Submitted on: 17/08/2014

Service-Oriented Architecture for automatic markup of documents. An use case for legal documents

Francisco Adolfo Cifuentes-Silva
Servicios y Sistemas de Información en Red, Biblioteca del Congreso Nacional de Chile, Valparaíso, Chile
E-mail address: fcifuentes@bcn.cl

Abstract:

The problem of information extraction and automatic markup of plain text to XML, has been resolved partially in a specific domain of legal documents. Techniques such as named entity recognition, hierarchy detection of text sections and others has led to partially identify and retrieve different kind of information inside non structured documents.

In this paper we introduce different interconnected components, the NLP techniques used on each component and the workflow needed for processing a plain text document and to generate a new full marked XML version of the document. The generated XML complies with the schema legal standard Akoma-Ntoso and is highly enriched with named entities, semantic URIS, structural sections, lists and elements sequences, between others. As an use case we analyze the experience of the Library of Congress of Chile in the context of the 'History of Law project' and Parliamentary Labor, where these architecture had a key role in order to accomplish the final product and results of processing and marking up different types or models of documents used in the legislative process.

Keywords: Linked Open Data, Semantic Web, Akoma-Ntoso, Machine Learning, e-parliament.

1 Antecedentes

La Biblioteca del Congreso Nacional de Chile, a lo largo de su historia, ha comprometido en su labor los más altos estándares de calidad en la generación de su acervo de productos y servicios. En estos términos, las herramientas tecnológicas han tenido un rol clave, ya que a través de su utilización se ha logrado mejorar procesos productivos permitiendo ampliar la oferta de productos y servicios al congreso y a la ciudadanía de manera continua.

Es bajo este interés constante de utilizar tecnologías innovadoras, que desde hace años se ha desarrollado un importante trabajo en el área de las tecnologías Web, y en particular de la Web Semántica y Open Data, trabajo que actualmente está en pleno proceso de producción.

Para hacer antecedente, podemos distinguir temporalmente tres fases de implantación de la Web Semántica en la BCN:

1. **Fase inicial (2008-2010):** en esta fase la BCN presenta el proyecto Leychile[1], la base de datos de normas legales chilenas. Leychile contiene más de 270.000 normas legales desde 1810 a la fecha asegurando su completitud y textos actualizados publicados en el diario oficial desde 1990 en formato digital (HTML y XML), dejando abierta la base de datos como Open Data a toda la comunidad nacional. Adicionalmente en este periodo, la BCN libera un portal de interoperabilidad[2] donde ofrece una amplia gama de servicios Web gratuitos para la comunidad.

2. **Fase de exploración (2010-2011):** en esta fase la BCN desarrolla la primera ontología en RDF utilizando los datos de normas legales [1] publicados en Leychile. Se publica el primer dataset de normas legales, el portal de datos[3] y a modo de prueba de concepto se desarrollan aplicaciones basadas en tecnologías de Web Semántica como RDF y SPARQL y lo más llamativo, Linked Open Data. La evaluación de esta fase es exitosa y se decide incorporar estas tecnologías al desarrollo de nuevos proyectos.

3. **Fase de explotación (2012-actualidad):** en esta fase se define a nivel técnico la ejecución de diversos proyectos basados en tecnologías de Web Semántica, considerando en ello tanto la generación e incorporación de nuevos datos a la base de datos Linked Open Data pública, como la generación de nuevos productos basados en esta tecnología.

A partir de la fase de explotación, podemos hacer referencia a los proyectos Labor Parlamentaria e Historia de la Ley en su dimensión tecnológica, los cuales explican el dominio de aplicación de la arquitectura que provee marcaje automático.

1.1 Labor Parlamentaria e Historia de La Ley

El proyecto Labor Parlamentaria, desde ahora LP, tiene como objetivo la recopilación automatizada de toda la actividad legislativa realizada por un parlamentario a lo largo de su carrera como tal, y que haya sido registrada en algún medio impreso perteneciente al poder legislativo, tal como un diario de sesión parlamentaria, un informe de comisión, un oficio u otro documento. Para lograr tal fin, es necesario contar con una base de datos altamente granular, en donde figure a nivel de parlamentario la actividad registrada en los distintos documentos en los cuales interviene de una u otra forma.

