Reporte de caso

REVISTA Ingeniería

Una Mirada a la Web de los Datos. Caso de Estudio: Consumo de Servicios CKAN

A View of the Web of Data. Case Study: Use of Services CKAN

Jhon Francined Herrera-Cubides[1] **,** ***Paulo Alonso Gaona-García***[*1] **,** ***Kevin Gordillo-Orjuela***[1]

[1]Facultad de Ingeniería de la Universidad Distrital Francisco José de Caldas. Bogotá-Colombia
Correspondencia: pagaonag@udistrital.edu.co

Recibido: 05/07/2016. Modificado: 05/09/2016 Aceptado: 12/10/2016

Resumen

Contexto: Se busca llevar a cabo el análisis, conexión y uso de los servicios ofrecidos por Comprehensive Knowledge Archive Network (CKAN, por sus siglas en inglés), con el fin de evaluar criterios base para obtener referentes preliminares de estudio sobre el estado de la web de los datos, a través de la exploración y acceso de los dataset publicados en el repositorio de datos abiertos DataHub.io.

Método: Empleamos los servicios ofrecidos por CKAN para la consulta y descarga de los dataset publicados en Datahub.io, para lo cual presentamos una serie de procesos llevados a cabo para analizar los datos descargados. La propuesta se compone de tres actividades clave: (1) revisión y análisis de las plataformas; (2) configuración y uso de los servicios prestados por la API; y (3) descarga y revisión de la información obtenida.

Resultados: Se configuraron y desplegaron los servicios requeridos, a través de CKAN, con el fin de llevar a cabo las consultas y descargas respectivas de dataset. Se procesó y analizó la información obtenida de los JSON descargados, permitiendo hacer un análisis preliminar comparativo, de la información obtenida acerca del comportamiento de la web de los datos.

Conclusiones: CKAN es una herramienta potente para gestionar catálogos de datos, permitiendo manejar una descripción de los datos y otra información relevante, tanto para las organizaciones que publican como para las personas que consultan dicha información, tales como categorías de organizaciones, formatos en que se encuentra disponible los datos, propietario de los datos, el tipo de licenciamiento de las publicaciones, enlaces a otros datos, entre otros datos, pertinentes para llevar a cabo un análisis de la web de los datos.

Palabras clave: CKAN, Datahub, JSON, Linked Data, web de datos.

Idioma: Español

Citación: J. F. Herrera, P. A. Gaona, K. Gordillo, "Una Mirada a la Web de los Datos. Caso de Estudio: Consumo de Servicios CKAN" INGENIERÍA, vol. 22, no. 1, pp. 46-64, 2017.

http://dx.doi.org/10.14483/udistrital.jour.reving.2017.1.a07

Abstract

Context: In order to assess basic criteria so as to obtain preliminary guidelines on the current state of the Web of Data, we analyze the connection and use of the services offered by CKAN - Comprehensive Knowledge Archive Network; the analysis is conducted through exploration and connection to datasets published in the datahub.io open data repository.

Method: We use the services offered by CKAN for consultation and downloading datasets published in Datahub.io, we propose a procedure carried out to analyze the downloaded data. The proposal consists of three key activities: (1) review and analysis of platform, (2) Setting up and using the services provided by the API and (3) download and review of the information obtained.

Results: The required services offered by the platform CKAN were configured and deployed, in order to carry out queries and downloads related to each dataset. The obtained information was processed and analyzed from the downloaded JSON, allowing a comparative preliminary analysis of the information regarding the behavior of the Web of Data.

Conclusions: CKAN is a powerful tool to manage data catalogs. This tool can handle a description of the data and other relevant information, from organizations that publish to people who query such information. These queries provide information as categories of organizations, data formats and owners, the type of publication licenses, links to other data, among other which are relevant to perform an analysis of the Web data.

Keywords: CKAN, Web of Data, DataHub, JSON, Linked Data.

1. Introducción

La iniciativa propuesta por Tim Berners Lee [1] asociados a la vinculación de datos mediante Linked Data promete resolver los problemas asociados con el análisis y la interoperabilidad de los datos vinculados a recursos relacionados. En la actualidad existen varios ejemplos que permiten el uso de datos vinculados en diferentes áreas de conocimiento tales como la medicina, la geografía, la bibliotecología, entre otras. Por ejemplo, en el campo de la educación, la vinculación de los recursos educativos de diferentes repositorios de conocimiento es útil en la Web, permite el intercambio, la búsqueda y la navegación de objetos de aprendizaje [2]. Varios proyectos como Europeana [3], LinkedUp [4], [5], y Linked Education [2], han adoptado el enfoque de Linked Data y su objetivo responde a vincular conjuntos de datos educativos.

Por su parte, el proyecto Linked Open Data (LOD, por sus siglas en inglés) es otra iniciativa orientada a aplicar los principios de la web semántica con el fin de conectar recursos. LOD utiliza tecnologías como Resource Description Framework (RDF, por sus siglas en inglés) [6] y Uniform Resource Identifier (URI, por sus siglas en inglés), junto con un conjunto de principios denominados “datos vinculados”. Varias fuentes de datos, como Wikipedia, están ahora disponibles para los desarrolladores, que claramente se benefician de los conjuntos de datos vinculados con base en un modelo de datos común [3].

Datahub, sujeto de análisis en el presente artículo, es una plataforma para la gestión de datos, ba-

sada en CKAN - Comprehensive Knowledge Archive Network [17] - [18], que es una herramienta para la gestión y publicación de colecciones de datos (Datasets) en un ambiente web, utilizado por diferentes gobiernos, nacionales y locales, instituciones de investigación, entre otras (denominadas "organizaciones" en CKAN), que recogen una gran cantidad de datos. A través de sus servicios, los usuarios pueden buscar y encontrar los datos que necesitan.

La motivación de este trabajo es presentar, a través de un estudio de caso basado en Datahub, cómo a partir de la conexión y el consumo de datos que pueda ofrecer el Application Programming Interface (API, por sus siglas en inglés), en este caso CKAN [17], se pueden obtener variables para la realización de un análisis preliminar sobre la visión de la web de los datos. Dicho caso estudio busca propiciar herramientas, para la identificación de tendencias en el modelo de la web de los datos, tomando como base datos históricos presentes en Datahub.io, como plataforma de gestión de datos, en el marco del proyecto de investigación sobre vinculación de datos que se viene adelantando.

El resto del siguiente artículo se encuentra organizado de la siguiente manera: en la primera parte de se encuentra el estado del arte, donde se revisan los fundamentos de la temática explorada. Posteriormente, en la sección III se presenta el planteamiento metodológico utilizado para la exploración de los conjuntos de datos. En la sección IV, se presenta el desarrollo metodológico, junto con los hallazgos encontrados, en la sección V se presenta el análisis de resultados. Finalmente, en la sección VI se presentan las conclusiones y trabajo futuros.

2. Estado del arte

Gracias al modelo de web de los datos, varias fuentes de datos, como Wikipedia, están ahora disponibles para los desarrolladores, que claramente se benefician de los conjuntos de datos vinculados con base en un modelo de datos común [7]. Por su parte, el proyecto DBpedia [8] dispone de un conjunto de datos que actualmente se consideran como el eje central y más significativo entre los conjuntos de datos LOD [9]. En la web se encuentran iniciativas LOD tales como las plataformas de gestión de datos, entre ellas:

- Datahub (https://datahub.io/), objeto de examen en este artículo, la cual permite buscar y publicar datos, crear y gestionar grupos de Dataset, entre otras funcionalidades [10].
- Junar (http://junar.com/), plataforma de datos abiertos en la nube que facilita la publicación de datos por parte de gobiernos, empresas u otras organizaciones [11].
- Socrata (https://www.socrata.com/), plataforma escalable de publicación de datos en la nube que facilita la creación de iniciativas de datos abiertos sostenibles, ofreciendo un amplio conjunto de funcionalidades [11].
- data.gov.uk To Go (http://guidance.data.gov.uk/), kit del gobierno de Reino Unido para poner a disposición del público en general, su plataforma de publicación de datos de forma que cualquiera pueda desplegar una plataforma similar, preocupándose únicamente de adaptar la apariencia externa final [11].

