UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Visión Electrónica

Más que un estado sólido

https://revistas.udistrital.edu.co/index.php/visele

VISIÓN ELECTRÓNICA

A RESEARCH VISION

Confidence level evaluation of LOD resources on CKAN instances

Evaluación del nivel de confianza de los recursos LOD en instancias CKAN

Jhon Francined Herrera-Cubides [1], ***Paulo Alonso Gaona-García*** [2], ***Carlos Enrique Montenegro-Marín*** [3], ***Álvaro Varón-Capera*** [4]

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Enviado: 12/01/2019
Recibido: 15/02/2019
Aceptado: 30/03/2019

Keywords:
CKAN
Linked Open Data
Machine Learning
Open Data
TensorFlow
Visual Analytics

Palabras clave:
CKAN
Linked Open Data
Aprendizaje de Máquina
Datos Abiertos
TensorFlow
Analítica Visual

ABSTRACT

Linked Open Data has been an initiative aimed at offering principles for the interconnection of data through machine-readable structures and knowledge representation schemes. At present, there are platforms that allow consuming LOD resources, being CKAN one of the most relevant on a large community made up of governmental organizations, NGOs, among others. However, the resources consumption lacks minimum criteria to determine their validity such as level of trust, quality, linkage and usability of the data; aspects that require a previous systematic analysis on the set of published data. To support this process of analysis and determination of the mentioned criteria, this paper has as purpose to present a method that allows analyzing the dataset current state obtained from the different instances published in CKAN, with the aim of evaluating the levels of trust that can offer from their sources. Finally, it presents results, conclusions and future work from the use of the tool for the dataset consumption belonging to certain instances ascribed to the CKAN platform.

RESUMEN

Linked Open Data ha sido una iniciativa orientada a ofrecer una serie de principios para la interconexión de datos mediante estructuras legibles por máquinas y esquemas de representación de conocimiento. En la actualidad existen plataformas que permiten consumir este tipo de recursos LOD, siendo CKAN una de las más relevantes sobre una gran comunidad conformada por organizaciones gubernamentales, ONGs, entre otras. Sin embargo, el consumo de estos recursos carece de criterios mínimos para determinar la validez de los mismos tales como: nivel de confianza, calidad, vinculación y usabilidad de los datos; aspectos que requieren de un análisis sistemático previo sobre el conjunto de datos publicados. Para apoyar este proceso de análisis y determinación de los criterios mencionados, el presente artículo tiene como propósito presentar un método que permita analizar el estado actual de los dataset obtenidos desde las distintas instancias publicadas en CKAN, con el propósito de evaluar los niveles de confianza que pueden ofrecer desde sus fuentes de origen. Finalmente, presenta resultados, conclusiones y trabajo futuro a partir del uso de la herramienta para el consumo de conjuntos de datos pertenecientes a ciertas instancias adscritas a la plataforma CKAN.

[1] Ph.D. (c) In Engineering, MSc. In Computer and Systems Engineering, BSc. In Systems Engineering, Universidad Distrital Francisco José de Caldas, Colombia. Current position: Professor at Universidad Distrital Francisco José de Caldas, Colombia, and active member of GIIRA research group. E-mail: jfherrerac@udistrital.edu.co.

[2] Ph.D. In Information of Engineering and Knowledge, MSc. In Information Sciences and Communications, BSc. In Systems Engineering, Universidad de Alcalá, Spain. Current position: Professor at Universidad Distrital Francisco José de Caldas, Colombia, and active member of GIIRA research group. E-mail: pagaonag@udistrital.edu.co.

[3] Ph.D. In Systems and Informatics Services for Internet, MSc. In Information Sciences and Communications, BSc. In Systems Engineering, Universidad de Oviedo, Spain. Current position: Professor at Universidad Distrital Francisco José de Caldas, Colombia, Director of GIIRA research group. E-mail: cemontenegro@udistrital.edu.co.

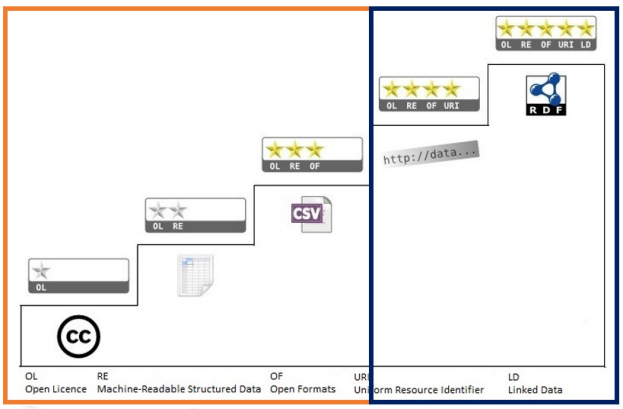
[4] BSc. In Systems Engineering, Universidad Distrital Francisco José de Caldas, Colombia. Current position: Multibrands group, Colombia. E-mail: avaron@multibrandscol.com.

