Tecnologías de la Web Semántica aplicadas a la gestión de información gubernamental: reflexiones sobre el impacto de su aplicación

Leandro Mendoza[1,2] and Alicia Díaz[2]

[1] CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
[2] LIFIA, Facultad de Informática, Universidad Nacional de La Plata, Argentina

Resumen En los últimos años las iniciativas de datos abiertos han incentivado a organismos gubernamentales a publicar información con el objetivo de contribuir a la transparencia, participación y colaboración, lo que ha dado lugar al surgimiento de nuevos retos en los procesos de gestión de la información. En este sentido, las tecnologías de la *Web Semántica* y el paradigma de *datos enlazados* proponen soluciones innovadoras para enfrentar estos desafíos y su adopción para mejorar los procesos que involucran la gestión de datos es una tendencia a nivel mundial. El objetivo de este trabajo es dar una perspectiva general de las tecnologías de la *Web Semántica* y los efectos de su aplicación en la gestión de información gubernamental considerando las etapas del ciclo de vida de los datos y las dimensiones para evaluar su calidad en cada etapa

1. Introducción

En los últimos años el concepto de *gobierno abierto* ha tomado gran relevancia en muchos países de todo el mundo como un medio esencial de comunicación entre gobiernos y ciudadanos con el objetivo de fomentar la *transparencia*, *participación* y *colaboración*. Los ejes principales de este movimiento incluyen la apertura de procesos y el uso de plataformas de participación ciudadana (*open action*) y la apertura de datos públicos (*open data*). Esto último implica la publicación de información del sector público en formatos que permitan su reutilización. Para que esto sea posible la información proporcionada debe ser completa, accesible, no restringida y su reutilización debe estar regulada por licencias libres que garanticen un uso adecuado.

Actualmente, gobiernos y agencias estatales de países que adhieren a las políticas e ideas de *datos abiertos* han comenzado a publicar una gran cantidad de información a través de portales Web de sus organismos oficiales[3]. Esta iniciativa no solo aplica a entidades gubernamentales sino que también ha sido adoptada por un número creciente de organizaciones tanto del ámbito público como privado dando lugar a una una gran cantidad y diversidad de datos en diferentes formatos que se extiende por

[3] Global Open Data Index `http://index.okfn.org/`

una amplia gama de dominios como educación, economía, agricultura, salud, bio-informática, etc. Frente a este escenario, uno de los principales desafíos es lograr mecanismos escalables que permitan publicar y reutilizar estos datos de manera eficiente. Para lograr esto, garantizar la calidad de los datos es fundamental y en consecuencia, los procesos para la gestión de la información precisan de tecnologías que permitan implementar herramientas eficaces para evaluar, detectar y eventualmente corregir problemas en los datos. En este sentido, las tecnologías de la *Web Semántica* proponen herramientas innovadoras para enfrentar estos desafíos, entre las que se incluyen:

- Lenguajes estándares para la especificación de conocimiento con una semántica bien definida (RDF, RDFS, OWL, etc.).
- Lineamientos que definen principios básicos para la publicación e interconexión de datos estructurados en la Web (*datos enlazados* o *Linked Data*).
- Mecanismos para detectar, evaluar y eventualmente corregir problemas en la calidad de los datos que pueden alinearse con modelos estándares como el ISO/IEC 25012.

Incentivadas por los beneficios que estas nuevas tecnologías pueden aportar, organizaciones gubernamentales que adhieren a la iniciativa de datos abiertos evolucionaron en la forma de publicar sus datos y comenzaron a adoptar el paradigma de *datos enlazados* en el marco de la *Web Semántica* [1]. En 2014, un relevamiento[4] realizado sobre conjuntos de datos que fueron publicados siguiendo los principios de *datos enlazados* determinó que las áreas que aportan mayor cantidad de datos son la Web Social (datos sobre perfiles personales), información gubernamental, publicaciones científicas o académicas y ciencia. En el área gubernamental, algunos ejemplos de entidades que siguen estos lineamientos son el gobierno de los Estados Unidos [2] y Reino Unido [3], entre otros. El desarrollo de portales Web dedicados es el principal medio elegido para proveer datos públicos y concentrar todas las actividades relacionadas con esta área [4]. En la República Argentina diferentes organismos a nivel Nacional, Provincial y Municipal cuentan con sus portales dedicados a la divulgación de datos del sector público con el objetivo de favorecer la transparencia. Algunos ejemplos son el portal oficial de datos abiertos[5] y el portal de datos de la Ciudad Autónoma de Buenos Aires[6], entre otros. A pesar de que la cantidad de conjuntos de datos abiertos continúa creciendo a nivel local, las tecnologías de la *Web Semántica* aún no han sido explotadas por organismos argentinos para enfrentar los desafíos y problemas que la gestión de la información pública conlleva.