- Plataforma de Gobierno Abierto OGoov (http://www.ogoov.com/es/), ofrece una serie de funcionalidades combinables entre si según la orientación o iniciativas relacionadas con el gobierno abierto que se deseen priorizar: datos abiertos, transparencia y participación [11].

Estas entre otras plataformas, han permitido una mayor visibilidad de datos compartidos y han facilitado la participación de iniciativas como LOD Cloud [9], para la visualización de las organizaciones y proveedores de contenidos que han liberado y enlazado sus datos.

Como parte de un proyecto de investigación de tesis doctoral en curso, orientado en el dominio de la vinculación de datos, se plantea recabar información preliminar sobre el contenido de la nube LOD a 2016. Con tal fin, y dado que en experiencias como [9] donde se trabajó con DBPedia, para el presente estudio se propone analizar la plataforma de gestión de datos libres Datahub.io, de la Open Knowledge Foundation [12] – OKFN. Esta fundación tiene como visión que el conocimiento crea poder para muchos, no para unos pocos, los datos nos libera para tomar decisiones informadas sobre la forma en que vivimos, lo que compramos y quien recibe nuestro voto; la información y el conocimiento son accesibles, aparentemente, a todo el mundo. Por otra parte, Datahub.io, como una de las plataformas de gestión de datos y repositorio internacional de datos abiertos reconocidas [25], es una de las que más Dataset aporta a la conformación del Linked Open Data Cloud Diagram [9].

CKAN maneja un backend construido en Python, y un frontend construido en JavaScript. También usa el framework web Pylons y SQLAlchemy como ORM, con PostgreSQL como motor de base de datos. Tiene una arquitectura modular que permite desarrollar extensiones para proporcionar características adicionales, tales como harvesting o carga de datos, visualización de múltiples vistas, diferentes extensiones, un JSON API para leer, escribir y hacer consultas a los Dataset, soportado en más de 40 lenguajes, entre otras funcionalidades [23] - [24].

Al ser de código abierto, licenciado bajo términos GPL v3.0 Affero GNU, los usuarios CKAN pueden adaptar sus servicios, como lo han hecho, entre otros [13]:

- Africa's Largest Volunteer Driven Open Data Platform (https://africaopendata.org/).
- data.amsterdam.nl (http://data.amsterdam.nl/).
- Buenos Aires Data (http://data.buenosaires.gob.ar/).
- Paraguay Digital (https://www.datos.gov.py/).
- La plataforma cívica de datos abiertos de México (http://datamx.io/).
- Registro de conjuntos de datos abiertos en Noruega (http://data.norge.no/).

Ahora bien, para los propósitos de CKAN, los datos se publican en unidades denominadas "DataSet". Un Dataset es un paquete de datos —por ejemplo, podría ser las estadísticas de la delincuencia para una región, las cifras de gasto para un departamento gubernamental o las lecturas de temperatura de varias estaciones meteorológicas—. Cuando los usuarios buscan datos, los resultados de búsqueda que se obtiene son Dataset individuales. Un Dataset contiene dos componentes [13]:

- Información o ”metadatos”sobre los datos. Estos datos deben proveer la siguiente información:

  a) Título, único a través de CKAN, de modo que sea breve pero específica. Por ejemplo: ”densidad de población del Reino Unido según la región.[es] mejor que ”Las cifras de población”.
  b) Descripción, información que la gente necesita saber cuándo se utilizan los datos.
  c) Etiquetas, etiquetas que ayuden a la gente a encontrar los datos y la vinculan con otros datos relacionados.
  d) Licencia, información de la licencia para que se sepa cómo se pueden utilizar los datos.
  e) Organización, elegir quien es el propietario del Dataset.

- Un número de ”recursos” , que contienen los datos en sí. CKAN no le importa en qué formato están los datos (hoja de cálculo CSV (comma-separated values) o Excel, XML, PDF, archivo de imagen, RDF, entre otros). CKAN puede almacenar el recurso internamente o almacenar un enlace al recurso en sí, ubicado en otra parte en la web. Para los recursos se debe proveer la siguiente información:

  a) Nombre, un nombre para el recurso.
  b) Descripción, una breve descripción del recurso.
  c) Formato, el formato de archivo del recurso, por ejemplo, CSV, XLS, JSON (JavaScript Object Notation), PDF, etc.
  d) Visibilidad, un Dataset público puede ser visto por cualquier usuario del sitio. Un Dataset privado solo puede ser visto por los miembros de la organización propietaria del Dataset y no se mostrará en las búsquedas realizadas por otros usuarios.
  e) Autor, el nombre de la persona u organización responsable de la producción de los datos.
  f) Correo electrónico del autor.
  g) Correo electrónico del responsable de mantenimiento.
  h) Campos personalizados, si desea adicionar más datos.

Considerando lo anteriormente descrito, en Datahub.io, se registra 822 organizaciones (propietarios de los Dataset), creadas con previa autorización de un administrador, a través del envío de una solicitud. Estas organizaciones pueden crear, administrar y publicar colecciones de conjuntos de datos y pueden tener miembros tales como administradores (que añaden usuarios y gestionan la organización), y editores (que solo pueden añadir conjuntos de datos de la organización). Dicha distribución en organizaciones provee las siguientes características:

- Se proporciona una estructura de permisos y autorizaciones mucho más rica (en torno a la organización), que ofrece a los usuarios un mayor control sobre quién puede o no, editar y añadir conjuntos de datos.

- Proporciona una estructura orientada a la organización, para la presentación y la búsqueda de bases de datos.

- Ayuda a controlar problemas de spam, proporcionando más control sobre quién añade conjuntos de datos.

Dentro de las organizaciones incluidas en Datahub.io se encuentran las que se presentan en la tabla I, junto con el total de Dataset publicados por cada una de ellas.

Tabla I. Cantidad de dataset publicados por Organización [10]

Organizacion	Cantidad	%	Organizacion	Cantidad	%
Global	3418	36,8	Opendata.cz	46	0,5
London Datastore Archive	612	6,6	Civil Society	46	0,5
Paises	1163	12,5	BudgIT Information Technology	43	0,5
Open Hampton Roads	531	5,7	Sport	38	0,4
Bio2RDF	378	4,1	IEE VIS	38	0,4
ie-ckan-net	272	2,9	Ayuntamiento de Zaragoza	37	0,4
It-ckan-net	263	2,8	Linked Education Cloud	36	0,4
Bioportal	243	2,6	AKSW	33	0,4
no-ckan-net	232	2,5	Where Does My Money Go?	31	0,3
Linking Open Data Cloud	210	2,3	Bibliographic Data	30	0,3
leeds-datamill-archive	201	2,2	US State Spending and Revenue Data	29	0,3
LODCloud2014	196	2,1	Ontology Engineering	29	0,3
Economics Datasets	148	1,6	Urban Design Studio	26	0,3
cz-ckan-net	147	1,6	Tetherless World Constellation	26	0,3
OpenSpending	142	1,5	Negawatt Challenge	26	0,3
International Budget Partnership	83	0,9	eagle-i	25	0,3
DAL	79	0,9	School of Data	25	0,3
OWLG	63	0,7	OpenSpending Cameroon	25	0,3
ClimateData	54	0,6	Open Knowledge Brasil	24	0,3
Statutarni mesto Brno	48	0,5	DataID	23	0,2
Occupy	47	0,5	Wikimedia	22	0,2
Linking Open Data	47	0,5	International Food Policy Research Institute	22	0,2
			Energy Data	22	0,2

Como se observa en la tabla I, la categoría “Global” es la que maneja la mayoría de Dataset publicados. Lo anterior se debe al proceso de migración realizado por la plataforma, dejando en esta categoría a aquellos Dataset que no eran parte de ninguna organización. Por otro lado, se observa que la siguiente categoría de mayor publicación es la correspondiente a “países”. De igual forma, en Datahub.io se encuentran publicados 10.900 Dataset.
Es importante anotar que hay organizaciones que:

- No tienen ningún Dataset publicado.
- Tiene un archivo de prueba de carga, que no configurar ningún tipo de Dataset.
- Tiene uno o más de un Dataset publicado, en diferentes formatos (que van desde CSV hasta PDF; XML, Sparql - SPARQL Protocol and RDF Query Language, etc.), y con diferentes tipos de licenciamiento.
- Tiene publicados otros archivos diferentes a un Dataset.