Si bien un enfoque para dar solución a este objetivo puede ser buscar uno a uno los documentos en los cuales un parlamentario ha intervenido, esto implicaría una muy baja eficiencia a la hora de requerir generar múltiples recopilaciones para distintos parlamentarios. De otra manera, si por cada documento se lograra identificar todas las secciones sobre las cuales algún parlamentario ha intervenido, y posteriormente se guardara esa información de manera que pueda ser consultada cuando se requiera, la solución permitiría optimizar

[1] http://www.leychile.cl
[2] http://llevatelo.bcn.cl
[3] http://datos.bcn.cl

recursos en múltiples dimensiones. Este último ha sido el enfoque que se adoptó para la implementación del proyecto LP. A través de un flujo de trabajo (workflow) en el cual se definen distintos tipos de usuario y participan distintas herramientas, se ingresan y procesan documentos del proceso legislativo en formato texto y se obtienen documentos XML en formato Akoma-Ntoso[4] altamente enriquecidos, los cuales permiten identificar con alto nivel de detalle cada aparición del parlamentario en el contexto de su labor como tal.

El proyecto LP ha tenido como primer objetivo el procesamiento de los documentos generados en ambas cámaras (Cámara de Diputados y Senado) durante las legislaturas comprendidas entre los años 1965 y 1973, habiendo concluido recientemente esta fase de manera exitosa, con lo que en la actualidad se está definiendo el procesamiento de un nuevo periodo legislativo.

Por otro lado, el proyecto Historia de La Ley, desde ahora HL, tiene como objetivo la recopilación automatizada de toda la actividad desarrollada en el poder legislativo asociada a una ley en particular, de manera tal de poder rescatar todas las modificaciones, versiones y debates en torno a un proyecto de ley desde que ingresa al congreso hasta que finalmente es publicado como ley en el diario oficial. El conjunto de todos los documentos relacionados a la tramitación de una ley, desde que se ingresa como proyecto de ley hasta que se publica, lo hemos denominado *bitácora*.

Este sistema toma real importancia en el contexto judicial a la hora de realizar la interpretación de una ley cuando se debe generar un fallo, principalmente porque de no existir antecedentes relacionados con las partes de una norma en uso, los jueces deben recurrir al criterio o interpretación propia.

En términos operativos, el proyecto HL está actualmente en fase de poblamiento. Esto significa que tanto el workflow como las herramientas relacionadas están operativas y se están agregando diariamente nuevas bitácoras y documentos asociados distintas leyes.

1.2 Procesamiento previo

Como se ha mencionado anteriormente, la primera fase de implementación del proyecto LP se ha basado en la incorporación de las legislaturas comprendidas entre 1965 y 1973. Por tratarse de un periodo histórico en donde no se contaba con medios electrónicos de procesamiento y almacenamiento de documentos, no obstante se mantenían los documentos archivados en papel, esta fase del proyecto requirió la implementación de una etapa previa de digitalización en imagen y posterior aplicación de reconocimiento óptico de caracteres (OCR)[5]. Una vez convertido el documento en papel a formato texto, cada documento ingresó a una fase de control de calidad, en donde un equipo de analistas comparó el documento en papel con el documento en texto generado, corrigiendo el contenido en los casos en que el OCR tuvo un bajo nivel de precisión.

Posterior a esta etapa, los documentos estuvieron listos para ser ingresados al workflow de LP. En relación a HL, las leyes procesadas corresponden a normas actuales en donde sí se cuenta con los documentos en formato digital, por lo cual esta etapa no fue necesaria.

[4] http://www.akomantoso.org/
[5] http://es.wikipedia.org/wiki/Reconocimiento_óptico_de_caracteres

1.3 Análisis de los documentos legislativos

Teniendo ya disponibles los documentos de texto listos para procesar, una parte importante del trabajo asociado al marcaje automático ha sido la identificación de las directrices de redacción legislativas [3] asociadas a los distintos tipos de documento procesados.

Sólo por nombrar algunos, dentro de los tipos de documento procesados se encuentran los siguientes:

- **Diarios de sesión**: corresponde al registro fidedigno de los debates en sala asociados a cada sesión de cada cámara, el cual es registrado por funcionarios específicamente capacitados para tal función.

- **Informes de comisión**: corresponden a documentos generados en distintas comisiones parlamentarias con la finalidad de estudiar algún tema específico relacionado con la legislación.

- **Oficios**: corresponden a documentos generados por diversos entes de la contingencia política (personas u organismos) los cuales tratan algún tema particular que sirve de antecedente tanto en LP como en HL.