Citar este artículo como: J. F. Herrera-Cubides, P. A. Gaona-García, C. E. Montenegro-Marín and A. Varón-Capera, "Confidence level evaluation of LOD resources on CKAN instances" , Visión electrónica, algo más que un estado sólido, vol. 13, no. 2, p.p 233-246, july-december 2019, https://doi.org/10.14483/22484728.15158

1. Introduction

Linked Open Data - LOD allows user to link and exploit data from various sources, freely and without licensing restrictions [1]. LOD is based on: a) Linked Data, as a set of design principles to share interconnected data that are readable by the machines; it generally focus on tools that provide meaning to data, and ontologies that provide meaning to the terms [2, 3]. b) Open Data, which focuses on use, reuse and redistribution of data, therefore, these data must be available and must be in public domain or have licensing conditions that allow user to use data as he wishes without restrictions [2].

Figure 1: 5 Stars Scheme [4].



In LOD, you should take into account:

- Approach 5 Stars scheme for data publication, a scheme which a great deal of dataset does not necessarily reach the last 2 stars (Figure 1), where linkage actually occurs [5, 6].
- Have repositories that publish dataset and metadata [7, 8], that facilitate searches (in different knowledge domains) to interested people [9]. In addition, they are configured as open access repositories [10]. All of this topics in order to carry out the resources exploitation.

On the other hand, repositories work under platforms for data management. CKAN - Comprehensive Knowledge Archive Network [11, 12] is a tool for management and publication of data collections (dataset) in a Web environment. CKAN allows to producers publish data with linked resources, and for consumers, through services offered by this platform, they can consume dataset according to searches made.

Dataset exploitation, through platform services such as CKAN, faces challenges such as comprehensibility of obtained results, visualization of query results, dealing with formats, licensing, trust, heterogeneity and interoperability [13], access and other aspects studied in researches such as [14, 15]. These researches raise the need for tools to detect possible quality problems and ambiguities produced by redundancy, inconsistencies and lack of completeness of data and links. Open link problems and strategy in order to solve this problem are described in [16, 17]. Quantity of linked data available as of July 2009 and the number of links between RDF dataset are shown in [18]. Different efforts are made through use of Fuzzy Logic as a strategy to evaluate the quality of dataset are presented in in [19, 20]. Finally, a statistics study about the structure and content of the LOD cloud is presented in [21].

With the growing dataset publication, process how to consume these datasets quickly and efficiently is evident as a challenge [22], which exposes the lack of a tool which consumer can evaluate the exploited data confidence, based on basic information of resources [7], such as licenses, formats, among others, and be able to make decisions about consumed dataset. To address this requirement, Visual Analytics offers strategies to identify quickly hidden trends, patterns and anomalies, as well as to explore and collect general information from large data spaces and reach quickly points of interest [22].

Making use of visual data analytics, this research focuses on development of a framework for dataset consumption in different CKAN instances [11, 12], which allows retrieving pertinent information for visualization process. The purpose of this framework is to have elements that allow analyzing and evaluating metadata quality, thus promoting generation of confidence in data consumption.

This article is organized as follows: the state of the art is described in Section 2, where references of the proposed topic are reviewed. Subsequently, the methodology and methodological design used to explore the data sets are described in section 3. The methodological development is exposed in section 4, which includes a brief description of proposed tool for analysis of resources. The obtained results are presented in section 5. The result analysis and discussions are described in Section 6. Finally, conclusions and future work are argumented in section 7.

2. State of the art

For data publication processes, Berners Lee [4] proposed a five-level scheme for linked data. According to this scheme, contexts such as the Open Government Data, have proposed adding an unofficially star in order to classify dataset described using metadata. Strategy that has been implemented in open government data [23, 24].

Framed in LOD principles and schemes, different projects such as: a) "DataHub LOD Datasets" [25], which is a tool that analyzes a specific repository, providing an integrity level of each dataset depending on the metadata description of a specific CKAN instance; b) LOD-Vader [26], which focuses on validating if a dataset is active, representing it graphically by adding its link address.