Dentro de la literatura se encuentran algunos trabajos relacionados tales como [19], quienes describen el problema de los enlaces abiertos y la estrategia para resolver este problema. Por su parte, en [20] se muestra las cantidades de datos vinculados disponibles a partir de julio de 2009 y el número de enlaces entre conjuntos de datos RDF; y en [21] se presenta un estudio de estadísticas acerca de la estructura y contenido de la nube LOD.

Es así como a partir de este panorama, el siguiente documento tiene como propósito presentar un estudio de caso sobre la configuración, acceso y consumo de datos que pueda ofrecer el API CKAN [17], con el fin de poder consultar una muestra de los Dataset publicados en Datahub, seleccionados de forma aleatoria, con los cuales se podrá identificar y extrapolar, de manera preliminar, las tendencia del modelo de la web de los datos a partir de los datos históricos presentes en Datahub.io, como plataforma de gestión de datos. Análisis que se plantea en etapas posteriores del proyecto que enmarca esta investigación. El motivo de usar esta API se debe a que CKAN está escrito en Python y hace uso de una variedad de framework de código abierto, incluido Pylons, el cual es una combinación de varios framework de código abierto integrados que forman la base para aplicaciones de nivel empresarial basados en la web.

El almacenamiento y gestión de datos de CAKN incluye el almacenamiento de archivos, gestión de metadatos, y la gestión de datos estructurados. Además, ofrece un mecanismo de plug-in, que permite a los desarrolladores extender rápidamente la funcionalidad de CKAN. De igual forma, CKAN posee la capacidad de soportar características geográficas, así como la exposición de metadatos de acuerdo con el catálogo estándar Open Geospatial Consortium (OGC, por sus siglas en inglés) y Catalog Service for the Web (CSW, por sus siglas en inglés). Por último, CKAN implementa funciones de limpieza cruciales tales como el logueo y gestión de usuarios [22].

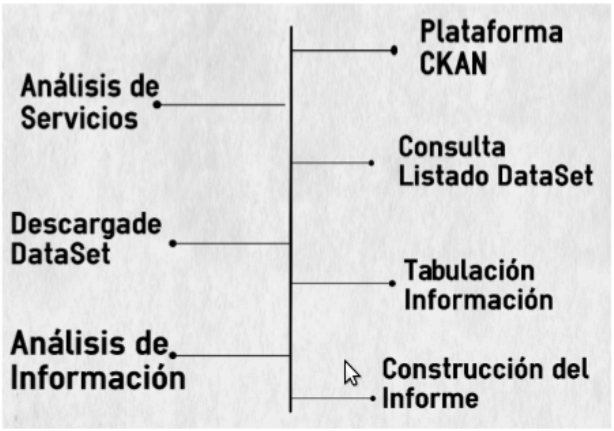


Figura 1. Metodología desarrollada para el proyecto

3. Planteamiento metodológico

Para llevar a cabo el proceso de conexión y descarga de los Dataset publicados en Datahub.io, se realizaron una serie de actividades tal como se describe en la figura 1.

Tal como se presenta en la figura 1 las actividades más relevantes son:

- Se revisó la plataforma CKAN, con el fin de establecer los servicios que ésta provee, al igual que la forma de acceder a ellos. Para esta labor se hizo uso del CKAN's API guide [14]- [15].
- Posteriormente se estudió la forma de conectarse a los servicios prestados por CKAN, con el fin de descargar los Dataset publicados en Datahub.io [14]- [15].

- Haciendo uso de los servicios de CKAN, se consultó el listado de los nombres de los Dataset publicados.
- Dada la cantidad de Dataset publicados, Datahub usa la estrategia de publicar grupos de Dataset (en sus respectivos formatos) por páginas, para un total de 545 páginas. Teniendo en cuento la distribución anterior, se descargaron archivos JSON, formato de archivo que se descarga localmente, los cuales contienen los Dataset publicados por página en la plataforma de gestión de datos.
- Por último, con el fin de revisar la integridad de la descarga, se llevó a cabo un análisis comparativo de la información, haciendo uso de herramientas que permitieran visualizar la data descargada, con la revisión manual de algunos Dataset publicados.
- Construcción del presente artículo, de acuerdo a los resultados obtenidos.

4. Materiales y métodos

Dado que el presente documento se basa en la descripción de un caso de estudio acerca de la conexión y consumo de los servicios ofrecidos por CKAN y, con el fin de que el procedimiento seguido pueda ser replicado a *posteriori*, a continuación se describen el proceso desarrollado con el fin de poder consumir los servicios del API, y por ende consultar los Dataset, dispuestos en páginas, para posteriormente ser descargados en archivos JSON, los cuales permitirán llevar a cabo un análisis preliminar de las tendencias del modelo de la web de los datos. El proceso seguido, junto con los recursos necesarios para la conexión y consumo se describe a continuación.

4.1. Recursos para la conexión a Datahub.io – CKAN API:

Datahub.io, como sitio desarrollado usando la API de CKAN, permite hacer uso de las funcionalidades que provee la API como consultar los metadatos (información acerca del dato alojado),

Tabla II. Recursos utilizados en la Conexión a CKAN

Recurso	Descripcion	Caracteristicas del recurso
Ubuntu desktop o server 12.04 o superior	Distribucion Linux, basada en Debian GNU/Linux. Sistema Operativo predominantemente enfocado a maquinas de escritorio, aunque tambien proporciona soporte para servidores	Las consultas realizadas en este proceso de exploracion se realizaron usando Ubuntu 15.10
Python 2.7 o superior	Python es un lenguaje de programacion de alto nivel, interpretado y de codigo abierto	Aunque se puede reemplazar con cualquier lenguaje que permita interactuar con el shell de Ubuntu. Para esta exploracion se uso Python, dada su afinidad con Ubuntu
Visualizador de archivos .json	Archivos JSON – JavaScript Object Notation: formato estandar abierto que utiliza texto legible para transmitir objetos de datos que consisten en pares atributo-valor. Para visualizar este tipo de archivos, en Google Chrome o Firefox, se debe instalar el plugin jSONView. De igual forma se pueden visualizar con un editor que permita la edicion de este tipo de archivos.	Para esta tarea se puede usar Sublime text 2/3 o cualquier editor que facilite la lectura
HTTPIE	Cliente HTTp de linea de Comando, utilizada para la comunicación e interaccion entre usuarios y web Service. Provee un comando HTTP simple que permite realizar peticiones HTPP usando sentencias simples. Es usado para realizar pruebas, depuracion e interaccion con servidores HTTP. La mayoria de las distribuciones de Linux proporcionan un paquete que se puede instalar mediante el gestor de paquetes del sistema.	Proporciona un comando HTTP que permite el envio de peticiones HTTP arbitrarias utilizando una sintaxis simple y natural. Aunque HTTPIE es compatible con Python 2.6 y 2.7, se recomienda instalar HTTPie en una version superior de de Python siempre que sea posible. Esto asegurará que algunas de las caracteristicas HTTP mas recientes, como SIN, funcionen de mejor forma.

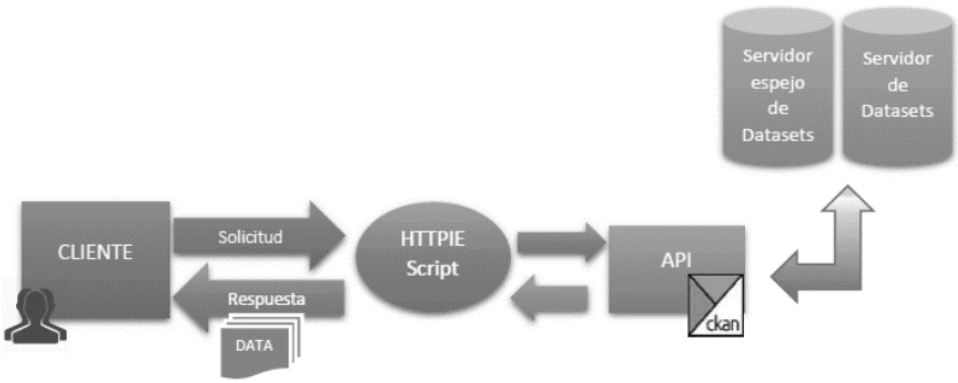


Figura 2. Interacción de Recursos de Conexión a CKAN

descargar la lista de los Dataset alojados, información de organizaciones, etiquetas que identifican determinado Dataset, entre otras funciones.