Afortunadamente, dentro de un mismo periodo, los documentos presentan regularidad en el uso de normas de redacción, particularmente en estructura, estilo y nombres de secciones. Esto ha sido de gran apoyo en la definición de estrategias de recuperación e identificación de secciones de texto para posterior marcaje automático.

Habiendo puesto en antecedente todos estos puntos, procederemos a definir la arquitectura orientada a servicios asociada al marcaje automático.

2 Arquitectura TI de la solución

El contexto del marcaje automático se da dentro de una de las fases del workflow de producción de HL y LP. Nuestro workflow ha sido implementado basado en una arquitectura orientada a servicios sobre tecnologías Web, en donde la interoperabilidad, la alta cohesión y bajo acoplamiento han sido claves para la construcción y extensión del entorno.

Desde una perspectiva técnica, una decisión acertada fue la adopción del protocolo REST[6] para la definición de las políticas de interoperabilidad. Aprovechando la simplicidad de uso que este mecanismo provee, hemos integrado la mayor parte, si es que no el total de nuestras herramientas y servicios de apoyo tanto en LP como en HL. Es el caso de todas las componentes relacionadas al marcaje automático, en donde el mecanismo de interoperabilidad definido ha sido REST, solo considerando como excepción a nuestro endpoint SPARQL, el cual es un servicio Web HTTP.

En términos del proceso de marcaje automático, éste permite transformar un documento de texto plano en un documento XML altamente enriquecido en formato Akoma-Ntoso. Por cada tipo de documento a marcar, se ha diseñado un esquema XSD intermedio definido por sus secciones específicas. Muchas de estas secciones son muy similares entre sí, por lo

[6] http://es.wikipedia.org/wiki/Representational_State_Transfer

que en muchos casos la lógica de implementación es similar. Por ejemplo, algunas secciones con lógica reutilizable son la portada del documento, el índice o secciones en donde se declara una lista de asistentes, la cual se denomina *asistencia*.

La idea de generar una representación intermedia basada en un esquema que no es Akoma-Ntoso, se justifica en la premisa de facilitar al máximo la implementación de las componentes. Esto es principalmente porque el esquema Akoma-Ntoso es un esquema XSD altamente complejo, por lo que realizar una traducción directa implicaría un esfuerzo estimado mucho mayor.

Otra de las razones del por qué definimos usar un esquema intermedio es que aunque nuestro proceso de marcaje hoy en día se centra en obtener documentos Akoma-Ntoso, el procedimiento no está limitado a este esquema ya que permite extender la funcionalidad a nuevos esquemas de manera ilimitada. Esto nos permitiría en la eventualidad, la utilización de nuevos estándares asociados a distintos contextos de uso.

3 Componentes de la arquitectura

A continuación se describen los principales componentes del proceso de marcaje automático que permitirán obtener un documento XML enriquecido, sin embargo, de acuerdo a la naturaleza de los distintos tipos de documento a procesar, es que existen distintos flujos de aplicación de estas componentes.

Vale decir que cada uno de los componentes mencionados a continuación está implementado mediante un servicio Web específico que cumple la funcionalidad de manera atómica. Un diagrama que da cuenta del flujo tradicional asociado al procesamiento del diario de sesiones se presenta en la imagen 1.

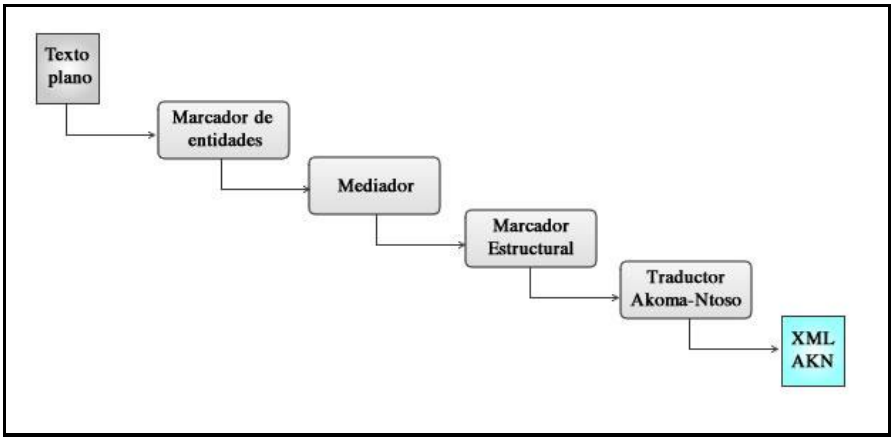

Imagen 1: Flujo de ejecución para el marcaje automático

3.1.1 Marcador de entidades

Este componente, también denominado NER por sus siglas de *Named-Entity Recognition* realiza el reconocimiento de entidades nombradas en el texto tales como: personas, organismos, documentos, fechas, eventos, roles o localidades, dentro de otros. Para esto, se