[4] `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/`
[5] `http://www.datos.gob.ar/`
[6] `http://data.buenosaires.gob.ar/`

En las próximas secciones se describen las problemáticas y desafíos asociados a la publicación de datos gubernamentales (sección 2) y se realiza un breve repaso del las tecnologías de la *Web Semántica* y *datos enlazados* (sección 3). Luego se describe el ciclo de vida de *datos enlazados* (sección 4) y las dimensiones de calidad que pueden ser aplicadas para evaluar estos datos (sección 5). Finalmente, conclusiones y sugerencias se desarrollan en la sección 6.

2. Relevancia y problemática

Los principales desafíos que deben enfrentar organismos gubernamentales (y no gubernamentales) en la gestión de datos públicos están relacionados con la dificultad de publicar gran cantidad de información de calidad y dar soporte al reuso efectivo de esa información. Actualmente los datos públicos proporcionados están pensados y estructurados para el consumo de usuarios pero no son adecuados para aplicaciones o agentes de software inteligentes que requieran buscar y procesar esta información de manera automática. Algunos de los factores que contribuyen a esta limitación son:

- **Volumen y diversidad.** Los datos publicados son proporcionados a través de diferentes formatos o medios (p. ej., CSV, XML, HTML, Servicios Web, etc.), cubren una amplia gama de dominios (p. ej., sector público, economía, educación, salud, etc.) y están bajo normativas o licencias diferentes dependiendo de la entidad responsable o propietaria de esos datos.

- **Información no estructurada.** Los datos publicados son proporcionados sin procesamiento previo (*raw data*), no cuentan con metadatos suficientes que describan su estructura y significado (p. ej., falta de información sobre los campos de una base de datos o sobre el esquema utilizado para modelarlos) lo que dificulta su interpretación sin ambigüedades, su procesamiento automático y su reutilización.

- **Interoperabilidad.** Los datos publicados representan información aislada sobre una actividad de algún organismo particular que por lo general no puede ser integrada, combinada o relacionada fácilmente con otras bases de datos existentes (internas o externas) lo que dificulta su reutilización y el descubrimiento de nueva información que puede surgir al cruzar datos de diferentes dominios.

- **Calidad.** Los datos publicados no cuentan con especificaciones que garanticen un nivel de calidad adecuado para un uso específico. En muchos casos no cuentan con información confiable sobre su procedencia, es muy difícil determinar si están bajo alguna licencia de reuso específica, quién es la entidad responsable de los mismos, cómo fueron extraídos y de dónde, con qué frecuencia son actualizados o si son sometidos a algún proceso de evaluación de calidad antes de ser publicados.

3. Web semántica y datos enlazados

El concepto de *Web Semántica* fue concebido por Berners-Lee como una extensión de la *World Wide Web* (WWW) en donde la información tiene asociado un significado bien definido (semántica) que permite a los usuarios y a las computadoras buscar, combinar y procesar contenidos [5]. Desde un punto de vista práctico, el concepto de *Web Semántica* refiere al conjunto de tecnologías y estándares que permiten a las computadoras o agentes de software inteligentes comprender el significado de la información disponible en la Web [6]. Mas allá de las diferentes perspectivas que se puedan encontrar en las definiciones todas coinciden en que el significado de los recursos en la Web debe estar especificado explícitamente; no es suficiente que los datos estén codificados usando una sintaxis bien definida sino que también se requiere que tengan asociada una semántica que describa su significado y determine cómo deben ser interpretados.

El lenguaje estándar para representar la información dentro del contexto de la *Web Semántica* es RDF[7]. Las unidades más pequeñas de información son declaraciones (*statements*) del tipo "sujeto, predicado, objeto" y reciben el nombre de *tripletas*. Cada tripleta representa un hecho específico y un conjunto de tripletas conforman un grafo RDF que se corresponde con un bloque de información o base de conocimiento. Todo lo que pueda describirse utilizando RDF (sujetos y objetos que representen entidades concretas o abstractas) se denomina *recurso* y las propiedades (predicados) definen relaciones entre recursos. Estos recursos y propiedades son identificados de manera única en la Web mediante la utilización de URIs (*uniform resource identifier*). RDF es la base fundamental sobre la que se sustentan las ideas de la *Web Semántica* y representa para este paradigma lo que HTML representa para la Web convencional. Sobre RDF se definen otros lenguajes de especificación de conocimiento como RDFS[8] y OWL[9] que agregan expresividad y permiten definir vocabularios para dominios específicos o modelar conceptos y relaciones complejas (taxonomías, jerarquías de clases, etc.) a través de ontologías. Esto es posible debido a que los estándares mencionados tienen una semántica bien definida [7] basada en fundamentos teóricos como la teoría de grafos y la lógica descriptiva lo que posibilita, entre otras cosas, la implementación de mecanismos de inferencia mediante razonadores para la generación de nueva información. La información contenida en conjuntos de datos que se describen utilizando RDF, RDFS u OWL puede dividirse en dos grupos: información sobre el modelo de datos o a nivel de esquema (p. ej., una jerarquía de clases) e información que describe recursos utilizando estos modelos o a nivel de instancia (p. ej., información que especifica recursos como miembros de una clase). Por lo general, los datos en RDF se almacenan en repositorios denominados *triplestores* y son accedidos utilizando un lenguaje de consulta específico llamado *SPARQL*[10]. El me-