El proceso para realizar consultas sobre la API plantea el uso de los siguientes recursos (Tabla II).

La figura 2 se muestra la interacción de los recursos necesarios para conectarse a los servicios de CKAN.

4.2. Proceso de instalación y consulta al API:

Instalación HTTPie: una vez ubicado en el entorno Ubuntu, se procede a abrir una terminal y ejecutar el comando:

sudo apt-get install httpie

La figura 3 muestra los resultados de la ejecución del citado comando

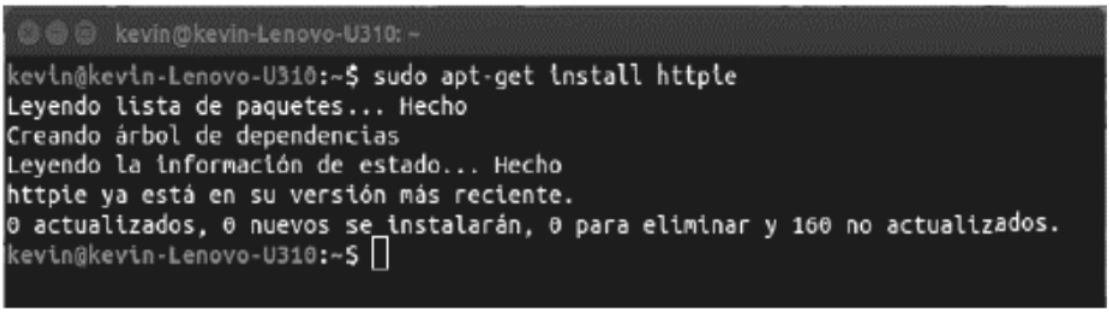


Figura 3. Instalación de HTTPie

Una vez ha terminado el proceso de instalación, ya se pueden realizar peticiones a servidores http, en este caso CKAN. La documentación de HTTpie se puede encontrar en el repositorio https://github.com/jkbrzt/httpie.

Ahora bien, la estructura básica para una consulta HTTPie se formula de la siguiente manera:

http [Flags] [Method] URL [Item [Item]]

Consulta de los DataSet: CKAN provee métodos para consultar los repositorios del sitio que ha sido desarrollado con su API. Para hacer uso de dichos métodos basta con agregar al sitio principal (que haga uso de CKAN) la sintaxis:

/api/3/action/ [método del CKAN]

Con el fin de aclarar lo descrito anteriormente, se hará una consulta a Datahub.io de todos los nombres de los Dataset que tiene y, por defecto, el servidor devolverá el resultado en una lista con estructura tipo JSON. Para realizar este proceso se abre una terminal y se digita:

http https://datahub.io/api/3/action/package_list

La figura 4 muestra los resultados de la ejecución del citado comando.

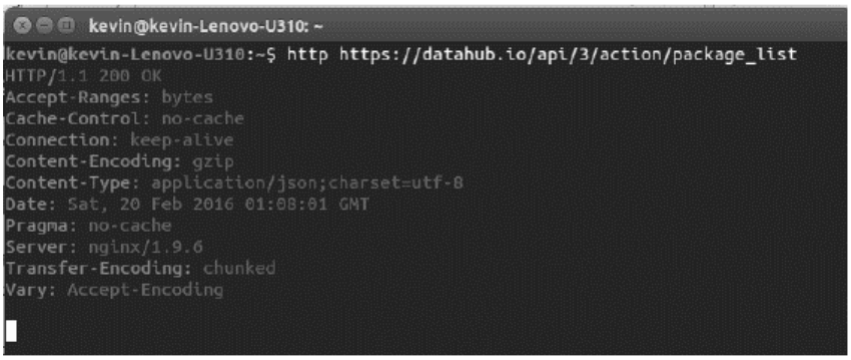


Figura 4. Consulta del Listado de Nombres de DataSet

Seguido a esta ejecución, se desplegará dentro de la misma consola los nombres de todos los Dataset de Datahub.io. Si adicionalmente se quiere descargar la información consultada y volcarla a un archivo, se debe agregar las banderas *--download* y *-o*, incluidas en HTTpie. La bandera *--download* permite descargar el archivo; y la bandera o permite nombrar y ubicar el archivo descargado. Por ejemplo, para descargar la lista anterior se realiza el proceso descrito en la figura 5.

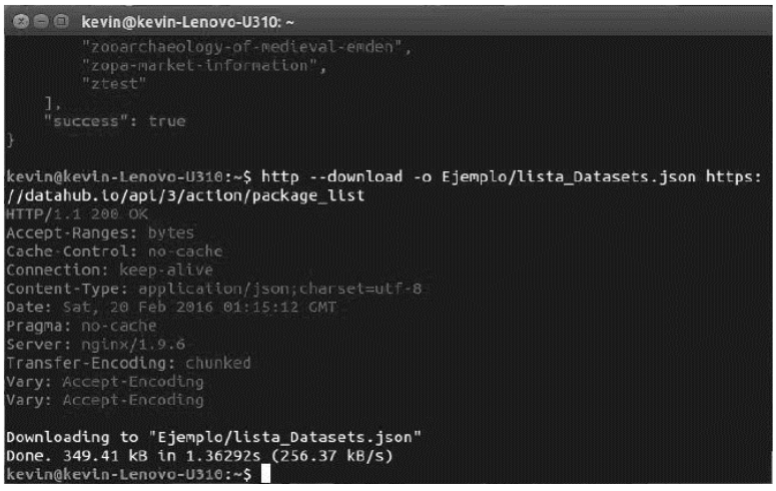


Figura 5. Descarga de la Consulta de Nombres de DataSet

Como resultado de esta ejecución se descarga el archivo con el nombre “lista_Datasets.json”, en el folder “Ejemplo”, previamente creado.

Por otro lado, para descargar todos los Dataset de Datahub.io, se puede hacer uso del método *current_package_list_with_resources*. Sin embargo, este método solo descarga de a diez Dataset si no se le especifica algún tipo de parámetro. Adicionalmente, su capacidad máxima de descarga de Dataset es de 1000, es decir que el archivo de descarga tendrá como máximo 1000 Dataset. Por ende, para descargar los Dataset publicados en la plataforma se escribe un script en Python, que haga uso de un ciclo *for*, que permita descargar los Dataset completos de cada página hasta completar 535 páginas. En la figura 6 se muestra el script y el resultado de su ejecución.

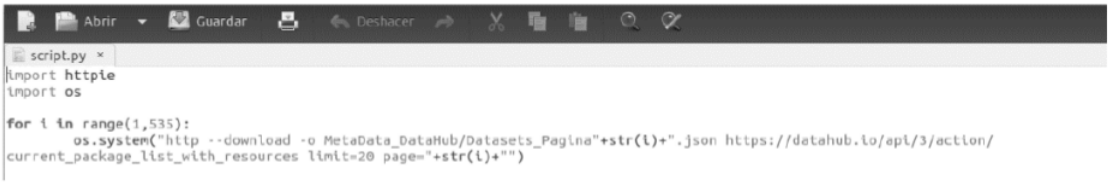


Figura 6. Archivo Python para descarga de DataSet

Para ejecutar el script desde la terminal, se debe ubicar en el directorio donde se va a guardar el archivo y se ejecuta el comando Python [Nombre del archivo], que para este caso corresponde: Python script.py

El resultado de la iteración es la generación de un archivo por página (535 archivos), como se muestra en la figura 7.