ha utilizado una implementación basada en el Stanford NER[7], el cual utiliza como técnica principal un clasificador CRF[8]. Este componente se ha adaptado al lenguaje castellano y se han implementado mecanismos de entrenamiento asociado a dominios acotados. Como resultado de su ejecución, este componente entrega una nueva versión del documento en pseudo XML (ya que carece de un nodo raíz), en donde se identifican mediante etiquetas aquellas palabras o conjunto de palabras asociadas a una entidad particular. La imagen 2 muestra un trozo de texto marcado por esta herramienta.

```
 PROYECTOS DE LEY violencIa que ha golpeado a las familias El
▼<entity>
   <body bestClass="ROL" knownFraction="1">Secretario</body>
   <class value="ROL" probability="0.2201459321"/>
 </entity>
 , señor
▼<entity>
   <body bestClass="PER" knownFraction="1">Raúl Guerrero Guerrero</body>
   <class value="PER" probability="0.83023132"/>
 </entity>
 y ei rrosecretario,
▼<entity>
   <body bestClass="PER" knownFraction="0.5">don Fernando Parga Santelices</body>
   <class value="PER" probability="0.9555160392"/>
 </entity>
 .
```

Imagen 2: Texto marcado con el marcador de entidades

3.1.2 Marcador estructural

El objetivo de este componente es segmentar el texto plano o texto con entidades en bloques estructurales del documento, tales como páginas, capítulos, secciones del índice y también sub secciones, definiendo tantos niveles de marcaje estructural como sea necesario. Para la implementación de esta fase se hace clave el análisis estructural de los distintos tipos de documento, debido a que de esta manera se han podido identificar los diversos patrones que comparten los documentos de un mismo tipo. Desde el punto de vista técnico, se han utilizado básicamente dos mecanismos que proveen un alto nivel de efectividad (sobre un 99% en dominio cerrado), los cuales se han utilizado tanto por separado como en combinación en distintos tipos de documento.

El primer mecanismo se basa en la definición de un esquema XSD del documento, denominado esquema de marcaje previo. Por cada elemento definido en este esquema de marcaje previo se implementan una o más reglas que van a permitir identificar el elemento. Cada regla está compuesta a su vez de un conjunto variable de expresiones regulares de segmentación y de otro conjunto de expresiones regulares de identificación. Adicionalmente, cada regla tiene asociado uno o varios algoritmos que permiten aplicar estas expresiones regulares en combinación, permitiendo identificar unívocamente una sección.

El segundo mecanismo mencionado, se basa en el descubrimiento de secuencias de elementos consecutivos o listas ordenadas de elementos. Para ello también se identifican los símbolos de enumeración que pueden ser secuencias de números, letras, números romanos, bajo todas sus variantes, y posteriormente se aplica un algoritmo que permite identificar el árbol de

[7] http://nlp.stanford.edu/software/CRF-NER.shtml
[8] http://en.wikipedia.org/wiki/Conditional_random_field

secuencias que define la estructura con mayor encaje dentro de los elementos definidos.

Durante el proceso, sea cual sea el mecanismo de reconocimiento, el marcador estructural genera una representación interna en objetos que posteriormente es traducida a XML y validada bajo el esquema de marcaje previo, entregando como resultado el XML estructural del documento.

3.1.3 Mediador

El objetivo de este componente es permitir identificar unívocamente cada entidad reconocida por el NER asignando una URI a cada elemento. Para ello, el mediador se conecta con la base de datos RDF a través del endpoint SPARQL que esta provee, lugar de donde obtiene distintas listas de pares URI/Etiqueta asociadas a los distintos tipos de entidad. Mediante la aplicación de algoritmos de comparación de cadenas de caracteres, el mediador es capaz de establecer con alta probabilidad la similitud entre una entidad definida en el texto con una etiqueta formal de un recurso RDF existente en la base de datos, permitiendo identificar de forma unívoca lo declarado en el texto. De esta manera, si por ejemplo en el texto el marcador de entidades identifica un nombre de persona como entidad persona, el mediador irá a buscar a la base de datos la URI de la persona que tenga el nombre más similar a las palabras reconocidas en el texto. Si bien, esto permite un nivel de acierto aceptable, en algunos casos no es suficiente del todo ya que existen etiquetas se repiten en distintos contextos. A partir de esto, y como mecanismo de apoyo a la generación de aciertos, es que para cada ejecución del mediador se definen parámetros de inicialización denominados "Contexto". Con estos parámetros, el mediador es capaz de aplicar distintas heurísticas asociadas a cada tipo de entidad, teniendo como dato el contexto en el que se está reconociendo una entidad particular. Un ejemplo de dato de contexto puede ser un año asociado al espacio temporal de aplicación. Bajo el supuesto de que tanto un padre como un hijo tengan el mismo nombre y apellido, y que además sean políticos, el mediador podrá, gracias a este dato adicional, identificar de quién se está hablando asignando la URI correcta al marcaje.

3.1.4 Traductor Akoma-Ntoso

Este componente permite generar un documento XML marcado en un esquema ad-hoc en un documento XML marcado en el esquema Akoma-Ntoso. Si bien se podría pensar que realizar la transformación a Akoma-Ntoso podría ser resuelta a través de hojas de transformación XSLT, nuestra opinión es que tal tarea resulta demasiado compleja de implementar usando sólo esta tecnología, elevando la complejidad de la hoja de transformación hasta hacerla inmantenible. Por esta razón, optamos por la implementación de una aplicación de software que integra distintos métodos de programación que apoyan la traducción XML (tales como DOM, XSLT, cadenas de caracteres y POO), facilitando la implementación, extensión y mantención del módulo acorde a los distintos requerimientos.

Akoma-Ntoso ha sido seleccionado para el marcaje de documentos legislativos principalmente por ser un esquema desarrollado de manera mancomunada entre distintos organismos internacionales bajo el alero del grupo OASIS[9] (dentro de los cuales se encuentra la BCN como miembro activo), permitiendo recoger las mejores prácticas de la industria informática, y las distintas visiones y requerimientos aportados por cada uno de ellos, produciendo un estándar de alta calidad.

[9] https://www.oasis-open.org/

3.1.5 Triplestore RDF

Este componente corresponde a la base de datos Linked Open Data, la cual almacena triples RDF que son de dominio público a través de nuestro endpoint SPARQL[10]. Esta contiene el total de datos de base para la producción de los documentos XML como también el total de datos producidos por el proceso posterior de extracción de datos desde los documentos XML marcados. Como ejemplo de los datos de base para la producción podemos nombrar la información de parlamentarios o de localidades geográficas, información que es utilizada cada vez que en un documento XML se hace referencia a alguna de estas entidades.

Desde la perspectiva del contenido del triplestore RDF, se pueden identificar dos dimensiones almacenadas:

A. **Modelos**: básicamente este conjunto está formado por las distintas ontologías y vocabularios utilizados para estructurar y articular la información almacenada, como también para describir algunos aspectos del negocio, los que en su mayoría son usados en diversas aplicaciones. Un ejemplo de ello es la ontología de Biografías Parlamentarias[11], la cual describe la forma en que se estructura y relaciona la información básica de los parlamentarios, los cargos que éstos ejercen, los periodos temporales asociados a un cargo y los organismos a los cuales estas entidades se interconectan.

   Finalmente, vale decir que en su mayoría estas ontologías han sido diseñadas por la BCN tomando como premisa la reutilización y extensión de vocabularios existentes, tal como lo indican las mejores prácticas [2].

B. **Datos**: este conjunto corresponde a los distintos datasets que se han recopilado desde diversas fuentes, tales como los listados de parlamentarios, cargos, organismos del estado, información acerca de las sesiones parlamentarias y legislaturas entre muchas otras; como también al gran volumen de datos generado a diario por sistemas como Leychile, del cual diariamente se obtiene la información sobre las normas legales publicadas, o los mismos sistemas HL y LP.

   Evidentemente, cada dataset incorporado en el triplestore RDF ha debido ser previamente procesado, estructurado y transformado a los modelos descritos anteriormente previo a su carga.

4 Contexto de uso

Fuera del contexto del marcaje automático encontramos otros componentes de la arquitectura que también son dignos de mención ya que van a proveer parte de la lógica necesaria para la generación de productos. Dentro de éstos, los más importantes se describen brevemente a continuación.

[10] http://datos.bcn.cl/sparql
[11] http://datos.bcn.cl/ontologies/bcn-biographies/doc/

- **Servicios Web de operación**: un conjunto de servicios Web que permiten al workflow la ejecución de operaciones de negocio y persistencia en el triplestore RDF y otros servicios externos.

- **Editor XML**: herramienta Web de edición del XML marcado automáticamente, mediante la cual los usuarios realizan el control de calidad y enriquecimiento de los textos ingresados al workflow.

- **Servicio GRDDL**[12]: también llamado servicio de publicación, permite extraer desde los documentos XML Akoma-Ntoso, triples RDF, las cuales son agregadas al triplestore RDF.

- **Visualización y búsqueda**: aplicaciones mediante las cuales la información generada se presenta al usuario. Dentro de estas también se encuentran los productos finales de LP y HL.

5 Conclusión

La utilización de tecnologías de Web Semántica y en particular de una arquitectura orientada a servicios han tenido un positivo efecto en el desarrollo de los proyectos LP y HL. En este sentido, tareas como el marcaje en XML de los documentos se han logrado optimizar reduciendo los tiempos de procesamiento considerablemente, y consecuentemente, se han logrado mejorar considerablemente los tiempos de generación de productos. Para hacernos una idea, se registra que en el pasado, el tiempo necesario para generar una historia de la ley de forma manual por un analista legislativo estaba en torno a las 45 horas de trabajo. En nuestro escenario basado en tecnologías de Web Semántica este tiempo se reduce a un máximo de 2,5 horas, de las cuales el tiempo necesario para marcaje automático en XML se reduce a segundos.

6 Trabajo futuro

Chile es un país en el cual la estructura del poder legislativo se define por un Congreso Nacional bicameral, en donde existe la Cámara de Diputados (o también denominada cámara baja) y el Senado (o también denominada cámara alta). Asimismo, conforme a la Ley Orgánica del Congreso Nacional[13], la Biblioteca del Congreso se define como un órgano común que presta servicio a las dos corporaciones, haciendo parte de la estructura del parlamento. En un esfuerzo por disponibilizar y transparentar la información generada diariamente, ambas cámaras más la BCN han recientemente inaugurado un nuevo portal de datos abiertos del Congreso Nacional[14] en donde, dentro de otra información, se está brindando acceso público a los diarios de sesiones actuales de la Cámara de Diputados y el Senado en formato XML.

En este sentido y de acuerdo a su disponibilidad, la BCN proyecta en un futuro cercano la incorporación de los documentos publicados en el portal de datos abiertos del Congreso al workflow de HL y LP con el fin de comenzar a trabajar con un insumo de datos de mayor

[12] http://www.w3.org/TR/grddl/
[13] Ley N° 18.918
[14] http://opendata.congreso.cl

calidad, esto ya que el formato publicado es perfectamente adaptable al esquema utilizado en la BCN, y en consecuencia permite un nivel de recuperación de información similar.

A nivel técnico, el flujo de procesamiento para marcaje automático comenzará en la etapa de traducción Akoma-Ntoso, y a partir de este punto se trabajará en un flujo de proceso levemente diferente aunque con los mismos componentes definidos en las etapas anteriores.

En otro contexto, la BCN actualmente está en plena fase de formulación de nuevos proyectos asociados al uso de marcaje de documentos en sus diversas áreas de trabajo, con la idea de dejar disponible la información generada en el trabajo diario de la institución en forma de Linked Open Data. La aplicación de esta idea se fundamenta en dos líneas base: la disponibilización de los datos para la generación interna de nuevos productos y servicios tanto para la comunidad parlamentaria como para la nación; así como dejar disponible a la ciudadanía el mayor volumen de datos posibles en pos de la generación de valor, de manera tal que la BCN pueda ser vista en su faceta de Biblioteca Semántica como un real aporte a la sociedad nacional e internacional.

Agradecimientos

El autor agradece a la Biblioteca del Congreso Nacional y a sus autoridades la oportunidad para presentar este artículo, así como también agradece el valioso aporte de todos quienes han contribuido en el desarrollo del proyecto como en la revisión y mejoramiento de este.

Referencias

[1] Cifuentes-Silva F., Sifaqui C. and Labra-Gayo J. Towards an architecture and adoption process for linked data technologies in open government contexts: a case study for the Library of Congress of Chile. I-Semantics 2011

[2] Hyland B, Atemezing G., Villazón-Terrazas B. Bests practices for Publishing Linked Data. Enero 2014.

[3] Palmirani M. XML Legislativo: Principios e instrumentos técnicos. Oct. 2012