In addition, tools such as one proposed in [27], use machine learning to determine prediction and scalability models, in order to increase accuracy of their models. This system has scalability characteristics for linked data recommendation systems through parallelization and stacking using MapReduce. Researches such as [28, 29], propose models for electronic acquisitions by data mining, and use of LOD Mining through semi-supervised learning methods, based on machine self-learning, to describe labels on SPARQL resources. As a result of this exploration, it is identified that there are tools for structure and validation reviewing of dataset, but it is evident as a research problem:

- A lack of strategies for visual analytics of open data, which are consumed by instances that manage data, such as those provided by CKAN [30],
- Low standardization in the processing of metadata information, that generates trust problems in linked data;

With the lack of visual analytics tools for data consumption, that collaborate in confidence level determination provided by the three Open Schema levels that are verifiable in published metadata; this inspection and analysis must be done manually, after downloading metadata of CKAN instance. Regarding machine learning [31], it is used in LOD applications, but tools are not identified, which use this technology to determine data reliability or data quality through resource consumption managed by CKAN platform.

Based on this problem, VACIT: Visual Analytics for CKAN Instances [32] is presented in this research. This tool is based on:

- Visual Analytics: rational analysis science supported by a visual and interactive interface [33]. This technique allows decision-making combining human flexibility, creativity and expertise, with enormous capacity of storage and machine process, in order to find solutions to the most complex problems. Therefore, using advanced visual information systems, people can interact with it to make better-informed decisions [34].
- Machine Learning [31, 35], as a method of data analysis, automates analytical model construction. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

These technologies will allow a graphic representation of matedata information, belonging to dataset stored in different CKAN instances. This visualization will support the data analysis, contributing elements for the confidence levels determination, associated with variables characterization of the first 3 linked data scheme levels, and the linking behavior of dataset on selected instances. The methodology used for the construction of this tool is presented in the following section, as well as strategies to carry out element integration that allow to determine visual analysis on dataset selected for the study.

3. Methodology

In order to support the process of determining reliability levels and quality of LOD instances, through resources consumption using CKAN platform, a quasi-experimental methodology is defined. Under this methodology, a methodological design is proposed, whose initial stage (obtaining inputs) consists of: 1) Analysis of services offered by CKAN, 2) Connection strategy design to the CKAN instances, and finally, 3) resource consumption from selected instances. This initial stage allows to obtain information about dataset structure when it is consumed, allowing to generate a standard for dataset exploitation published in CKAN instances. As a later stage, deployment of the proposed tool, poses following four stages (Figure **2**):

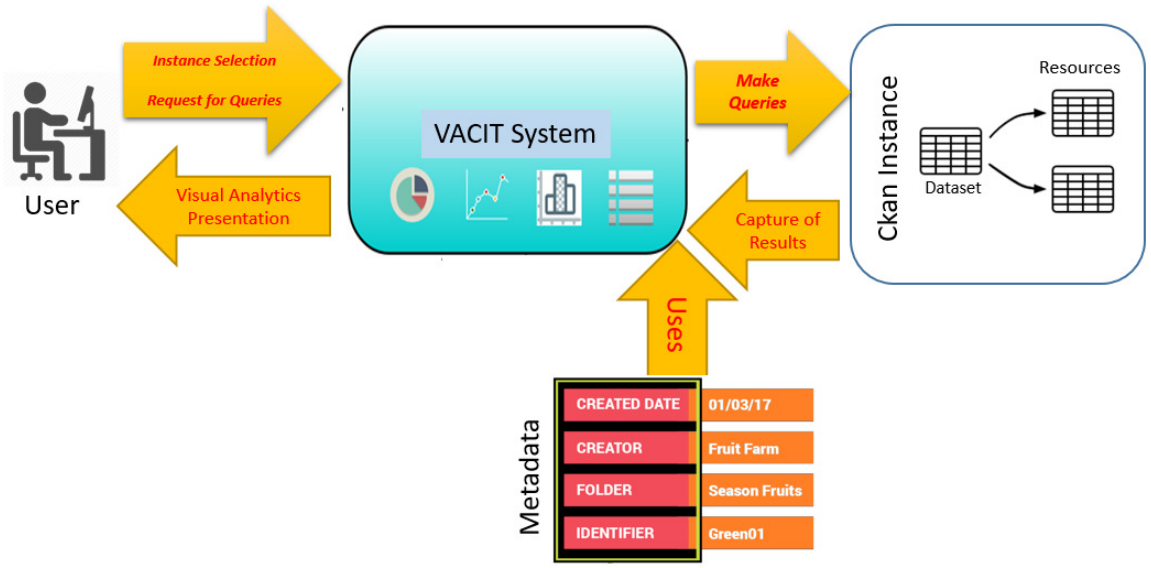
1. Metadata consumption of CKAN instances: Consumption made through use of CKAN API and later data storage.
2. REST Service: This service allows connection between tool front-end and data of the instances.
3. Implementation of Machine Learning module, to evaluate the concordance level of labels: Consumption of unsupervised Machine Learning libraries is developed, determining a concordance

level of tags corresponding to each dataset of an instance.

