Ministerio de Tecnologías de la Información y las Comunicaciones
República de Colombia
www.mintic.gov.co

ANEXO No.3: EL LENGUAJE COMÚN DE INTERCAMBIO DE INFORMACIÓN EN EL CONTEXTO DE LOS DATOS ABIERTOS EN COLOMBIA

LINEAMIENTOS PARA LA IMPLEMENTACION DE DATOS ABIERTOS EN COLOMBIA

CONVENIO INTERADMINISTRATIVO DE COOPERACION N° 308 DE 2011

CINTEL
CENTRO DE INVESTIGACION DE LAS TELECOMUNICACIONES

Coordinación de Investigación, Políticas y Evaluación
Estrategia de Gobierno en línea

Bogotá, DC., Diciembre de 2011

FORMATO PRELIMINAR AL DOCUMENTO

Título:	*El Lenguaje Común de Intercambio de Información en el contexto de Datos de Gobierno Abiertos en Colombia*				
Fecha elaboración	*2011-12-07*				
Sumario:	Este documento es un Anexo del documento titulado "Lineamientos para la Implementación de Datos Abiertos en Colombia" y la temática es la competencia de usar el Lenguaje Común de Intercambio de Información en el contexto de la apertura de los datos públicos del Gobierno Colombiano.				
Palabras Claves:	*Interoperabilidad, Lenguaje de Intercambio, Datos Abiertos, Colombia*				
Formato:	DOC		Lenguaje:		Español
Dependencia:	Ministerio de Tecnologías de la Información y las Comunicaciones: Programa Agenda de Conectividad. Estrategia de Gobierno en línea. *Coordinación de Investigación, Políticas y Evaluación*				
Código:	N/A	Versión:	2.0	Estado:	Final
Categoría:	Documento técnico				
Autor (es):	Centro de Información de las Telecomunicaciones - CINTEL		Firmas:		
Revisó:	Enrique Cusba García Líder de Investigación y Políticas Programa Agenda de Conectividad- Estrategia de Gobierno en línea				
Aprobó:	Ana Carolina Rodríguez Coordinadora Coordinación de Investigación, Políticas y Evaluación. Estrategia Gobierno en Línea				
Información Adicional:	No adicional				
Ubicación:	Programa Gobierno en línea				

CONTROL DE CAMBIOS

VERSIÓN	FECHA	DESCRIPCIÓN
1.0	2011-11-01	Versión inicial
1.1	2011-11-30	Se incluye la explicación de conceptos claves en el ámbito semántico
2.0	2011-12-26	Versión final

TABLA DE CONTENIDO

1. INTRODUCCIÓN

El objetivo del presente documento es describir la utilización del Lenguaje Común de Intercambio de Información en el contexto de la apertura de los datos públicos del Gobierno Colombiano y brindar las bases para la evolución de dicho lenguaje en este contexto.

Con este objetivo en mente, es necesario explicar primero:

- ¿Qué son datos abiertos?
- ¿Qué se persigue con la apertura de datos?
- ¿Quiénes son los destinatarios de los datos abiertos?
- ¿Cómo se abren los datos?

Aclarados estos puntos, se analizará en qué situaciones y en qué medida el Lenguaje Común de Intercambio de Información puede dar respuesta a las necesidades de apertura de los datos del Gobierno Colombiano.

2. DATOS ABIERTOS

2.1 ¿Qué son Datos Abiertos?

Dentro de su estrategia de gobierno abierto para los próximos años, el Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia, ha tomado la decisión de definir y construir una Plataforma de Datos de Gobierno Abiertos. Esta decisión se enmarca en la filosofía y movimiento social[1] de "Datos de Gobierno Abiertos" con los objetivos de impulsar la innovación y el crecimiento económico, mejorar la eficiencia administrativa y contribuir al incremento de la democracia.

En síntesis, la apertura de datos trata de identificar los conjuntos de documentos, información y datos públicos, gestionados por la Administración, y de implementar las soluciones técnicas necesarias para ponerlos a disposición de los sistemas de información de terceras partes, como por ejemplo:

- Aplicaciones para ciudadanos que consumen dicha información.
- Servicios de sectores industriales que añaden valor a la información y, con el resultado, producen y venden nuevos productos o servicios (industria denominada "infomediaria"[2]).
- Sistemas de Administraciones que ingresan los datos, la información y los documentos para el desarrollo de sus funciones públicas.