[7] `https://www.w3.org/RDF/`
[8] `https://www.w3.org/TR/rdf-schema/`
[9] `https://www.w3.org/2001/sw/wiki/OWL`
[10] `https://www.w3.org/TR/rdf-sparql-query/`

dio más utilizado para hacer estas consultas es a través de servicios Web que reciben el nombre de *SPARQL endpoints*.

A modo de ejemplo, la figura 1 muestra un fragmento de la estructura organizativa del Poder Ejecutivo Nacional (PEN) de la República Argentina que ha sido modelada utilizando los conceptos mencionados previamente. Los datos a nivel de esquema representan una jerarquía de clases que modelan tres tipos de organismos (OrganizacionGubernamental, Ministerio y Secretaría) y la relación entre clases se establece a través del predicado `rdfs:subClassOf`. Utilizando este esquema se modelan cuatro organismos específicos: *Poder Ejecutivo Nacional* (PEN), *Secretaría General de la Presidencia* (SGP), *Jefatura de Gabinete de Ministros* (JGM) y *Presidencia de la Nación* (PN). Luego, los recursos que representan a estas entidades se especifican como instancias de las clases previamente definidas a través del predicado `rdf:type`. Relaciones de pertenencia y propiedades se describen mediante un vocabulario ilustrativo de dominio específico utilizando los predicados `pertenece_a`, `depende_de`, `nombre` y `sitio_web`. Todos los recursos y propiedades que aparecen en el diagrama (con excepción de los valores de propiedades que son cadenas de texto) se identifican mediante URIs (p. ej., `http://example.org/recurso/pen`, `http://example.org/vocabulario/sitio_web`, `http://example.org/vocabulario/Secretaria`, etc.).

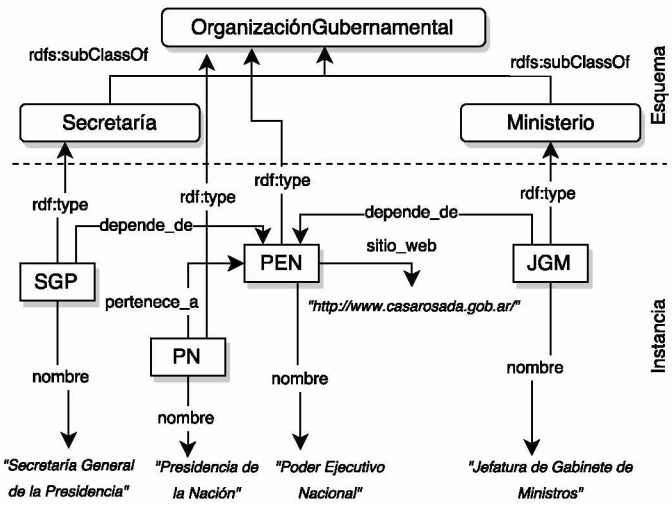

Figura 1. Ejemplo de grafo RDF que describe organismos y relaciones

El concepto de *datos enlazados* (o *Linked Data*) hace referencia a un conjunto de buenas prácticas[11] para la publicación de datos estructurados en la Web y para establecer enlaces entre ellos con el objetivo de conformar

[11] `https://www.w3.org/DesignIssues/LinkedData.html`

una gran red de datos interconectados (*Web of Data*). Originalmente, el término fue propuesto por Tim Berners-Lee y se basa en cuatro principios básicos: i) utilizar URIs para identificar recursos, ii) utilizar HTTP URIs para que se pueda acceder a esos recursos, iii) cuando se accede a un recurso, proveer información utilizando estándares como RDF o SPARQL y finalmente, iv) establecer enlaces entre las URIs de los recursos para facilitar el descubrimiento de información. Utilizando estos principios se ha propuesto un sistema de estrellas que determina que tan abierto y usable es un conjunto de datos. Los principios se describen y organizan de menor a mayor en 5 puntos, cada punto incluye las características del punto anterior y tiene asignado una cantidad de estrellas en función de los requerimientos alcanzados por los datos:

★ Los datos están publicados en la Web bajo una licencia abierta y con formatos no estructurados (p. ej., PDF, CSV, HTML, imágenes, etc.)