4. Implementation of Visual Analytics module: This module is built through use and implementation of graphic libraries for representation of metadata analysis, coming from the CKAN instances.

In addition to this design, it is proposed: a) statistical data representation obtained from instances, for visualization and graphic representation of dataset and its resources; b) Testing this tool, through test execution corresponding to consumption and visualization of resources, dataset and instances using a Web application, tool back-end and metadata download process.

Figure 2: Methodological Design.



Source: own.

4. Methodological development

The proposed methodological design works two fronts for VACIT tool implementation: Front- End design and Back-End design.

4.1. *Back – End design*

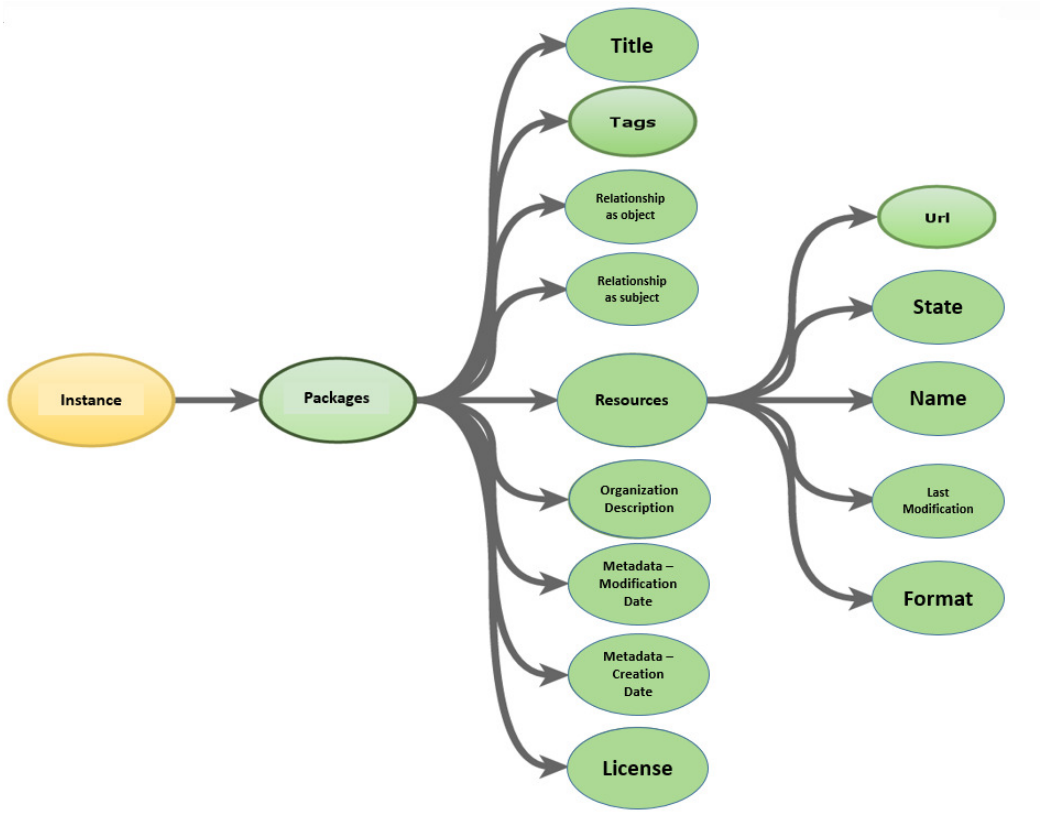
This component carries out CKAN API connection, in order to perform the LOD resources consumption that are published in Ckan instance analyzed. It allows deploying the server so that data can be consumed. Back-End design consists of the following phases:

a. Metadata Download of CKAN Instances: an application was developed in Python, under Linux environment, to download dataset provided by CKAN API. This application allows you to store .json files locally, with data obtained from queries to repositories (Figure 3). Local storage is carried out in order to optimize tool-processing time, since when consuming an instance online; its data quantity would generate an exponential behavior in the response time variable.

Figure 3: Data Consumption of Minnesota Instance [36].

Taking into account changes offered by CKAN API (Version 2.8), data has passed from a SPARQL point [37], to access of instance dataset through a defined structure, which is broken down into a JSON type file [38], which provides data from CKAN instance (Figure 4).

Figure 4: Dataset structure of a CKAN Instance.