2.2 ¿Qué se persigue con la apertura de datos?

La apertura de datos del Gobierno es uno de los facilitadores más potentes para el desarrollo de lo que se está denominado "**Gobierno Abierto**". La apertura de los datos confirma la obtención de beneficios en tres ejes clave:

1) **Transparencia**. Abrir datos sobre sus actuaciones supone:

[1] Ver el espíritu del movimiento en http://www.opengovdata.org/home/8principles. Sede del grupo de trabajo creado en la cumbre de Sebastopol (California), en 2007, donde se desarrollaron los 8 principios de los Datos de Gobierno Abiertos.

[2] "El término "informediario" fue acuñado por consultores de McKinsey y los profesores John Hagel III y Marc Singer en su libro NetWorth. Aunque originalmente el contexto donde se originó el término era diferente, el modelo de negocio de la industria infomediaria reconoce la existencia de valor en los datos y busca actuar como agente de confianza para proveer la oportunidad y los medios para monetizar la adición de valor a los datos originales, así como para que sus consumidores extraigan beneficio de los resultados proveídos por su intermediación. Hoy en día se utiliza ampliamente para referirse a un conjunto de empresas que sacan provecho económico de la explotación de los datos originalmente gestionados por las Administraciones Públicas. Ver, por ejemplo, el "Estudio de Caracterización del Sector Infomediario Español", disponible en la siguiente URL: http://www.aporta.es/web/guest/estudioRISP2011.

- Un ejercicio de **democracia** por parte del gobierno y
- Un instrumento para la **rendición de cuentas**

2) **Eficiencia**. La apertura metodológica de datos de calidad puede conducir a:
   - La **racionalización** de los **procesos administrativos,** y a
   - La **automatización** de la **colaboración interadministrativa**
3) **Innovación y crecimiento económico**. La reutilización de los datos por parte del Sector Público impulsa:
   - La creación de **nuevos servicios y productos** de valor añadido y, consecuentemente,
   - **Nuevos negocios y puestos de trabajo**.

2.3 ¿Quienes son los destinatarios de los datos abiertos?

Dados los propósitos mencionados en el apartado anterior, los datos suelen tratarse con finalidades de agregación de valor, se trate de valor informativo, de transformación y presentación o de servicio.

Es por ello que se distingue muy claramente entre el consumidor final –la persona que aprovecha la información- y el consumidor intermedio, aquel o aquellos que prepara(n) los datos para su consumo final.

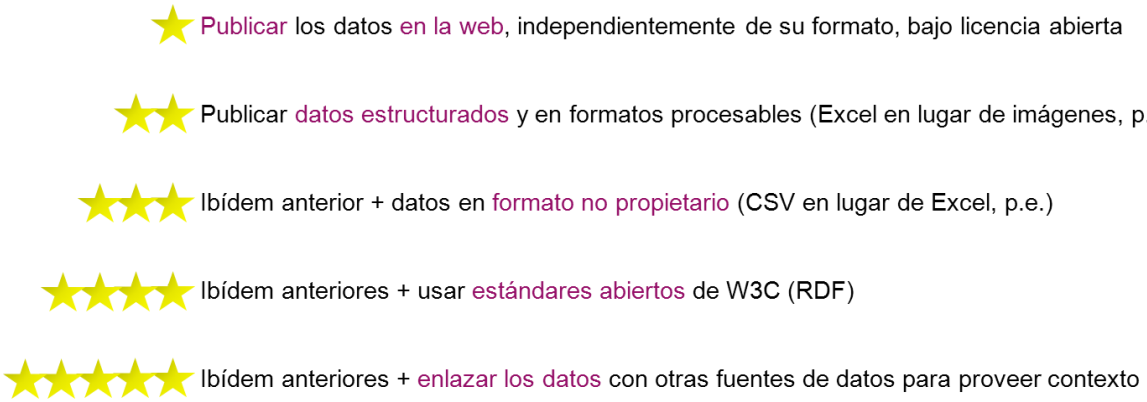
Estos intermediarios usan aplicaciones informáticas para la mecanización industrializada de la agregación de valor a los datos originales. Lo que ha dado lugar a un tipo de industria denominado “Sector Infomediario”.

