ↑ LOD2 – Creating Knowledge out of Interlinked Data." [Online]. Available: http://lod2.eu/Project/WIQA.html. [Accessed: 24-Nov-2011].
30. ↑ J. Euzenat, F. Scharffe, and A. Zimmerman, "Expressive Alignment Language and Implementation," Knowledge Web Network of Excellence (EU-IST-2004-507482), Project Deliverable D2.2.10, 2007.
31. ↑ B. Haslhofer, "A Web-based Mapping Technique for Establishing Metadata Interoperability," PhD, 2008.
32. ↑ 49,0 49,1 49,2 J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres, and S. Decker, "Towards dataset dynamics: Change frequency of linked open data sources," 2010.
33. ↑ N. Popitsch, B. Haslhofer, and E. M. Roochi, "An evaluation approach for dynamics-aware applications using linked data," in Database and Expert Systems Applications (DEXA), 2010 Workshop on, 2010, pp. 188-192.
34. ↑ "Dataset Dynamics (dady) vocabulary | DERI Vocabularies." [Online]. Available: http://vocab.deri.ie/dady. [Accessed: 15-Dec-2011].
35. ↑ R. Shaw, R. Troncy, and L. Hardman, "Lode: Linking open descriptions of events," The Semantic Web, pp. 153-167, 2009.
36. ↑ 53,0 53,1 "Talis changeset vocabulary." [Online]. Available: http://vocab.org/changeset/schema.html. [Accessed: 15-Dec-2011].
37. ↑ S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller, "Triplify: light-weight linked data publication from relational databases," in Proceedings of the 18th international conference on World wide web, 2009, pp. 621-630.
38. ↑ "Ontology Metadata Vocabulary - OWL 2 Change ontology." [Online]. Available: http://omv.ontoware.org/OWLChanges. [Accessed: 15-Dec-2011].
39. ↑ N. Noy, A. Chugh, W. Liu, and M. Musen, "A framework for ontology evolution in collaborative environments," The Semantic Web-ISWC 2006, pp. 544-558, 2006.
40. ↑ "The atom syndication format. rfc4287," 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4287. [Accessed: 15-Dec-2011].
41. ↑ "Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0." [Online]. Available: http://www.openarchives.org/OAI/openarchivesprotocol.html. [Accessed: 15-Dec-2011].
42. ↑ G. Tummarello, C. Morbidoni, R. Bachmann-Gmür, and O. Erling, "RDFSync: efficient remote synchronization of RDF models," in Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, 2007, pp. 537-551.

43. ↑ "Ping the Semantic Web.com - Share your RDF data with the World!" [Online]. Available: http://pingthesemanticweb.com/. [Accessed: 15-Dec-2011].
44. ↑ S. Tramp, P. Frischmuth, T. Ermilov, and S. Auer, "Weaving a social data web with semantic pingback," Knowledge Engineering and Management by the Masses, pp. 135-149, 2010.
45. ↑ "Web Hooks - HTTP callback." [Online]. Available: http://wiki.webhooks.org. [Accessed: 15-Dec-2011].
46. ↑ 63,0 63,1 "SPARQL 1.1 Update - W3C Working Draft 12 May 2011." [Online]. Available: http://www.w3.org/TR/sparql11-update/. [Accessed: 15-Dec-2011].
47. ↑ A. Passant and P. N. Mendes, "sparqlPuSH: Proactive notification of data updates in RDF stores using PubSubHubbub," in Scripting for the Semantic Web Workshop (SFSW2010) at ESWC2010, 2010.
48. ↑ J. R. Hobbs and F. Pan, "An ontology of time for the semantic web," ACM Transactions on Asian Language Information Processing (TALIP), vol. 3, no. 1, pp. 66-85, 2004.
49. ↑ G. Correndo, M. Salvadores, I. Millard, and N. Shadbolt, "Linked Timelines: Time Representation and Management in Linked Data," in First International Workshop on Consuming Linked Data (COLD 2010), Shanghai, China, 2010.
50. ↑ G. Tonti, J. Bradshaw, R. Jeffers, R. Montanari, N. Suri, and A. Uszok, "Semantic Web languages for policy representation and reasoning: A comparison of KAoS, Rei, and Ponder," The Semantic Web-ISWC 2003, pp. 419-437, 2003.
51. ↑ 68,0 68,1 C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," International Journal on Semantic Web and Information Systems, vol. 4, no. 2, pp. 1-22, 2009.
52. ↑ L. Kagal, T. Finin, and A. Joshi, "A policy based approach to security for the semantic web," The Semantic Web-ISWC 2003, pp. 402-418, 2003.
53. ↑ "WebAccessControl/Vocabulary - W3C Wiki." [Online]. Available: http://www.w3.org/wiki/WebAccessControl/Vocabulary. [Accessed: 19-Dec-2011].
54. ↑ "WebAccessControl - W3C Wiki." [Online]. Available: http://www.w3.org/wiki/WebAccessControl. [Accessed: 19-Dec-2011].
55. ↑ 72,0 72,1 M. Sporny, S. Corlosquet, T. Inkster, H. Story, B. Harbulot, and R. Bachmann-Gmür, WebID 1.0: Web identification and Discovery. Unofficial draft (August 2010). 2010.
56. ↑ D. Brickley and L. Miller, "FOAF vocabulary specification," 2005.
57. ↑ O. Sacco and A. Passant, "A privacy preference ontology (PPO) for linked data," in Proceedings of the Linked Data on the Web Workshop, LDOW, 2011.
58. ↑ "W3C WebID Incubator Group." [Online]. Available: http://www.w3.org/2005/Incubator/webid/. [Accessed: 19-Dec-2011].
59. ↑ H. Story, B. Harbulot, I. Jacobi, and M. Jones, "FOAF+ SSL: RESTful authentication for the social web," in Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web (SPOT2009), 2009.
60. ↑ J. Hollenbach, J. Presbrey, and T. Berners-Lee, "Using rdf metadata to enable access control on the social semantic web," in Workshop on Collaborative Construction, Management and Linking of Structured Knowledge, 2009.
61. ↑ E. Rescorla, SSL and TLS: designing and building secure systems, vol. 1. Addison-Wesley Reading, Massachusetts, 2001.
62. ↑ L. Masinter, T. Berners-Lee, and R. T. Fielding, "Uniform resource identifier (URI): Generic syntax," 2005.
63. ↑ H. Mühleisen, M. Kost, and J. C. Freytag, "SWRL-based Access Policies for Linked Data," in Proceedings of the Second Workshop on Trust and Privacy on the Social and Semantic Web (SPOT2010), 2010, vol. 576.