★★ Los datos están publicados de manera estructurada y pueden ser procesados por algoritmos computacionales (p. ej., una tabla Excel en vez de una imagen).

★★★ Los datos están publicados en formatos no propietarios (p. ej., un archivo CSV en vez de una tabla Excel).

★★★★ Los datos publicados utilizan estándares Web para identificar recursos (HTTP URIs) y estándares sugeridos por la W3C (p. ej., RDF, SPARQL, etc.) para representar y proporcionar información sobre estos recursos.

★★★★★ Los datos publicados están enlazados con datos de otros proveedores. En la práctica, esto implica que el conjunto de datos incluye relaciones a recursos en conjuntos de datos externos.

Es importante destacar que el concepto de datos enlazados no necesariamente implica que estos datos sean abiertos o de utilización libre y gratuita. Al combinar los conceptos de datos enlazados (*Linked Data*) y datos abiertos (*Open Data*) surge lo que se conoce como datos abiertos enlazados (*Linked Open Data* o LOD).

4. Ciclo de vida de datos enlazados

Además de las guías de referencia que especifican buenas prácticas para la publicación de *datos enlazados* podemos encontrar trabajos que adaptaron estos lineamientos a metodologías generales [8] o a metodologías más específicas para escenarios concretos como la publicación de datos bibliográficos [9], datos estadísticos [10] o datos relacionados a organismos gubernamentales [11]. En términos generales, el proceso de publicación de datos enlazados es un ciclo iterativo e incremental que involucra decisiones administrativas (p. ej., decidir que datos se van a

publicar y con que fin), decisiones relacionadas a la gestión de la información (p. ej., determinar que requerimientos de calidad deben cumplir los datos), decisiones relacionadas con el diseño y la infraestructura (p. ej., definir modelos de datos y tecnologías a utilizar), etc. Considerando el escenario de organismos gubernamentales, las cinco actividades principales del ciclo de vida de *datos enlazados* propuesto por Villazon et al. en [11] pueden observarse en la figura 2 y se repasarán brevemente en las próximas subsecciones.

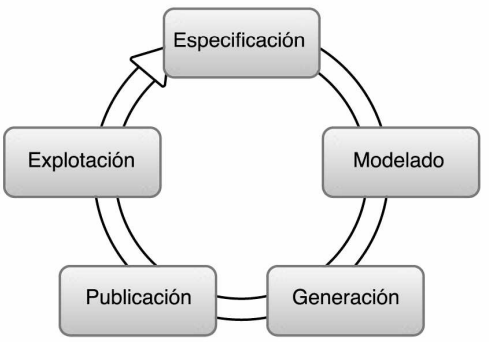

Figura 2. Ciclo de vida de *datos enlazados* gubernamentales propuesto por Villazon et al. en [11]

4.1. Especificación

La primera actividad de este ciclo está relacionada con la especificación de requerimientos y puede subdividirse en tres tareas:

- **Identificación y análisis de las fuentes de datos gubernamentales**. Esta tarea consiste en identificar y seleccionar la información que se desea publicar. Los datos seleccionados pueden ser información pública que aún no ha sido publicada o información que ya ha sido publicada y que se encuentra disponible a través de catálogos de datos o portales Web pero que no sigue los principios de *datos enlazados*. Una vez identificadas las fuentes de datos, es necesario recolectar toda la información disponible relativa a cada fuente: propósito de publicación, modelo o esquema de datos (metadatos, componentes conceptuales y relaciones) y detalles de implementación. Uno de los propósitos principales de estas tareas es identificar los *ítems* de interés del dominio, es decir, recursos cuyas propiedades y relaciones serán descriptas en el conjunto de datos a publicar.
- **Diseño de URIs**. Esta tarea consiste en definir el diseño de las URIs que se utilizarán para identificar los recursos. Es deseable que estas URIs sean simples, expresivas, estables, fáciles de manejar en el

idioma oficial del gobierno (si es posible), etc. En [11], puede encontrarse una guía para el diseño de URIs en ámbitos gubernamentales, basada en documentos previos del diseño de URIs [12] para escenarios más generales y en experiencias propias de los autores.

- **Definición de licencia**. Esta tarea consiste en determinar la licencia de los datos que serán publicados. Para datos gubernamentales, algunas de las licencias actualmente utilizadas son: UK Open Goverment Licence[12], Open Database Licence[13], Open Data Common Attribution Licence[14], Creative Commons Licences[15], etc.

4.2. Modelado

Una vez identificadas, seleccionadas y analizadas las fuentes de datos es necesario precisar que vocabularios u ontologías se utilizarán para modelar el dominio de esos datos. En este sentido, la principal recomendación es reutilizar las ontologías existentes tanto como sea posible. Sitios Web como *Linked Open Vocabularies* [16] o *Swoogle*[17] actúan como repositorios que facilitan la búsqueda de estos vocabularios y permiten tener una idea de cómo y dónde utilizarlos. Para modelar estos datos es posible combinar más de un vocabulario, extender un vocabulario existente o crear un vocabulario completamente nuevo en caso de que las opciones disponibles no cubran los requerimientos necesarios. Para este último caso, se recomienda la implementación de metodologías conocidas para el diseño de ontologías [13].