Para que dicha industria pueda funcionar, **es necesario que los datos que se abren se presenten en soportes informáticos y que sus contenidos sean estructurados** mediante lenguajes procesables por aplicaciones informáticas.

2.4 ¿Cómo se abren los datos?

Muchos gobiernos con proyectos de Datos Abiertos planifican la apertura de sus datos progresivamente en base a una escala conocida como las 5 estrellas de Berners-Lee (entre más estrellas más óptimo el tratamiento automático de los datos y mejores posibilidades de reutilización):

Figura 1: Clasificación 5 estrellas de Tim Berners-Lee



Fuente: Tim Berners-Lee, TED 2009

Como se mencionó, el destino de los datos es ser procesado por sistemas de información. De ahí que el **descubrimiento** de las fuentes de datos y la **navegación** entre los diferentes conjuntos de datos **por medios automáticos** constituyen un aspecto clave.

La internet de las páginas web actual, cuyo destinatario es el humano, no es una solución válida para impulsar la apertura de los datos. Si bien existe la posibilidad de analizar y extraer información de las páginas web, expresadas en formato HTML, esta técnica resulta cara e insostenible.

Una solución alternativa, impulsada por el mismo inventor de la web actual, Tim Berners-Lee, es la construcción de otro tipo de web denominada la **Web de los Datos**. En ella, los datos se representan mediante un lenguaje llamado **RDF** (Resource Description Framework).

RDF expresa conceptos y relaciones entre ellos con base en una estructura muy simple denominada "tripleta". Una Tripleta RDF está compuesta por:

Un Concepto (Sujeto) + Una relación (Propiedad) + Un Concepto (Objeto)

O bien, por:

Un Concepto (Sujeto) + Una relación (Propiedad) + Un Valor (Literal)

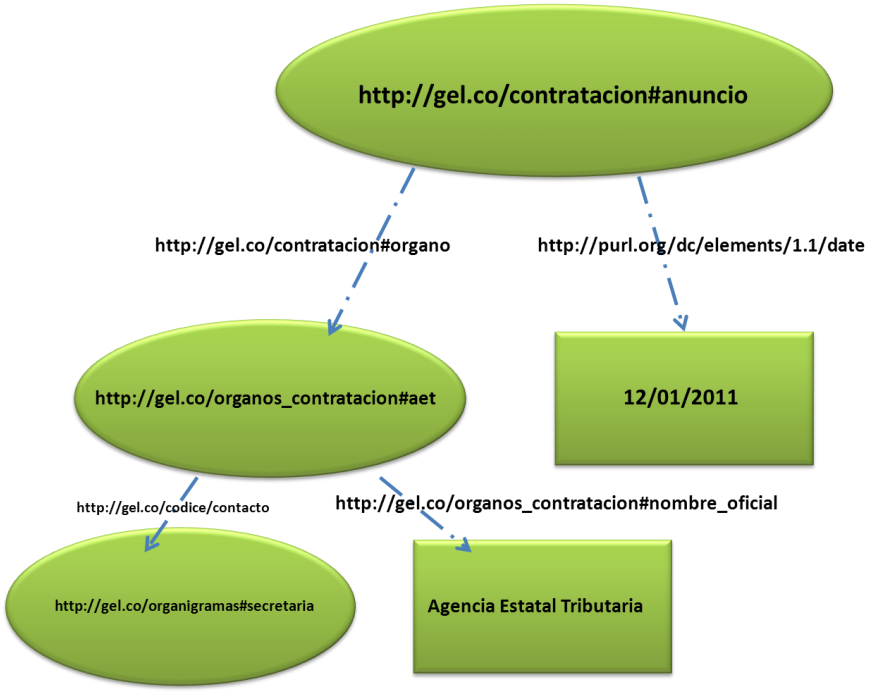
RDF permite expresar modelos conceptuales complejos construyendo "grafos[3]" de tripletas donde Conceptos, Propiedades y Valores se identifican, describen, y tipifican mediante URLs (direcciones únicas en Internet).

[3] Se llama grafo al conjunto de tripletas ya que éstas se organizan jerárquicamente a partir de una primera tripleta que contiene el nodo raíz de un árbol. De él cuelgan ramas u hojas: cada cada nodo del árbol es o bien el concepto objeto de la tripleta anterior (una rama del árbol) o bien un literal (la hoja de un rama del árbol).