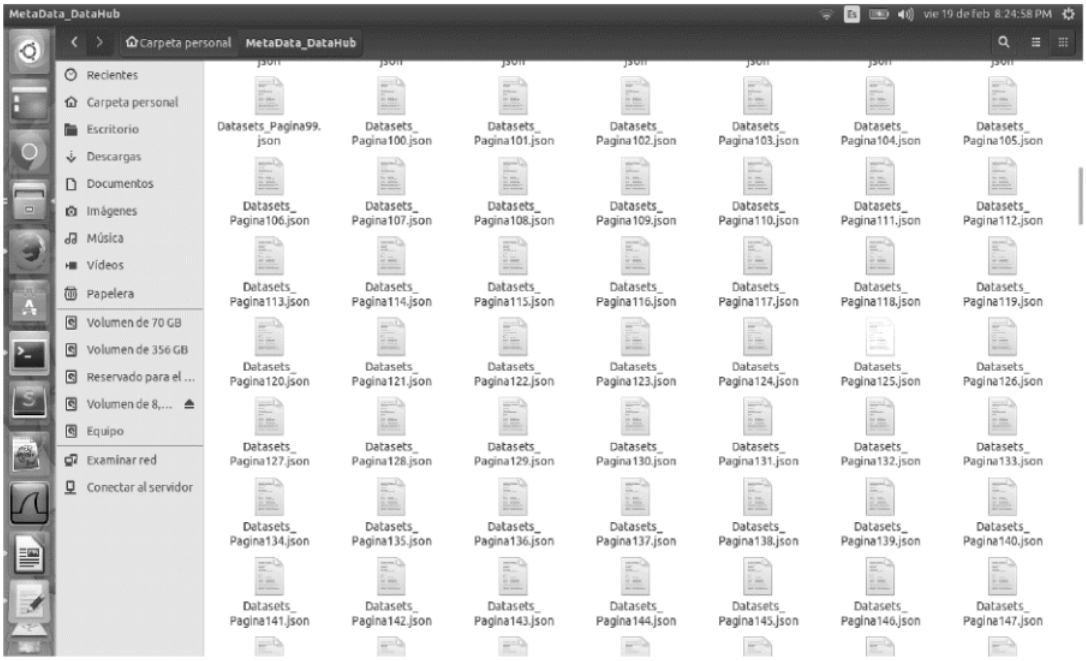


Figura 7. Archivos JSON descargados

5. Resultados y discusión

Como primer resultado obtenido en el proceso de conexión, se descargó el listado de Dataset publicados en la plataforma. La figura 8 muestra, al costado derecho el archivo JSON resultante de la consulta y al costado izquierdo el árbol de consulta resultante, donde cada valor de arreglo "result" corresponde a un nombre de Dataset consultado.

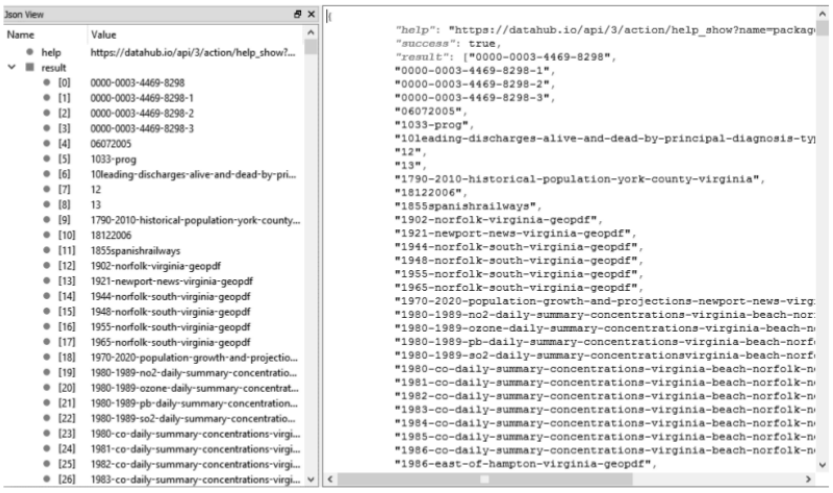


Figura 8. Listado de DataSet consultados

En total se obtuvieron 10694 entradas a la variable "result", que representa la misma cantidad de nombres de Dataset consultados.

Posterior a esta consulta, se procedió a la descarga del listado de Dataset, como se especificó en la sección de Métodos y Materiales, generando 545 archivos correspondientes al mismo número de páginas existentes en la plataforma. Con este producto obtenido, se procedió a comparar y verificar la información resultante del proceso de descarga:

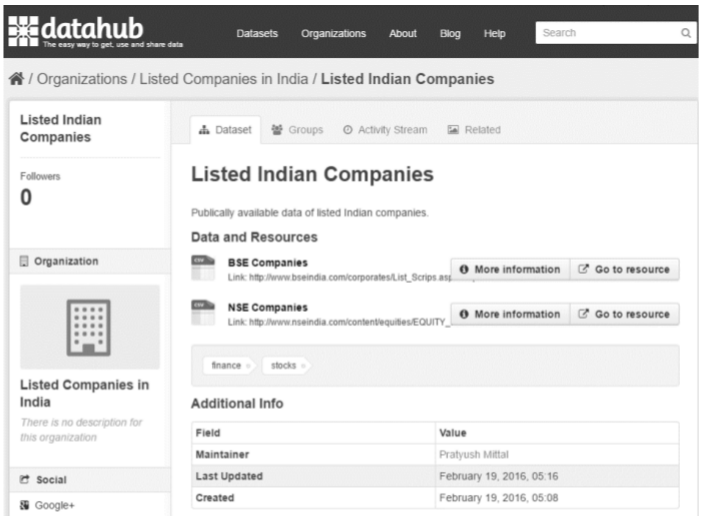


Figura 9. Organización consultada en Datahub.io

a. DataSet consultado en la plataforma de Datahub.io: para revisar manualmente el proceso se tomó como ejemplo y punto de referencia la Organización “Listed Indian Companies”, la cual tiene un Dataset publicado en dos formatos, como se observa en la figura 9.

b. En la figura 10 se observa el segmento del archivo JSON, que contiene la información de esta organización.

1 {"help": "https://datahub.io/api/3/action/help_show?name=current_package_list_with_resources", "success": true, "result": [{ "license_title": "Creative Commons Attribution", "maintainer": "Pratyush Mittal", "relationships_as_object": [], "private": false, "maintainer_email": "pratyushmittal@gmail.com", "num_tags": 2, "id": "b78af181-811b-4182-b624-82919ea05556", "metadata_created": "2016-02-19T05:08:41.492537", "metadata_modified": "2016-02-19T05:16:00.401035", "author": "", "author_email": "", "state": "active", "version": "", "creator_user_id": "ede20aa6-dda4-4535-a82e-58f4d48e9930", "type": "dataset", "resources": [{"mimetype": null, "cache_url": null, "hash": "", "description": "Link: http://www.bseindia.com/corporates/List_Scrips.aspx?expandable=1", "name": "BSE Companies", "format": "CSV", "url": "https://datahub.io/dataset/b78af181-811b-4182-b624-82919ea05556/resource/372c614e-064f-4521-8136-068366be578c/download/listofscrips.csv", "cache_last_updated": null, "package_id": "b78af181-811b-4182-b624-82919ea05556", "created": "2016-02-19T05:10:47.980032", "state": "active", "mimetype_inner": null, "webstore_last_updated": null, "last_modified": null, "position": 0, "revision_id": "1eb09546-0ee3-4089-a2b8-96b5b55aaace", "webstore_url": null, "url_type": "upload", "id": "372c614e-064f-4521-8136-068366be578c", "resource_type": null, "size": null}, {"mimetype": null, "cache_url": null, "hash": "", "description": "Link: http://www.nseindia.com/content/equities/EQUITY_L.csv", "name": "NSE Companies", "format": "CSV", "url": "https://datahub.io/dataset/b78af181-811b-4182-b624-82919ea05556/resource/ab8acf8a-323f-433c-9388-3ba1328fc490/download/equityl-1.csv", "cache_last_updated": null, "package_id": "b78af181-811b-4182-b624-82919ea05556", "created": "2016-02-19T05:11:40.260386", "state": "active", "mimetype_inner": null, "webstore_last_updated": null, "last_modified": null, "position": 1, "revision_id": "30f94f18-0df5-4d75-9470-b91f6f007614", "webstore_url": null, "url_type": "upload", "id": "ab8acf8a-323f-433c-9388-3ba1328fc490", "resource_type": null, "size": null}], "num_resources": 2, "tags": [{"vocabulary_id": null, "state": "active", "display_name": "finance", "id": "98b2af7c-d7f0-4cfd-b980-1339b805873b", "name": "finance"}, {"vocabulary_id": null, "state": "active", "display_name": "stocks", "id": "152a3b6f-f347-41af-8de6-535e51c28af6", "name": "stocks"}], "groups": [], "license_id": "cc-by", "relationships_as_subject": [], "organization": {"description": "", "created": "2016-02-12T07:56:03.046042", "title": "Listed Companies in India", "name": "listed-indian-companies", "is_organization": true, "state": "active", "image_url": "", "revision_id": "ff2d7339-2560-4c75-a5b4-e2191af950a8", "type": "organization", "id": "221a69a5-5609-431e-b246-acc2b4facf70", "approval_status": "approved"}, "name": "listed-indian-companies", "isopen": true, "url": "", "notes": "Publically available data of listed Indian companies.", "owner_org": "221a69a5-5609-431e-b246-acc2b4facf70", "extras": [], "license_url": "http://www.opendefinition.org/licenses/cc-by", "title": "Listed Indian Companies", "revision_id": "9e96feee-a26e-4059-8b6e-cc130e2308fb"}, {"license_title": "Creative Commons Attribution", "maintainer": "Heather Mann", "relationships_as_object": [], "private": false, "maintainer_email": "heather.mann@gmail.com", "num_tags": 4, "id": "b8321a70-4fa0-422f-841f-f76e99a4743a", "metadata_created": "2016-02-18T19:01:41.047555", "metadata_modified": "2016-02-18T19:30:19.946115", "author": "Heather Mann, Ximena Garcia-Rada, Lars Hornuf, Juan Tafurt, Dan Ariely", "author_email": "heather.mann@gmail.com", "state": "active", "version": "", "creator_user_id": "e59b6d7b-294e-4f74-b260-09952f249160", "type": "dataset", "resources": [{"mimetype": null, "cache_url": null, "hash": "", "description": "", "name": "Cross-cultural dishonesty study - Reports of roll outcomes on die task", "format": "CSV", "url": "https://datahub.io/dataset/b8321a70-4fa0-422f-841f-f76e99a4743a/resource/f475f4a6-2fdc-4824-9f6e-5ec23a682018/download/cross-cultural-dishones

Figura 10. Segmento del archivo JSON descargado en la consulta

c. Haciendo uso de herramientas visualizadoras de archivos JSON, se puede revisar la información contenida en el segmento JSON, tal como los muestran las figuras 11 y 12.

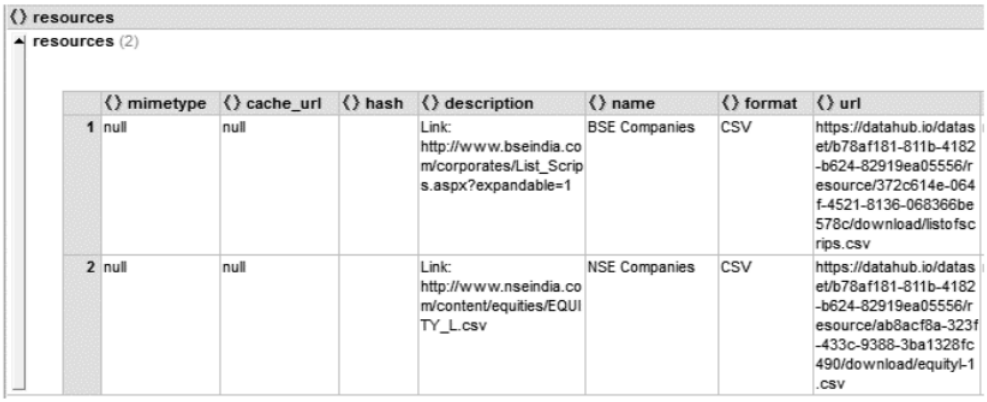
d. De igual forma, con este tipo de herramientas se pueden revisar parámetros de evaluación, tales como tipo de licenciamiento, última fecha de actualización, tipo de formatos publicados, entre otros que pueden brindar información relevante al momento de evaluar el estado de la web de los datos. La figura 13 permiten evidenciar la información brindada por este tipo de herramientas.

{} help	https://datahub.io/api/3/action/help_show?name=current_package_list_with_resources
{} success	true
result (20)	

	{} license_title	{} maintainer	{} relationships_a...	{} private	{} maintainer_email	{} num_tags
1	Creative Commons Attribution	Pratyush Mittal	relationships_a...	false	pratyushmittal@gmail.com	2
2	Creative Commons Attribution	Heather Mann	relationships_a...	false	heather.mann@gmail.com	4

Figura 11. Visualización del archivo JSON en modo cuadrícula

Figura 12. Visualización del archivo JSON en modo árbol



	mimetype	cache_url	hash	description	name	format	url
1	null	null		Link: http://www.bseindia.com/corporates/List_Scrips.aspx?expandable=1	BSE Companies	CSV	https://datahub.io/dataset/b78af181-811b-4182-b624-82919ea05556/resource/372c614e-064f-4521-8136-068366be578c/download/listofscrips.csv
2	null	null		Link: http://www.nseindia.com/content/equities/EQUITY_L.csv	NSE Companies	CSV	https://datahub.io/dataset/b78af181-811b-4182-b624-82919ea05556/resource/ab8acf8a-323f-433c-9388-3ba1328fc490/download/equityl-1.csv

Figura 13. Información obtenida de los DataSet

Como se observa en la figura 13, el Dataset consultado como ejemplo de verificación, tiene dos recursos publicados: BSE Companies y NSE Companies, ambos en formato CSV, bajo licencia cc-by, con fecha de última actualización el 19 de febrero de 2016.

e. Para la realización de un análisis preliminar de la información descargada, se seleccionó, de forma aleatoria, una muestra de 54 organizaciones y se procedió a revisar los Dataset publicados, 311 en total, con el fin de levantar la información acerca de dominios, formatos licenciamiento y última fecha de publicación de los Dataset.

Como resultado de este proceso se obtuvo la siguiente información:

• **Dominio de Datos:** En los 311 Dataset revisados, en la figura 14 se categorizan los dominios de la siguiente manera:

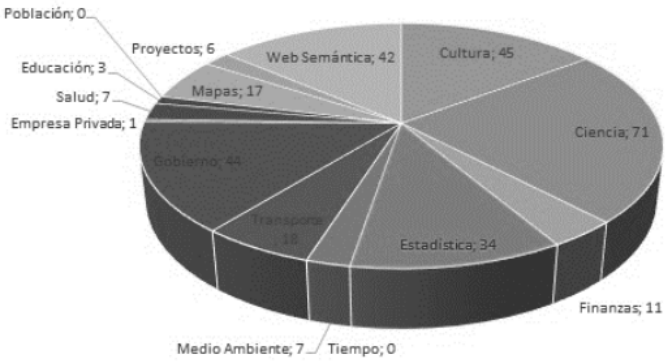


Figura 14. Dominio de Datos versus Cantidad de DataSet

En la revisión de dominio de publicación se encuentra que los dominios de tipos de datos abiertos con mayor publicación son:

- Ciencia, con un 23 %, en la cual los mayores subdominios son arqueología, química y bioinformática.
- Cultura, gobierno y web semántica, con un 14 %.
- Estadística, con un 11 %.

• **Formatos de publicación:** como se observa en la revisión preliminar realizada en Datahub.io (Tabla III), los DataSet están cargados en la plataforma en diferentes formatos (algunos de las organizaciones cargan más de un tipo de formato).