Source: own.

b. Creation of REST Service: Deploying a REST server [39], consumption of dataset downloaded previously is done, which will be treated in visualization tool. As shown in Table 1, CKAN API provides methods to consume resources at HTTP level through micro framework for Python Flask [40]. Using this micro framework, CORS headers [41] are assigned to a script to be consumed at Web level. For this purpose, chords and necessary methods for application's deployment were determined. Methods that allow consuming data from the tool's Front-End, are implemented by Python scripts.

c. Implementation of Machine Learning Module for the label association level: In classical machine learning, complexity and divergence are controlled by "black box principle". It is expected that each machine learning method suit a simple mold: the input is a table of instances, described by several characteristics with a target value to predict. The output is a model that predicts the target value [42]. A common vision of the intersection of machine learning and linked data is that machine learning can provide inference when traditional methods based on logic fail [43]. For this research, an automatic learning perspective is adopted: linked data is seen as simply a new form of data.

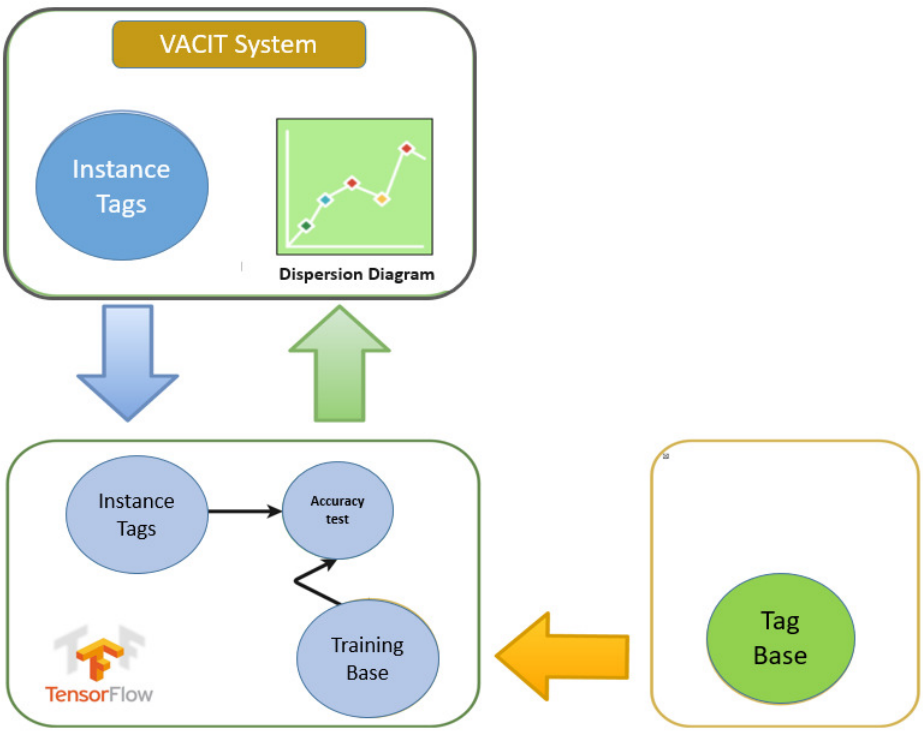
On the other hand, tags are a set of keywords that manage to describe the dataset and specify its content. These tags are a relevant field in LOD resources consumption [23], since they allow relating dataset of different instances with a knowledge domain. For this reason, an unsupervised machine learning module was implemented (Figure 5), which determines labels accuracy level depending on their description. Using of TensorFlow [44], which is a framework for machine learning, proposed model was implemented, based on a specific instance of the platform (training base). This in order to determine the label concordance of other instances with the base instance.

Table 1: Methods of the CKAN API Model.

Resource	Method	Request	Response
Dataset Register	GET		Dataset-List
Dataset Register	POST	Dataset	
Dataset Entity	GET		Dataset
Dataset Entity	PUT	Dataset	
Group Register	GET		Group-List
Group Register	POST	Group	
Group Entity	GET		Group
Group Entity	PUT	Group	
Tag Register	GET		Tag-List
Rating Entity	GET		Rating
Dataset Relationships Register	GET		Pkg-Relationships
Dataset Relationship Entity	GET		Pkg-Relationship
Dataset Relationships Register	POST	Pkg-Relationship	
Dataset Relationships Entity	PUT	Pkg-Relationship	
Dataset's Revisions Entity	GET		Pkg-Revisions
Revision List	GET		Revision-List
Revision Entity	GET		Revision
License List	GET		License-List

Source: own.

Figure 5: Machine Learning Module - Label Matching Level.