4.3. Generación

El principal objetivo de esta actividad es tomar las fuentes de datos seleccionadas durante la *especificación* y transformarlas a RDF acorde a las ontologías definidas en la actividad de *modelado*. Esta generación puede dividirse en tres tareas:

- **Transformación.** Esta tarea consiste en la generación de tripletas RDF a partir de las fuentes de datos. Es deseable que esta generación sea lo más completa posible (todos los datos de la fuente de datos original deben estar disponibles en RDF) y se ajuste a las ontologías o vocabularios seleccionados en la etapa de modelado (los datos de la fuente original deben corresponderse con la estructura o semántica del modelo de datos a utilizar). Actualmente existe una gran variedad de herramientas utilizadas para convertir datos a RDF según los tipos de datos de la fuente original.

[12] `http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/`
[13] `http://opendatacommons.org/licenses/odbl/`
[14] `http://opendatacommons.org/licenses/by/`
[15] `https://creativecommons.org/licenses/?lang=es`
[16] `http://lov.okfn.org/dataset/lov/`
[17] `http://swoogle.umbc.edu/`

- **Limpieza de datos.** Esta tarea consiste en la detección y corrección de errores sintácticos y semánticos. Algunos de estos errores pueden ser heredados de la fuente de datos original y otros pueden surgir durante alguna de las tareas del ciclo de vida de *datos enlazados*. Por este motivo, corregir errores en los datos es una tarea que puede extenderse a otras actividades del ciclo y está estrictamente relacionada con varias de las dimensiones de calidad que se mencionan en la sección 5.
- **Enlazado**: Esta tarea consiste en establecer relaciones o enlaces entre recursos del conjunto de datos gubernamentales y otros recursos de conjuntos de datos publicados bajo los principios de *datos enlazados* que pueden ser gubernamentales o no. La generación de estos enlaces (tarea conocida como *Link Discovering*) puede realizarse en forma manual, semi-automática o automática y comprende tres pasos principales: i) identificar posibles conjuntos de datos a relacionar, ii) descubrir relaciones entre recursos del conjunto de datos gubernamental y los identificados en i) y iii) validar las relaciones que se han descubierto. En [14] puede encontrarse una descripción exhaustiva de diferentes herramientas y *frameworks* para el descubrimiento de enlaces en el contexto de *datos enlazados*.

4.4. Publicación

Una vez generados los datos en RDF es necesario publicarlos para que puedan ser accedidos. Esta actividad comprende tres tareas principales:

- **Almacenamiento del conjunto de datos**. Esta tarea consiste en determinar como serán almacenados y accedidos los datos. La forma habitual de almacenamiento de datos en RDF es a través de repositorios denominados *triplestores* que ya proveen mecanismos para almacenar y consultar estos datos (p. ej., a través de SPARQL *endpoints*). Una de las herramientas más utilizadas como *triplestore* es *OpenLink Virtuoso* [18].
- **Publicación de meta-datos**. Esta tarea consiste en describir meta-información sobre el conjunto de datos: fecha de publicación, autores o responsables, información de procedencia, vocabularios u ontologías que utiliza, con que otros conjuntos de datos está relacionado, etc. Actualmente, el vocabulario más utilizado para describir conjuntos de datos es VoID[19].
- **Permitir el descubrimiento del conjunto de datos**. Esta tarea consiste en hacer visible el conjunto de datos a los potenciales usuarios mediante la publicación de meta-datos (generados en la tarea anterior) en catálogos generales como *Datahub*[20] o en catálogos específicos de datos abiertos gubernamentales[21].

[18] http://virtuoso.openlinksw.com/
[19] https://www.w3.org/TR/void/
[20] `https://datahub.io/`
[21] `http://opengovernmentdata.org/data/catalogues`

4.5. Explotación

El objetivo final de abrir los datos gubernamentales es dar transparencia y fomentar el acceso público y reuso de la información. Por lo tanto, esta actividad se refiere al desarrollo de aplicaciones (*Linked Data Applications* [15]) que exploten los datos generados combinando información de diferentes fuentes, descubriendo o generando nueva información y facilitando el acceso a través de interfaces visuales ricas y amigables.