Las URLs que identifican conceptos y propiedades definen lo que se conoce como "Espacios de nombres" (*Namespaces*). Realmente, lo que hace un espacio de nombres es delimitar un dominio de negocio. Lo que hace que un concepto sea único y que su significado no pueda ser confundido con el de otro concepto es que ha sido definido en un espacio de nombres concreto. Así, por ejemplo un concepto "anuncio" definido en el espacio de nombres reservado por GEL para la contratación administrativa será diferente del definido en un espacio de nombres donde se manejan conceptos publicitarios comerciales (anuncios televisivos por ejemplo). Y contendrá propiedades con semánticas (significados) diferentes.

Observe la siguiente ilustración:

Figura 2: Ejemplo de grafo RDF sencillo



La siguiente información[4] puede interpretarse como sigue:

- El grafo resultante de unir tripletas pertenece a un anuncio de licitación (nodo raíz: anuncio de licitación. El grafo de la ilustración estaría incompleto, obviamente. Solo representa unos pocos datos del anuncio. El espacio de nombres donde se definió este concepto es "gel.co/contratacion" (URL donde el Gobierno En Línea podría estar definiendo la semántica de los conceptos relacionados con la contratación).
- El anunció se creó en fecha 12 de Enero de 2011. Para describir esta propiedad fecha se ha utilizado un espacio de nombre estándar (Dublin Core, también utilizado por el Lenguaje Común de Intercambio de Información).

[4] Atención, este ejemplo es ficticio, se ha montado únicamente a efectos de ilustrar los conceptos descritos en esta parte del documento.

- El Órgano de Contratación es identificado mediante la URL http://gel.co/organos_contratación#aet, donde se identificarían y describirían los diferentes órganos de contratación.
- Como se ve a continuación "aet" es el identificador de un Órgano de Contratación cuyo nombre oficial es "Agencia Estatal Tributaria" (este nodo es pues un literal, una "hoja" final del árbol).
- La última tripleta por la izquierda reza así "aeat (la Agencia Estatal Tributaria) tiene por contacto a su Secretaría. La elipse que representa al objeto "Secretaría", a su vez, es el sujeto de otros conjuntos de tripletas cuyos objetos podrían ser otros conceptos (como dirección, persona principal de contacto, etc.) o directamente literales (como el número correspondiente a un teléfono de contacto, o una dirección de correo de contacto).

Es importante comprender que **RDF**:

- Por sí solo **no define nada**.
- **Necesita** de un **modelo conceptual** previo sobre el cual organizar las tripletas. El modelo le indica qué conceptos y propiedades existen y cómo se unen entre ellos. En el caso del Lenguaje Común de Intercambio de Información, tanto los conceptos definidos en él como la existencia de un catálogo de documentos ya definidos y usados constituyen este modelo.
- **Necesita** de un **espacio predefinido** para la identificación y descripción de los conceptos y las propiedades del modelo (un **espacio de nombres**) como pertenecientes a un dominio de negocio específico, es decir a un contexto con semántica propia.

La información representada mediante RDF suele escribirse en sintaxis XML. Es por ello que los contenedores por excelencia de datos representados en RDF suelen ser documentos XML o bases de datos con prestaciones XML.

A semejanza de los hipervínculos textuales utilizados en la web actual que permiten a los humanos "navegar" entre documentos distintos, las propiedades expresadas en RDF de un documento pueden apuntar a objetos contenidos en otros documentos RDF. Esto permite que las aplicaciones informáticas que leen documentos RDF vayan "saltando" (navegando) entre documentos distribuidos a lo largo y ancho de internet para la obtención y tratamiento de nuevos datos.

Al hecho (al resultado) de ir relacionando datos ubicados en diferentes documentos se conoce como "Datos Enlazados" ("Linked Data"). Y, en el contexto de los datos de gobierno abiertos, como Datos Abiertos Enlazados ("Linked Open Data").

El objetivo final de ir vinculando datos de fuentes diferentes consiste en acabar generando un nuevo tipo de web, ésta destinada a ser navegada por aplicaciones informáticas. A esta nueva Web se la denomina "la Web de los Datos". Algunos la ven como una inmensa base de datos distribuida.