Source: own.

As a training base, Datahub was chosen [45]. This platform has the greatest richness in the description of its domain labels. Additionally, Machine Learning model implemented has been an algorithm based on principal component analysis (PCA), which is applied to determine level of label association defined by CKAN, comparing these labels with those defined by each author in a dataset belonging to a specific instance. This algorithm was divided into the following phases:

- Correlation matrix Analysis: It allows determining if there are high correlations between the variables, since this is indicative of redundant information and, therefore, few factors will explain much of total variability between each label with respect to the training base.
- Factors Selection: It allows determining outliers or redundancy of labels, as well as descriptions of labels that do not provide information that allows relating an instance with its resources. On the other hand, labels that have relevant information or a valid label will have a higher value than any one before mentioned.
- Factorial matrix analysis: Once the main components are selected, they are represented as a matrix. Each element represents the factorial coefficients of the variables (the correlations between variables and main components). This matrix will have as many columns as main components and as many rows as datasets belonging to a specific instance.
- Factors Interpretation: A factor will be interpretable if it must have the following characteristics: a) The factorial coefficients must be close to 1, b) A variable must have high coefficients with only one factor, and c) There should be no factors with similar coefficients.
- Factorial scores calculation, they are calculated by the expression (equation 1):

$$X_{ij} = a_{i1} * z_{1j} + \ldots + a_{ik} * z_{kj} = \sum_{s=1}^{k} a_{is} * z_{sk} \quad (1)$$

The variables a correspond to coefficients and variables Z corresponds to standardized values that variables have in each of sample subjects.

4.2. *Front – End design*

This module is also called the visualization component. It is responsible for interacting with user, providing graphical interface for selecting instances and Visual Analytics modules, which perform statistical processes to represent information of each dataset. For this purpose, implementation of Visual Analytics Module was carried out, in order to provide tools to support identification of trust in the published dataset in each instance. For this purpose, VACIT uses bar charts, pie charts, data tables and other elements of visual analytics, in components considered relevant in the dataset richness: organization description, Author, Licenses, Dataset resource formats, Relationships as object and subject, and Links of resources. Angular framework is used to carry out this implementation [46], and Javascript libraries are consumed, which allow a data representation in a graphic form, after a statistical process, such as: Percentage of omitted Authors data, Quantity of resources with a specific format, Type of licenses of each Dataset, Dispersion plot of dataset / tags concordance.

4.3. *Tool architecture*

Components implemented inside the VACIT tool are described in Figure 6.

The minimum technological environment used for the implementation of these components is presented in Table 2.

Table 2: Software requirements for VACIT Tool.

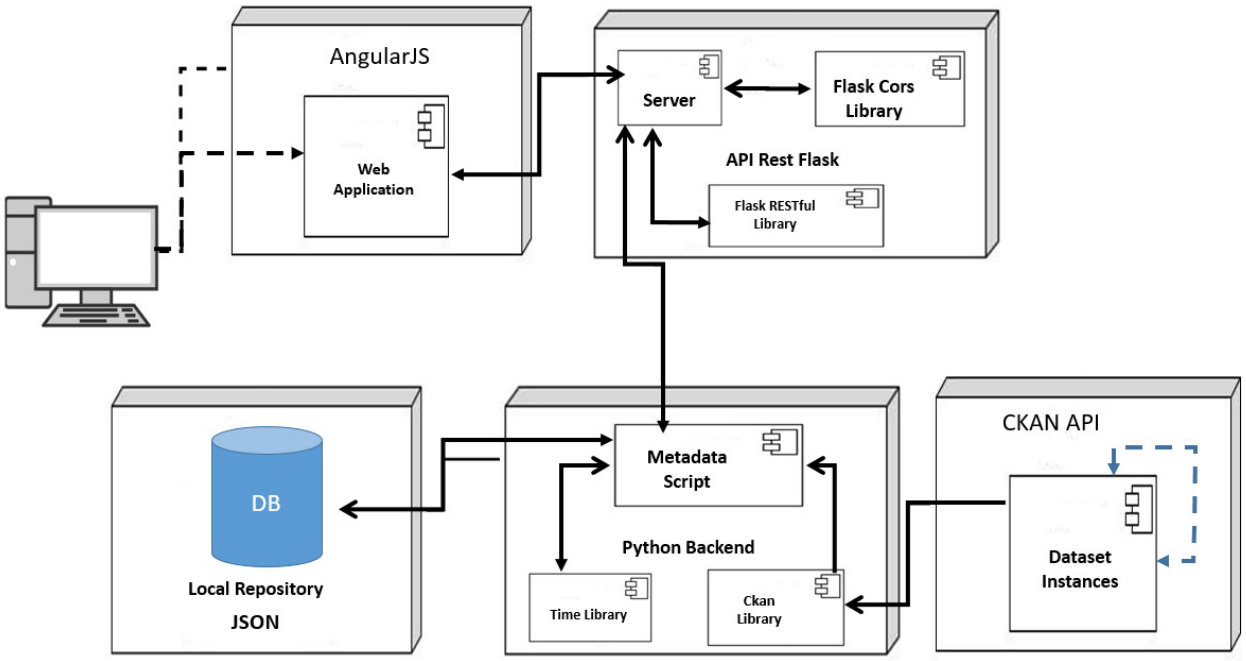
Technology	Version
Python	>=2.7 ó >=3.5
Pip (python framework)	8.0.1
Termcolor (python library)	Any version
Flask (python library)	Any version
Flask - Cors (python library)	Any version
Flask - restful (python library)	Any version
Nodejs	8.0 or higher
Angular	2.4 or higher