5. Calidad de datos enlazados

El concepto de *calidad de los datos* comúnmente se relaciona a la medida en que los datos son "adecuados para el uso" (*fitness for use*), es decir, que los datos sean de utilidad o cumplan con los requerimientos para una aplicación o caso de uso específico. Actualmente, herramientas basadas en tecnologías de la *Web Semántica* aportan mecanismos innovadores para gestionar la calidad de los datos y pueden integrarse al ciclo de vida de la información. En [16] los autores realizaron un análisis exhaustivo de *frameworks* y herramientas para la evaluación de calidad enfocados principalmente en su aplicación a *datos enlazados*. En el mencionado trabajo se identifican 18 dimensiones (o atributos) de calidad agrupadas en cuatro categorías (accesibilidad, intrínsecas, contextuales y representación - ver figura 3) y 69 métricas diferentes (cuantitativas y cualitativas). A continuación, se repasan brevemente las definiciones de cada uno de estos grupos y dimensiones.

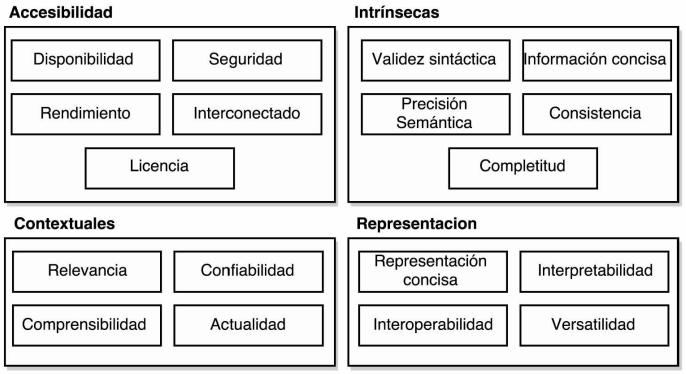

Figura 3. Dimensiones de calidad para datos enlazados propuestas por Zaveri et. al en [16]

5.1. Accesibilidad

Las dimensiones o atributos de calidad que conforman este grupo están relacionados con los mecanismos de acceso y autenticidad del conjunto

de datos:

- **Disponibilidad.** Refiere a los mecanismos de acceso al conjunto de datos y si estos funcionan correctamente. Incluye el análisis de accesibilidad mediante puntos de acceso SPARQL (*endpoints*) o RDF *dumps* (archivos para ser descargados), acceso a información sobre recursos mediante sus URIs, etc.
- **Seguridad.** Refiere a los mecanismos que aseguran la información contra alteraciones y usos indeseados. Incluye el análisis del uso de firmas digitales para documentos RDF o datos entregados a través de SPARQL *endpoints* y la verificación de autenticidad a través de la información disponible sobre la procedencia del conjunto de datos: autores, colaboradores, editor y origen de la fuente de original.
- **Rendimiento.** Refiere a la eficiencia con la cual son procesados y entregados los datos. Incluye el análisis de los tiempos de latencia (p. ej., tiempo que se tarda en retornar los datos requeridos por una consulta SPARQL), la cantidad máxima de requerimientos de datos que se pueden atender al mismo tiempo o la capacidad de mantener el rendimiento de estas variables a medida que el conjunto de datos escala en tamaño o complejidad.
- **Interconectado.** Refiere al grado en que entidades que representan un mismo concepto (o similar) están relacionadas entre si, ya sea dentro de un mismo conjunto de datos o entre conjuntos de datos externos. Incluye el análisis, validación y verificación (sintáctica y semántica) de enlaces entre recursos.
- **Licencia.** Refiere a la información disponible sobre normas que regulan la reutilización del conjunto de datos. Incluye la especificación de la licencia adoptada en forma comprensible para los usuarios y también en formatos que posibiliten el procesamiento automático por agentes de software.

5.2. Intrínsecas

En este grupo se encuentran las dimensiones de calidad que son independientes del contexto del usuario:

- **Validez sintáctica.** Refiere al grado con que un documento RDF es válido con respecto a la especificación del lenguaje. Incluye el análisis de errores en los tipos de datos, detección de caracteres sintácticamente inválidos, uso incorrecto de vocabularios, etc.
- **Precisión semántica.** Refiere al grado con que los datos representan efectivamente hechos del mundo real. Incluye el análisis de valores atípicos (*outliers*) y semánticamente incorrectos a través de técnicas basadas en métodos estadísticos o en la definición de reglas (p. ej., dependencias funcionales) que imponen restricciones a los datos.
- **Consistencia.** Refiere a la medida en que la información contiene (o no) contradicciones (lógicas o formales) con respecto a los mecanismos de representación de conocimiento e inferencia utilizados.

- **Completitud.** Refiere al grado en que la información requerida para una tarea o aplicación específica está disponible en el conjunto de datos.
- **Información concisa.** Refiere a la minimización de información redundante tanto a nivel de esquema (redundancia en propiedades o clases) como a nivel de instancia (información adicional sobre recursos que no es necesaria).