En la revisión de estos formatos, se puede observar que la plataforma no maneja un estándar unificado en cuanto a los formatos subidos, lo anterior añadido a que los formatos que se cargan no presentan una estandarización, lo que se puede observar al encontrarse como XLS, XLSX, Hoja de Cálculo, Excel, SpreedShet, MS Excel.

De igual forma, se identifica que hay publicaciones de Dataset realizadas en formatos no estructurados como PDF, PNG. Como se observa en la tabla III, en Datahub.io se pueden cargar diferentes formatos de un Dataset, encontrando por ejemplo que hay 1389 publicaciones que tiene formato de hojas de cálculo en sus formatos de publicación, lo que representa que un 13 % de los Dataset publicados y, como formatos propietarios, requieren herramientas que no son públicas.

También se identifica que varias organizaciones le han apostado a depurar sus publicaciones, escalando en el modelo propuesto por Berners-Lee [1], ofreciendo datos en formatos estructurados, abiertos al público y que permiten llevar a cabo el proceso de vinculación. Sin embargo, en este proceso también se observa en Datahub, situaciones como la disparidad de formatos de publicación, algunos de los cuales no permiten la interoperabilidad semántica, limitándose a la descripción sintáctica de la información o, por el contrario, publicando formatos que dificultan el procesamiento por las máquinas.

Tabla III. Cantidad de formatos de publicación, obtenidos de la muestra analizada

Formato	Cantidad	%	Formato	Cantidad	%
CSV	1049	11,4	application/zip	101	1,1
XLS	803	8,7	example/turtle	95	1,0
Example rdf+xml	750	8,1	XML	92	1,0
Api/sparql	674	7,3	Json	90	1,0
text/csv	403	4,4	GeoJSON	90	1,0
meta/void	356	3,9	TopoJSON	88	1,0
application/rdf+xml	337	3,6	application/octet-stream	87	0,9
Csv	319	3,5	application/json	81	0,9
PDF	298	3,2	Xml	72	0,8
XLSX	251	2,7	RDF	68	0,7
ODS	248	2,7	application/x-ntriples	66	0,7
HTML	224	2,4	Pdf	58	0,6
Html	215	2,3	rdf/turtle	57	0,6
Aspx	210	2,3	Other	53	0,6
application/pdf	190	2,1	Data File in EXCEL	49	0,5
Spreadsheet	185	2	ZIP	48	0,5
text/turtle	176	1,9	JPG	48	0,5
meta/rdf-schema	165	1,8	Shp	46	0,5
text/html	157	1,7	JSON	46	0,5
SHP	144	1,6	Zip	45	0,5
application/vnd.ms-ex-cel	140	1,5	application/trig	44	0,5
KML	132	1,4	PNG	42	0,5
meta/sitemap	123	1,3	mapping/owl	411	0,4
Xls	101	1,1	example/rdfa	40	0,4
			GML	39	0,4

• **Fecha de última actualización:** en cuanto a la fecha de última actualización, teniendo en cuenta que bajo los Principios de Open Data [16], estos datos deben ser "Actualizados", con el fin de que no pierdan su valor y sean precisos, como resultado de la exploración muestral realizada, se identifica que el 1 % de los mismos alcanzaría un grado de actualización frecuente, un 7,7 % un grado de actualización media, mientras que un 60 % presenta un grado de actualización deficiente.

En este aspecto es importante resaltar que varios de los dominios corresponden a datos estadísticos generados de periodos anteriores, los cuales no requerirían de una actualización permanente, sin embargo, esta disparidad en el proceso de actualización permite reflexionar acerca del grado de actualización de los Dataset publicados.

6. Conclusiones y trabajos futuros

CKAN es una herramienta potente para gestionar catálogos de datos, permitiendo manejar una descripción de los datos y otras informaciones relevantes, tanto para las organizaciones que publican como para las personas que consultan dicha información, tales como categorías de organizaciones, formatos en que se encuentra disponible los datos, propietario de los datos, el tipo de licenciamiento de las publicaciones, enlaces a otros datos.

Por otro lado, CKAN es una herramienta usada en muchos catálogos de datos abiertos, disponibles en Web. Las organizaciones que hacen uso de CKAN, tanto privadas como públicas, publican

sus datos haciendo uso de algún nivel de las recomendaciones de buenas prácticas de Open Data y las propuestas por Berners-Lee.

En cuanto a los servicios ofrecidos por CKAN, para la consulta y descarga de la data, son servicios consistentes, que ofrecen peticiones y respuestas en formato JSON, sin restricciones al público. Existen herramientas tales como Python, Java, Java, Ruby, PHP, que pueden interactuar con el API REST de CKAN, permitiendo obtener la información solicitada, la cual puede ser visualizada en diferentes formatos, haciendo uso de herramientas visualizadoras, que permiten hacer las consultas respectivas de formas más ágil.

Como lo plantea [24], es complejo seleccionar una plataforma sin realizar un estudio cuidados de las necesidades de los interesados. En el estudio citado, se plantea que CKAN se está convirtiendo en uno de los referentes en cuanto a la gestión de catálogos de fuentes de datos, pero que a su vez presenta desventajas como el mantenimiento de una plataforma tecnológica diferente al CMS usado para gestionar los datos, falta de soporte nativo de RDF para el enriquecimiento semántico de los datos, y el despliegue de una aplicación web diferente al portal CMS usado [26].

Por su parte, Datahub.io, como plataforma de gestión de datos, permite una publicación de datos ágil, y con pocas restricciones al usuario final, a través de una interfaz web sencilla. Por otro lado, si como desarrollador o investigador, se requiere a Datahub.io, para consultar o descargar sus Dataset, ofrece una como apoyo el CKAN API, interface que interactúa con diferentes herramientas, ofrece una variedad de servicios, para llevar a los procesos requeridos.

Ahora bien, dentro de las situaciones que se observan en las publicaciones realizadas en Datahub.io, corresponde a la interpretación que algunas organizaciones en Datahub le han dado al concepto de Datos Abiertos, tomando a la plataforma como una estrategia para publicar o promocionar escritos, artículos, eventos, etc., que no poseen ningún tipo de vinculación ni de contexto ampliado a los datos expuestos. Tal es el caso de la organización ‘Mente Clara’, donde sus 13 publicaciones son imágenes y PDF (artículos) acerca del budismo tántrico tibetano.

Otro aspecto identificado es el acceso a ciertos recursos de los Dataset tomados como muestra, dado que las URI no accedían, generando inconvenientes para el proceso de vinculación. Este tipo de problemas podría ser causado a la falta de actualización de los datos publicados en la plataforma, entre otras situaciones. Independiente del origen de dicho problema, uno de los factores claves de la web de los datos es el proceso de identificación y vinculación de datos que aporten contexto al dato publicado, y la falta de enlaces resolubles afecta el desarrollo de esta tecnología.

De otro lado, en el caso de los gobiernos, por ejemplo, se identifican publicaciones que responden a los principios de transparencia y apertura de información, pero que bajo los principios de Datos Abiertos y de Linked Data, no aportan contexto a la información expuesta, limitándose a publicar listados de cuadros en hojas de cálculo, algunos de los cuales no se actualizan de forma periódica. Un ejemplo de ello es por ejemplo la organización ‘Bolivia’ que registra seis Datasets, en los cuales todos tiene como última fecha de actualización hace en promedio dos años, y la información contenida en ellos corresponden a tablas de hojas de cálculo de informes de indicadores sociales, que no se complementan o añaden contexto a la información, con otros enlaces o datos de interés.

Como trabajos futuros del proceso de investigación, y dados los hallazgos identificados, se plantea: 1) llevar a cabo un análisis más detallado del estado de la Web de los datos, a través de herramientas de analítica visual, 2) realizar un análisis del estado de la Web de los datos en el área de recursos digitales abiertos mediante propuestas de accesos seguros a plataformas LCMS [27], objeto de estudio de la investigación, y por último, 3) el planteamiento de una propuesta metodológica para procesos de vinculación de datos de cara a los recursos digitales abiertos, analizando los hallazgos encontrados para su aplicación en entornos educativos.