3. EL LENGUAJE COMÚN DE INTERCAMBIO DE INFORMACIÓN Y DATOS ABIERTOS

En los apartados anteriores se ha dicho que la apertura de datos requiere de:

- Un lenguaje común para la modelación de los datos y
- Un vocabulario estándar para la expresión de los datos estructurados en base al modelo

El Lenguaje Común de Intercambio de Información proporciona un modelo común para todas las Entidades Públicas y pretende convertirse en el vocabulario estándar mediante el cual expresar y transportar la información entre las entidades.

El Lenguaje Común de Intercambio de Información, es el vocabulario diseñado por el programa de Gobierno en Línea del Ministerio de Tecnologías de la Información y las Comunicaciones, que permite la estructuración de los datos y su análisis por parte de aplicaciones informáticas.

"El intercambio de información entre organizaciones requiere de acuerdos entre éstas para definir las estructuras a partir de las cuales se intercambiarán los datos requeridos. Este proceso se hace dispendioso si es necesario llegar a acuerdos entre las diferentes organizaciones, cada una en procesos independientes. Para facilitar la definición de estas estructuras se definió el Lenguaje Común de Intercambio de Información. Este lenguaje brinda un significado y una estructura unificada sobre los datos, facilitando el entendimiento del negocio y el intercambio de información de la organización, facilitando su gestión y su relación con el ciudadano. Así mismo este Lenguaje Común de Intercambio de Información puede ser utilizado por las entidades para obtener la información de los ciudadanos, a través de formularios, o para entregar información estructurada."[5]

También se ha dicho que abrir datos es lograr que las aplicaciones informáticas sean capaces de leer e interpretar los contenidos de un documento.

Por lo tanto, por el hecho de reunir todos los documentos con datos públicos modelados en base al Lenguaje Común de Intercambio, expresados en el formato estructurado y catalogarlos y publicarlos en una sede web central, el Gobierno de Colombia podría declarar que está abriendo datos y su posición en la escala de calidad de apertura de datos sería de 3 estrellas, según la clasificación de Berners-Lee mencionada en apartados anteriores.

Siendo esto posible, y siguiendo el esquema progresivo de las 5 estrellas, en ningún momento debería dejar de publicarse acorde con los lineamientos y formatos definidos por el Lenguaje Común de Intercambio de Información.

El concepto de estandarización, sin embargo, no puede limitarse al ámbito nacional, ya que los datos Colombia son también de interés para el resto del mundo. Y viceversa, las aplicaciones que

[5] **Guía de Uso del Lenguaje Común de Intercambio de Información.** Programa agenda de conectividad estrategia de gobierno en línea. Segunda edición. Enero 2011.

tratan los datos de Colombia, los productos y servicios resultantes de dichas aplicaciones son del interés de Colombia.

Por lo tanto, si se desea abrir al máximo los datos del Gobierno de Colombia, lo recomendable será expresarlos, además, **en formatos lo más estándares posibles**. Dado que al día de hoy a nivel mundial ese estándar es RDF, lo recomendable es expresar la máxima cantidad posible de conjuntos de datos en dicho formato.

Si, además, se siguen las reglas de expresión y de vinculación de datos recomendadas por el Consorcio W3C se alcanzarían los niveles 4 y 5 de la Clasificación Berners-Lee.

Ahora bien,

- ¿Con base en qué modelo? y
- ¿A partir de qué fuente?

Parece más que razonable que si Gobierno en línea ya definió un Lenguaje Común de Intercambio de Información que permite modelar los dominios de negocio de las entidades públicas sea éste el que sirva para fundamentar la expresión de los conjuntos de datos mediante RDF. ¿Para qué inventar uno nuevo?

Por otro lado, la gran cantidad de datos ya expresado en el Lenguaje Común de Intercambio de Información es una fuente excelente para obtener datos RDF mediante transformación de formatos. Si el modelo subyacente en origen y el modelo en destino son el mismo (y aquí es el caso), la transformación de una sintaxis XML (en el origen la representación en XML del Lenguaje Común de Intercambio de Información) a otra (la de RDF-XML, en el destino) no es dificultosa, aunque pueda necesitar de operaciones adicionales para el enriquecimiento de la semántica en el destino. En otras palabras, el Lenguaje Común de Intercambio de Información puede ser un punto de partida para la evolución de los modelos RDF.