5.3. Contextuales

Las dimensiones de calidad de este grupo son aquellas que tienen una alta dependencia con el contexto:

- **Relevancia.** Refiere a que la información provista sea adecuada para los propósitos con que fue publicada y de interés para los usuarios que la utilizan. Incluye, por ejemplo, el análisis de la cantidad de recursos que se describen en el conjunto de datos y con que nivel de detalle.
- **Integridad.** Refiere al grado con que la información es aceptada como correcta, verdadera y creíble. Esta dimensión se relaciona con la *fiabilidad* de los datos y puede incluir varias sub-dimensiones: integridad en las declaraciones (*statements*) y recursos, integridad de los vocabularios utilizados, reputación del conjunto de datos, etc.
- **Comprensibilidad.** Refiere a la facilidad con que la información puede ser comprendida sin ambigüedades y utilizada por los usuarios. Factores que contribuyen a la comprensibilidad son: el uso de etiquetas legibles para clases, propiedades y metadatos, el diseño de URIs claras o que sigan un patrón de diseño, la especificación de los vocabularios y ontologías utilizados, proveer fragmentos del conjunto de datos y consultas SPARQL como ejemplos, etc.
- **Actualidad.** Refiere a cuán actualizado está un conjunto de datos con respecto a la utilización que se le va a dar. La actualidad de los datos puede analizarse considerando la fecha de publicación o generación, el periodo de tiempo en que los datos se mantienen válidos (volatilidad), la diferencia en los tiempos de actualización entre el conjunto de datos y la fuente de datos original, etc.

5.4. Representación

Las dimensiones de calidad de este grupo están relacionadas con aspectos de diseño y modelado de los datos:

- **Representación concisa.** Refiere a que los datos estén representados de forma correcta; compactos, bien formateados, en forma clara y completos. Esta dimensión está relacionada, por ejemplo, con el uso de URIs cortas y de primitivas RDF prolijas.
- **Interoperabilidad.** Refiere al grado en que el formato y estructura de los datos favorece la reutilización del conjunto de datos. Esta dimensión está relacionada con el reuso de términos y vocabularios relevantes que existen dentro del dominio en cuestión.

- **Interpretabilidad.** Refiere a aspectos técnicos de los datos que determinan si la información está representada utilizando notación adecuada y si es fácilmente procesable por computadoras. Esta dimensión esta relacionada con el uso de vocabularios auto-descriptivos, definiciones claras y lenguajes, símbolos, unidades y tipos de datos adecuados.
- **Versatilidad.** Refiere a que la información pueda estar disponible en diferentes representaciones (*serialization*) y de manera internacionalizada (diferentes idiomas).

6. Conclusión y sugerencias

Si bien existe una gran cantidad de datos públicos disponibles a través de la Web fomentada por la iniciativa de *datos abiertos*, el descubrimiento, reusabilidad y explotación efectiva de estos datos están limitados debido a que gran parte de esta información se encuentra en forma no estructurada, en formatos no estándares y por lo general no cuenta con especificaciones de calidad claras que aseguren que esté libre de errores o que cumpla con ciertas normas de calidad adecuadas para un uso específico. En consecuencia, los conjuntos de datos y los esquemas para modelarlos conforman islas de información que difícilmente pueden ser compartidos o interrelacionados, aun cuando pertenecen a sectores de una misma organización o describen recursos similares, lo que dificulta la integración, el descubrimiento de nueva información e incrementa los costos y tiempos destinados a la gestión de la información.

Las tecnologías de la *Web Semántica* se presentan como herramientas innovadoras para dar soporte a la solución de algunos de los desafíos que la gestión de la información en organismos gubernamentales conlleva. Si bien su potencial aún no ha sido explotado, algunos proveedores de datos tienden a la adopción de estas tecnologías que representan una evolución en la manera de gestionar y publicar información. El modelo RDF como lenguaje estándar permite que los datos provenientes de fuentes variadas y en diferentes formatos se puedan describir utilizando un único lenguaje fundamental, lo que favorece la interoperabilidad y el intercambio a través de la Web. La flexibilidad y expresividad de los lenguajes RDFS y OWL permiten modelar una gran diversidad de información de diferentes dominios. La existencia de vocabularios y ontologías abiertos actualmente disponibles en la Web que se describen mediante estos lenguajes facilitan la construcción de nuevos modelos ya que pueden ser fácilmente reutilizados o extendidos. Las metodologías para la generación de *datos enlazados* y los modelos de calidad de datos en este contexto pueden ser combinados para redefinir políticas de gestión de la información gubernamental que apliquen tanto a los datos públicos como a los datos internos de una organización. Desde el punto de vista práctico, la reutilización de técnicas existentes que aprovechan la semántica de los datos facilita la implementación de métricas y herramientas para detectar, evaluar y eventualmente corregir problemas en la información. En base a estas conclusiones, se sugiere:

- Fomentar el uso de estándares y la adopción de tecnologías de la *Web Semántica* para la gestión de información gubernamental, tanto para los datos internos (utilizados y compartidos dentro de una organización) como así también para los datos públicos. Esta adopción no implica un reemplazo de tecnologías actuales, sino que puede implementarse como extensión (a más alto nivel) de los mecanismos ya utilizados.