Agradecimientos

Esta investigación se lleva a cabo en el marco de la formación doctoral en Ingeniería, en la Universidad Distrital Francisco José de Caldas. De igual forma, la temática planteada se configura como una línea de investigación de Grupo GIIRA.

Referencias

[1] T. Berners-Lee, C. Bizer, T. Heath, "Linked data-the story so far". International Journal on Semantic Web and Information Systems, vol. 5, pp. 1-22, 2009. ↑ 47, 60

[2] S. Dietze, H. Yu, D. Giordano, E. Kaldouidi, N. Dovrolis, D. Taivi, "Linked Education: Interlinking Educational Resources and the Web of Data". 27th annual ACM symposium on Aapplied Computing, pp. 366-371, 2012. [En línea]. Disponible en: http://oro.open.ac.uk/31077/. ↑ 47

[3] B. Haslhofer, A. Isaac, "Data. europeana. eu: The Europeana Linked Open Data Pilot ". International Conference on Dublin Core and Metadata Applications, pp. 94-104, 2011. [En línea]. Disponible en: http://dcpapers.dublincore.org/pubs/article/view/3625/1851. ↑47

[4] D'Aquin, M. et al ., "Building the Open Elements of an Open Data Competition " D-Lib Magazine, vol. 20, p. 3, 2014. [En línea]. Disponible en: http://www.dlib.org/dlib/may14/daquin/05daquin.html. ↑ 47

[5] L. Project, "Linking Web Data for Education". [En línea]. Disponible en: http://linkedup-project.eu/. ↑ 47

[6] G. Klyne, J. Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax " . 2006. Edited by Brian Mcbride. [En línea]. Disponible en: http://www.citeulike.org/group/2170/article/532408. ↑ 47

[7] M. Hausenblas, "Exploiting Linked Data To Build Web Applications ". IEEE Internet Computing, vol. 13, no 4, p. 68. 2009. ↑ 48

[8] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, "DBpedia: A Nucleus for a Web of Open Data ". The Semantic Web , Busan, Korea, Springer, 2007. ↑48

[9] LOD-Cloud. "The Linking Open Data Cloud Diagram". S.f. [En línea]. Disponible en: http://lod-cloud.net/. ↑48, 49

[10] Datahub, "Datahub Project". S. f. [En línea]. Disponible en: http://datahub.io/dataset?tags=lod. ↑ 48, 51

[11] Ministerio de Hacienda y Administración Pública, Ministerio de Industria Energía y Turismo, "Plataformas de publicación de datos abiertos". [En línea]. Disponible en: http://datos.gob.es/sites/default/files/informe-herramientas-publicacion.pdf. ↑48, 49

[12] OKF, "Open Knowledge Foundation ". S. f. [En línea]. Disponible en: http://services.okfn.org/. ↑49

[13] CKAN, "The Open Source Data Portal Software". S. f. [En línea]. Disponible en: http://ckan.org/. ↑ 49

[14] CKAN, "API Guide-CKAN's API for developers". S. f. http://docs.ckan.org/en/latest/api/index.html. ↑ 52

[15] CKAN, "Ckan Wiki-CKAN Pages". S.f. [En línea]. Disponible en: https://github.com/ckan/ckan/wiki/_pages. ↑ 52

[16] J. Wonderlich, "Ten Principles for Opening up Government Information". 2010. [En línea]. Disponible en: http://sunlightfoundation.com/policy/documents/ten-open-data-principles/. ↑ 61

[17] CKAN, "CKAN API Guide " . S. f. [En línea]. Disponible en: http://docs.ckan.org/en/latest/api/. ↑ 48, 52

[18] J. Winn, "Open Data and the Academy: an Evaluation of CKAN for Research Data Management. (IASSIST 2013)". 28-31 Mayo 2013. [En línea]. Disponible en: http://eprints.lincoln.ac.uk/9778/1/CKANEvaluation.pdf. ↑

48
[19] E. Rajabi, S. Sánchez-Alonso, M.-A. Sicilia, “Analyzing broken links on the web of data: An experiment with DBpedia ” . Journal of the Association for Information Science and Technology , vol. 65, no 8, p. 1721–1727, 2014. [Onine]. Disponible en: http://onlinelibrary.wiley.com/doi/10.1002/asi.23109/abstract. ↑ 52
[20] C. Bizer, “The Emerging Web of Linked Data”. IEEE Intelligent Systems, vol. 24, no 5, pp. 87-92, 2009. [Onine]. Disponible en: http://lpis.csd.auth.gr/mtpx/sw/material/IEEE-IS/IS-24-5.pdf. ↑ 52
[21] HPI Institut, “State of LOD Cloud”. 2011. [En línea]. Disponible en: http://lod-cloud.net/state/. ↑52
[22] M. Allison, S. Richard, K. Patten, C. Caudill-Daugherty, A. Anderson, “Open Access to Geoscience Data for Exploration and Assessment”. 19 al 25 Abril 2015. [En línea]. Disponible en: http://www.geothermal-energy.org/pdf/IGAstandard/WGC/2015/33032.pdf. ↑ 52
[23] W. Mao, J. Jan, “Visualization of Open Data: a CaseStudy of Climate Data”. 36 Asian Conference on Remote Sensing. 19-23 de Octubre de 2015. Manila, Philippines. [En línea]. Disponible en: http://www.acrs2015.org/list-of-accepted-abstracts/. ↑ 49
[24] R. Carvalho, J. Aguiar, J. Rocha, C. Ribeiro, “A Comparision of Research Data Management Platforms: Architecture, Flexible Metadata and Interoperability ”. Junio de 2016. [En línea]. Disponible en: https://www.researchgate.net/publication/303918099_A_comparison_of_research_data_management_platforms_architecture_flexible_metadata_and_interoperability ↑ 49, 62
[25] CURE, “Infraestructura semántica basada en el paradigma de datos abiertos para la gestión de investigacion de las universidades españolas”. CRUE Universidades Españolas , 2016. [En línea]. Disponible en: http://tic.crue.org/wp-content/uploads/2016/07/Memoria-proyecto-H %C3 %A9rcules.pdf. ↑ 49
[26] Datos.gob.es, “Estudio de plataformas tecnológicas datos.gob.es”. S. f. Ministerio de Industria, Turismo y Comercio. [En línea]. Disponible en: http://datos.gob.es/sites/default/files/files/2_cms_01.pdf. ↑ 62
[27] P.A. Gaona-García, C. E. Montenegro y H.W. González, “Hacia una propuesta de mecanismos para la autenticidad de objetos de aprendizaje en plataformas LCMS”. Revista Ingeniería , vol. 19, no 1, p. 50-64. ↑ 63

Jhon Francined Herrera Cubides

Ingeniero de Sistemas, Corporación Universitaria del Meta; especialista en Construcción de Software para Redes, Universidad Autónoma de Colombia; magíster en Ingeniería-Sistemas y Computación, Universidad de los Andes; estudiante del Doctorado en Ingeniería, Universidad Distrital Francisco José de Caldas; profesor asistente en el área de Programación, Facultad de Ingeniería de la Universidad Distrital Francisco José de Caldas; adscrito al grupo de investigación “GIIRA”, donde realiza estudios sobre vinculación de datos. Correo electrónico: jfherrerac@udistrital.edu.co

Paulo Alonso Gaona-García

Ingeniero de Sistemas, Universidad Distrital Francisco José de Caldas; magíster en Ciencias de la Información y de las Comunicaciones, Universidad Distrital Francisco José de Caldas; profesor de tiempo completo adscrito a la Facultad de Ingeniería de la Universidad Distrital Francisco José de Caldas; doctor en Ingeniería de la Información y del Conocimiento, Universidad de Alcalá; sus áreas de interés se encuentran enfocadas en redes y comunicaciones, seguridad informática, e-learning, analítica visual de datos y Web semántica. Correo electrónico: pagao- nag@udistrital.edu.co

Kevin Gordillo Orjuela

Estudiante de Ingeniería de Sistemas adscrito al grupo de investigación GIIRA de la Universidad Distrital Francisco José de Caldas. Correo electrónico: ksgordilloo@correo.udistrital.edu.co