Pero, ¿y qué pasa con aquellos datos que nunca fueron expresados en el Lenguaje Común de Intercambio de Información y que se deseará abrir cómo conjuntos de datos RDF?

Existen diversas razones estratégicas y técnicas que justifican que los datos se modelen mediante el Lenguaje Común de Intercambio de Información para que, a continuación, se acaben transformando–automáticamente—en conjuntos de datos RDF. Algunas de ellas son:

1) GEL ha invertido un esfuerzo e inteligencia importantes en la producción de herramientas para la modelación, creación de mensajes (documentos). De cara a abrir datos en XML solo habría que evolucionar dichas herramientas para que una vez creados los mensajes GEL-XML éstos se transformaran además en RDF-XML. Es una evolución económica y muy beneficiosa, pues entre otras cosas tendría menor impacto en la gestión del cambio (sobretodo porque la automatización de las transformaciones haría transparentes las operaciones de conversión entre formatos y la introducción de nuevas tareas relacionadas con la apertura de los datos).

2) GEL diseñó un flujo administrativo para la aceptación y armonización de las nuevas necesidades de modelación y extensión del Lenguaje Común de Intercambio de Información, requerida por las entidades. Este proceso de homologación es el mecanismo para consolidar y normalizar el uso del Lenguaje Común de Intercambio de Información. Si éste lenguaje es consistente y ampliamente usado, los datos que se abran en base a este lenguaje gozarán de la máxima calidad, ya se expresen en el Lenguaje Común de Intercambio de Información, o en su versión transformada RDF. Por lo tanto, queda justificado continuar **impulsando y agilizando** al máximo los mecanismos de armonización ya existentes para los vocabularios de negocio expresados mediante el lenguaje.
3) La necesidad de abrir los datos, un proceso ágil de armonización y normalización del estándar del Lenguaje Común de Intercambio de Información, y la existencia de herramientas de transformación y de transferencia de datos, deberían a su vez alentar cada vez más a las entidades a utilizar el Lenguaje Común de Intercambio de Información como herramienta básica tanto para el intercambio de información entre ellas como para la apertura internacional de sus datos.

4. CONCLUSIONES

1) El esfuerzo realizado hasta la fecha por el Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia en definir un Lenguaje Común de Intercambio de Información, una arquitectura de información y un vocabulario extensible para el intercambio de información entre entidades puede aplicarse a:
    a) La modelación de nuevos conjuntos de datos, en formato GEL-XML
    b) Su transformación a formatos de datos abiertos RDF
    c) Su transferencia a plataformas de datos abiertos
    d) Su vinculación y publicación automatizadas en la Web de los Datos

2) Abrir los datos de la forma más eficaz significa proporcionar los datos en todos los formatos estructurados disponibles, simultáneamente. En otras palabras, se recomienda la publicación de datos en formato GEL-XML y, simultáneamente en formato RDF.

3) Esto requerirá la confección de nuevas herramientas. Estas nuevas herramientas deberían impulsar el **uso creciente** del Lenguaje Común de Intercambio para modelar y expresar los datos que se intercambian entre las entidades, algunos de los cuales tienen por destino ser publicados en la Web de los Datos.

4) En última instancia, la existencia de dichas herramientas **redundaría** en la **normalización del uso** del Lenguaje Común de Intercambio de Información por parte de **las entidades**.

5) En un entorno normalizado, no hay necesidad de descatalogar las buenas prácticas. Más bien al contrario, resulta conveniente evolucionarlas, volverlas más eficientes. Esto aplica especialmente al proceso de tramitación de nuevos requerimientos de modelación y extensión del Lenguaje Común de Intercambio de Información por parte de las entidades.

6) La evolución de la Web de los Datos en un futuro próximo se prevé que conduzca de la Web de los Datos, a la Web Semántica. La Web semántica se basa en los principios y tecnologías que los utilizados por la Web de los Datos y ésta, a su vez, en los datos enlazados (Linked Data).

En consecuencia, no debería haber un futuro acotado para el desuso del Lenguaje Común de Intercambio de Información **como instrumento de homologación y de facilitación** de la apertura de datos.