Source: own.

5. Obtained results

As it was described in the problem identification, a lack of tools to carry out a systematic and analytical study of ckan instances was determinate. Normally, this process is done manually, according to the description fields provided and their content, a process that can be extensive and cumbersome as there are more dataset and more resources within these datasets.

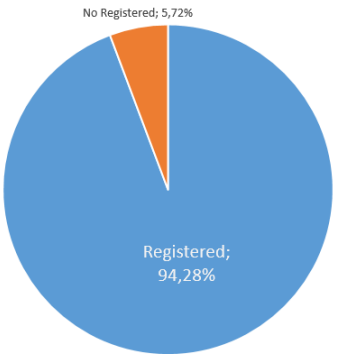
Figure 6: Architecture of the VACIT Tool.



Source: own.

To carry out this analysis, it starts with the use of a VACIT tool. This tool makes a representation of the information using statistical diagrams. These diagrams allow analysis process and decision-making using levels of trust, linkage, usability and quality levels of each open data provided by the organizations, which are attached to CKAN. As examples of the visual analytics module, a cake diagram is presented in Figure 7, which is resulting from statistical analysis for a specific CKAN instance. It shows the amount of dataset that have information of Author(s) in your description.

Figure 7: Authorship Percentage for a specific instance.

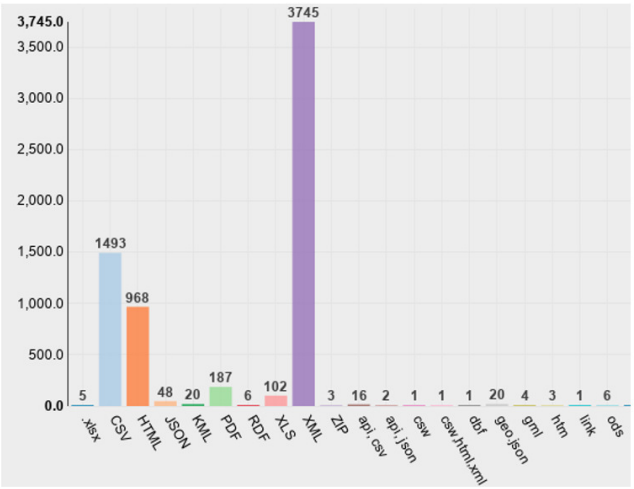


Source: own.

Different resource formats uploaded in a specific instance, with the number of resources with the same format, is shown in a bar plot in Figure 8.

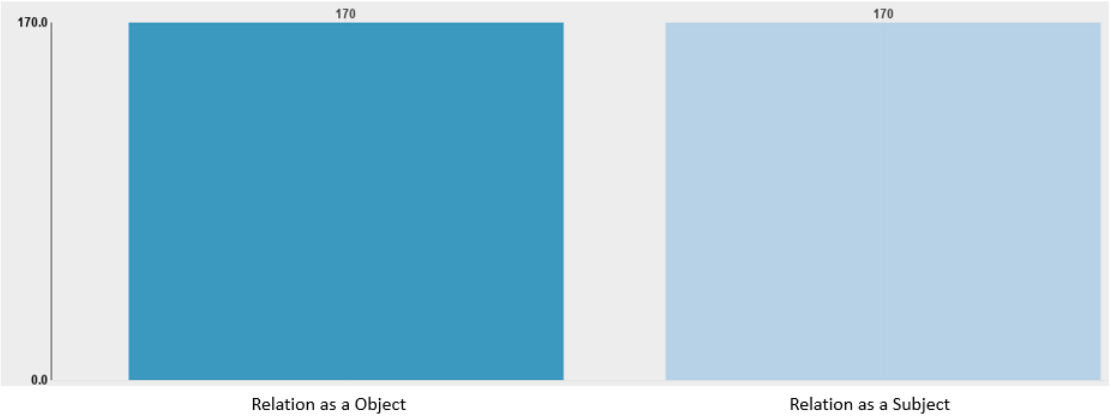
The number of relations as object (left bar) and relations as subject (right bar), from datasets belonging to the analyzed instance, are shown as bar chart in Figure 9.