- Adoptar los principios de *datos enlazados* para especificar e implementar en cada organización metodologías que cubran todo el ciclo de vida de la información (desde su generación hasta su publicación). Estas metodologías deben contemplar las políticas necesarias para garantizar la calidad de la información acorde a ciertos objetivos predefinidos y siguiendo un estándar o norma específico.

- Implementar portales Web dedicados a divulgar las políticas de una organización con respecto a la información que produce. Estos portales deben facilitar la búsqueda de conjuntos de datos relativos a la organización, documentación, guías, tutoriales y pueden servir como medio de divulgación de noticias y boletines.

Considerando estas sugerencias, se propone la creación del portal **LODArg** (Linking Open Data Argentina) como iniciativa general para fomentar la publicación de bases de datos abiertas de la República Argentina siguiendo los principios de *datos enlazados* y tecnologías de la *Web Semántica*. El objetivo es brindar soporte a las organizaciones que deseen adoptar estas nuevas tecnologías; compartir herramientas y experiencias relativas a esta área, promover el desarrollo de ontologías y vocabularios comunes, facilitar la búsqueda de conjuntos de datos y promover el desarrollo de aplicaciones basadas en estos datos.

Referencias

1. M. A. Hallo, M. M. Martínez-González, and P. de la Fuente Redondo, "Las tecnologías de linked data y sus aplicaciones en el gobierno electrónico," *Scire: representación y organización del conocimiento*, vol. 18, no. 1, pp. 49–61, 2012.
2. J. A. Hendler, J. Holm, C. Musialek, and G. Thomas, "Us government linked open data: Semantic.data.gov." *IEEE Intelligent Systems*, vol. 27, no. 3, pp. 25–31, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/expert/expert27.html#HendlerHMT12
3. J. Sheridan and J. Tennison, "Linking uk government data." in *LDOW*, ser. CEUR Workshop Proceedings, C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, Eds., vol. 628. CEUR-WS.org, 2010. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/ldow2010.html#SheridanT10

4. L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, and J. A. Hendler, "Twc logd: A portal for linked open government data ecosystems." *J. Web Sem.*, vol. 9, no. 3, pp. 325–333, 2011. [Online]. Available: http://dblp.uni-trier.de/db/journals/ws/ws9.html#DingLEDWLMGZSFMH11
5. T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.
6. L. Yu, *A Developer's Guide to the Semantic Web*. Springer, 2011.
7. P. Hitzler, M. Krötzsch, and S. Rudolph, *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, 2009.
8. S. Auer, J. Lehmann, A.-C. N. Ngomo, and A. Zaveri, "Introduction to linked data and its lifecycle on the web," in *Reasoning Web*, 2013, pp. 1–90. [Online]. Available: http://jens-lehmann.org/files/2013/reasoning_web_linked_data.pdf
9. D. Zengenene, V. Casarosa, and C. Meghini, "Towards a methodology for publishing library linked data," in *Bridging Between Cultural Heritage Institutions*, ser. Communications in Computer and Information Science, T. Catarci, N. Ferro, and A. Poggi, Eds. Springer Berlin Heidelberg, 2014, vol. 385, pp. 81–92. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-54347-0_10
10. I. Petrou, M. Meimaris, and G. Papastefanatos, "Towards a methodology for publishing linked open statistical data," *JeDEM-eJournal of eDemocracy and Open Government*, vol. 6, no. 1, pp. 97–105, 2014.
11. B. Villazón-Terrazas, L. Vilches, O. Corcho, and A. Gómez-Pérez, "Methodological guidelines for publishing government linked data," in *Linking Government Data*, D. Wood, Ed. Springer, 2011, ch. 2.
12. "Cool uris for the semantic web," DFKI GmbH, Technical Memo TM-07-01, February 2007, written by 29.11.2006. [Online]. Available: http://www.dfki.uni-kl.de/dfkidok/publications/TM/07/01/tm-07-01.pdf
13. M.-C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, *Ontology Engineering in a Networked World*, M.-C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, Eds. Berlin: Springer, 2012.
14. M. Nentwig, M. Hartung, A.-C. N. Ngomo, and E. Rahm, "A survey of current link discovery frameworks," *Semantic Web*, no. Preprint, pp. 1–18, 2015. [Online]. Available: http://www.semantic-web-journal.net/system/files/swj1117.pdf
15. M. Hausenblas, "Linked data applications," *First Community Draft, DERI*, 2009.
16. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," *Semantic Web Journal*, 2015. [Online]. Available: http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey