Construcción de un Catálogo Semántico para los Datos del Estado

Antecedentes

AGESIC ya construyó un Catálogo de Datos Abiertos[1] en donde se publica información sobre diferentes conjuntos de datos o *datasets* en diferentes formatos, típicamente tabulares (CSV,XLS,etc). Cada *dataset* tiene un conjunto de información asociada o **metadatos del dataset** (ej: la fecha de publicación, el organismo que lo publica y otros aspectos generales). Para algunos *datasets* se publica otro archivo que tiene una descripción de los campos o **metadatos de contenido**, pero esta información no siempre está disponible, o no está en formatos que permitan el procesamiento de su semántica por máquinas. Por ejemplo, los datos correspondientes a la encuesta continua de hogares del INE[2] tienen publicada la descripción de los campos por separado de los datos en una planilla de cálculo. Otros están publicados en XML aunque sin referencias conocidas a un XML Schema o DTD (ej: datos del Sistema de Precios al Consumidor[3]). En este último caso, es posible inferir bastante bien la estructura analizando una instancia de los datos pero de todos modos no existe información clara que permita relacionar un *dataset* con otro.

Descripción del Proyecto

El objetivo del proyecto es desarrollar estategias para facilitar la publicación de metadata de contenido. En esta información se debería incluir, entre otros:

- Categorías (clases) de objetos de los cuales se publica información. (Ej. Productos, Personas, Ubicación geográfica, etc)
- Nivel de Agregación de la información (Ej. Los datos del INE sobre personas son siempre agregados en función de diferentes dimensiones)
- Descripción de las clases y los atributos publicados.
- Relaciones entre diferentes clases y/o Propiedades. (Ej: La clase Pacientes que relacionada con los datos de Salud Pública es una subclase de la clase Personas relacionada con los datos del INE).

El desarrollo de estas estrategias consiste en al menos:

- Estudiar vocabularios y/o ontologías existentes para definir la información de los Datasets. Estudiar particularmente, el caso de la información que provee CKAN.
- Definir un vocabulario adecuado para representar la metadata de contenido de los Datasets. Este vocabulario debería ser una extensión de alguno compatible con la información que provee CKAN.
- Definir una ontología en OWL 2 que permita la descripción de la metadata de contenido de los datasets. Se debe tener en cuenta tanto la expresividad de la ontología como la computabilidad de la misma.
- Definir procedimientos adecuados que permitan obtener la metadata de contenido de los datasets con (relativamente) poco trabajo de parte del mantenedor de la información.
- Definir procedimientos adecuados para la publicación y explotación de esta metadata.

Alcance y Resultados

El alcance del proyecto estará regulado por los resultados intermedios que se obtengan. Sin embargo, es

1 http://www.catalogodatos.gub.uy

2 https://catalogodatos.gub.uy/dataset/ech2010

3 https://catalogodatos.gub.uy/es/dataset/precios

fundamental contar con uno o más informes que justifiquen la toma de decisiones.

También se espera disponer de:

- La descripción de un procedimiento, puede ser automático o guiado por el usuario, construya el la información de contenido dado el dataset como entrada.
- La descripción de un procedimiento, puede ser automático o guiado por el usuario, que dada la descripción de contenido de un dataset describe cómo se agrega esa descripción al catálogo.
- Una prueba de concepto de cada procedimiento.

Algunas preguntas que pueden quedar. (FAQ)

- ***Por qué es importante disponer de la metadata de contenido de los datasets?***

  Esta metadata permitiría:

  - Facilitar la búsqueda de los datasets que tienen información sobre determinados elementos de la realidad. (Ej: Información de personas está publlicada por el INE y Salud Pública). Puede parecer que este aspecto está cubierto con las "categorías de información" que maneja CKAN, sin embargo no es así. Las categorías de CKAN constituyen grandes areas pero no brindan información detallada (al menos como están usadas ahora) de las clases de objetos representados en los datasets.
  - Facilitar la combinación de datasets. Con esta metadata, debería facilitarse el trabajo de cruzar diferentes datasets publicados, eventualemnte, por organismos distintos. (Ej: cruzar la información de las enfermedades de información obligatoria de Salud Publica con datos socioeconómicos provistos por la encuesta de hogares del INE).

- ***Por qué definir un vocabulario y una ontología? No son básicamente lo mismo?***

  - Un ***vocabulario***, es sólamente un conjunto de términos con un uso definido para cada término. Ese uso está dado de diferentes formas. Algunos están dados como descripciones textuales, otros a través de una descripción en OWL 2, otros usando XMLSchema.
  - Una ***ontología***, se puede pensar como una estructura lógica similar a una estructura de primer orden o como un esquema de una base de datos. Allí se definen conjuntos de datos (en términos lógicos, predicados unarios) y relaciones entre esos predicados (en términos lógicos, predicados n-arios con n > 1) con restricciones de integridad entre todos ellos que regulan que instancias son válidas (en términos lógicos, modelos) y que instancias no lo son. Una ontología, induce un vocabulario. Pero la ontología me define la forma en debo interpretar ese vocabulario y cómo debo razonar con él.