64. ↑ I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, “SWRL: A semantic web rule language combining OWL and RuleML,” W3C Member submission, vol. 21, p. 79, 2004.
65. ↑ P. Orduña, A. Almeida, U. Aguilera, X. Laiseca, D. López-de-Ipiña, and A. G. Goiri, “Identifying Security Issues in the Semantic Web: Injection attacks in the Semantic Query Languages,” 2010.
66. ↑ E. Prud’Hommeaux and A. Seaborne, “SPARQL query language for RDF,” W3C working draft, vol. 4, no. January, 2008.
67. ↑ P. Orduña, A. Almeida, U. Aguilera, X. Laiseca, and A. Gómez-Goiri, “SPARQL/RDQL/SPARUL Injection.” [Online]. Available: http://www.morelab.deusto.es/code_injection/index.html. [Accessed: 15-Dec-2011].
68. ↑ Y. Gil and D. Artz, “Towards content trust of web resources,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5, no. 4, pp. 227-239, 2007.
69. ↑ “W3C Provenance Incubator Group Wiki - XG Provenance Wiki.” [Online]. Available: http://www.w3.org/2005/Incubator/prov/. [Accessed: 14-Dec-2011].
70. ↑ 87.0 87.1 “Provenance XG Final Report.” [Online]. Available: http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/#Proposed_Charter_for_a_Provenance_Interchange_Working_Group. [Accessed: 29-Nov-2011].
71. ↑ P. Buneman, S. Khanna, and T. Wang-Chiew, “Why and where: A characterization of data provenance,” Database Theory—ICDT 2001, pp. 316-330, 2001.
72. ↑ T. J. Green, G. Karvounarakis, and V. Tannen, “Provenance semirings,” in Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2007, pp. 31-40.
73. ↑ W. C. Tan, “Provenance in databases: Past, current, and future,” IEEE Data Eng. Bull, vol. 30, no. 4, pp. 3-12, 2007.
74. ↑ J. Zhao, G. Klyne, and D. Shotton, “Provenance and linked data in biological data webs,” in Proceedings of the 17th International World Wide Web Conference WWW2008 (Workshop: Linked Data on the Web LDOW2008), 2008, p. 2008.
75. ↑ R. Bose and J. Frew, “Lineage retrieval for scientific data processing: a survey,” ACM Computing Surveys (CSUR), vol. 37, no. 1, pp. 1-28, 2005.
76. ↑ Y. L. Simmhan, B. Plale, and D. Gannon, “A survey of data provenance in e-science,” ACM Sigmod Record, vol. 34, no. 3, pp. 31-36, 2005.
77. ↑ L. Moreau et al., “The open provenance model core specification (v1. 1),” Future Generation Computer Systems, vol. 27, no. 6, pp. 743-756, 2011.
78. ↑ “Provenance WG Wiki.” [Online]. Available: http://www.w3.org/2011/prov/wiki/Main_Page. [Accessed: 13-Dec-2011].
79. ↑ L. Moreau et al., “The PROV Data Model and Abstract Syntax Notation,” World Wide Web Consortium, W3C First Public Working Draft, Oct. 2011.
80. ↑ S. S. Sahoo and A. Sheth, “Provenir ontology: Towards a Framework for eScience Provenance Management,” in Microsoft eScience Workshop. Pittsburgh, USA, 2009.
81. ↑ O. Hartig, “Provenance information in the web of data,” in Proceedings of the Linked Data on the Web LDOW Workshop at WWW, 2009.
82. ↑ O. Hartig and J. Zhao, “Publishing and consuming provenance metadata on the web of linked data,” Provenance and Annotation of Data and Processes, pp. 78-90, 2010.
83. ↑ S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, “Dublin core metadata for resource discovery,” Internet Engineering Task Force RFC, vol. 2413, 1998.
84. ↑ J. Carroll, “Signing RDF graphs,” The Semantic Web-ISWC 2003, pp. 369-384, 2003.
85. ↑ G. Tummarello, C. Morbidoni, P. Puliti, and F. Piazza, “Signing individual fragments of an RDF graph,” in Special interest tracks and posters of the 14th international conference on World Wide Web, 2005, pp. 1020-1021.

86. ↑ O. Hartig and J. Zhao, “Using web data provenance for quality assessment,” in In: Proc. of the Workshop on Semantic Web and Provenance Management at ISWC, 2009.
87. ↑ S. F. . Sampaio, C. Dong, and P. R. . Sampaio, “Incorporating the timeliness quality dimension in internet query systems,” in of: Proceedings of the International Workshop on Web Information Systems Engineering, WISE, 2005, pp. 53-62.
88. ↑ Francisco Javier Nieto, Enric Sàrries, Sandra Jiménez and Davide Cerri, “Trust, Security, Dependability Platform Final Specifications”, 2010.
89. ↑ "James Boyle: A natural experiment - FT.com." [Online]. Available: http://www.ft.com/cms/s/2/4cd4941e-3cab-11d9-bb7b-00000e2511c8.html#axzz1gVBq9tvv. [Accessed: 14-Dec-2011].
90. ↑ "Open access - Wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/Open_access. [Accessed: 14-Dec-2011].
91. ↑ "What is free software? - GNU Project - Free Software Foundation (FSF)." [Online]. Available: http://www.gnu.org/philosophy/free-sw.html. [Accessed: 14-Dec-2011].
92. ↑ "Black Duck Open Source Resource Center | Black Duck Software." [Online]. Available: http://www.blackducksoftware.com/oss/licenses/. [Accessed: 15-Dec-2011].
93. ↑ "The GNU General Public License v3.0 - GNU Project - Free Software Foundation (FSF)." [Online]. Available: http://www.gnu.org/copyleft/gpl.html. [Accessed: 14-Dec-2011].
94. ↑ "Open Source Initiative OSI - The MIT License (MIT):Licensing | Open Source Initiative." [Online]. Available: http://www.opensource.org/licenses/MIT. [Accessed: 14-Dec-2011].
95. ↑ "Artistic License 2.0 | Open Source Initiative." [Online]. Available: http://www.opensource.org/licenses/artistic-license-2.0. [Accessed: 14-Dec-2011].
96. ↑ "GNU Lesser General Public License v3.0 - GNU Project - Free Software Foundation (FSF)." [Online]. Available: http://www.gnu.org/licenses/lgpl.html. [Accessed: 14-Dec-2011].
97. ↑ "Creative Commons." [Online]. Available: http://creativecommons.org/. [Accessed: 15-Dec-2011].
98. ↑ P. Miller, R. Styles, and T. Heath, "Open data commons, a license for open data," in Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008), 2008.
99. ↑ "Open Data Commons | Legal tools for Open Data." [Online]. Available: http://opendatacommons.org/. [Accessed: 15-Dec-2011].
100. ↑ "ODC Public Domain Dedication and License Summary | Open Data Commons." [Online]. Available: http://opendatacommons.org/licenses/pddl/summary/. [Accessed: 15-Dec-2011].
101. ↑ "ODC Attribution Summary | Open Data Commons." [Online]. Available: http://opendatacommons.org/licenses/by/summary/. [Accessed: 15-Dec-2011].
102. ↑ "ODC Open Database License (ODbL) Summary | Open Data Commons." [Online]. Available: http://opendatacommons.org/licenses/odbl/summary/. [Accessed: 15-Dec-2011].
103. ↑ "Digital rights management - Wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/Digital_rights_management. [Accessed: 15-Dec-2011].
104. ↑ Boletín Oficial del Estado, núm. 269, 8 de noviembre de 2011.

BizkaiLab

BFA DFB Bizkaiko Foru Aldundia Diputación Foral de Bizkaia

Deusto Universidad de Deusto Deustuko Unibertsitatea