Figure 8: "Resource Format" Query in an instance.



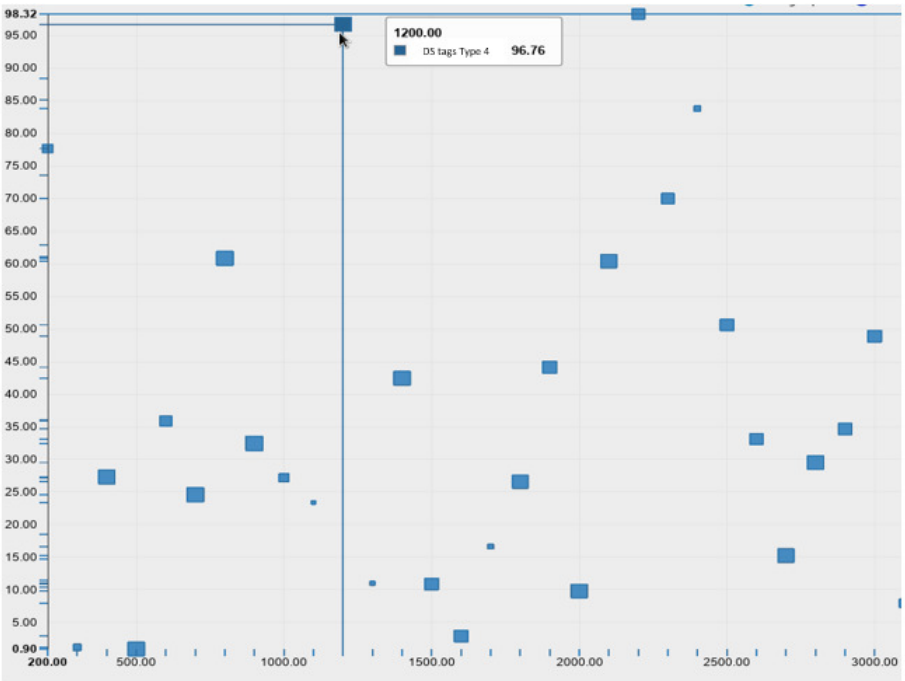
Source: own.

Figure 9: Relationships as subject and object of an instance.



Source: own.

Figure 10: Conformance level of the domain labels of each dataset belonging to a CKAN instance.



Source: own.

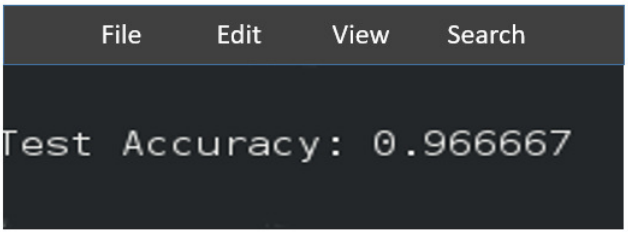
Making use of standardization in resource consumption, made by the platform, in specific case, consumption of domain labels of dataset; a dispersion plot can be generated, which indicates the relationship level of these labels related to other labels already defined in the platform. The process of determine the concordance level of each dataset, generates an output that Figure 10 is shown.

Dispersion plot represents relationship between concordance percentage of a dataset belonging to an instance (Y-axis) and each dataset identifier (X-axis). Figure size (blue square) is represented by the amount of resources that has dataset. However, this figure is omitted since it does not have a relevant role in the concordance level determination. This means that the further away from the X-axis a dataset is found, the higher the concordance level of its domain labels with

respect to the training base.

This dispersion plot represents graphically the evaluation process of concordance level of all dataset belonging to an instance. Console response of the Machine Learning module is shown in Figure 11, which corresponds to a specific dataset evaluation. With this diagram, it is possible, for those who use this tool and perform a visual analytic process for a specific instance, dataset richness evaluation and identification of one or more domains for which dataset has some relationship types.

Figure 11: Console-level output of a dataset match.



Source: own.

6. Result analysis and discussion

VACIT Back-End is aimed to obtain the LOD resources of instances in a local repository. These data are stored in .json files (Figure 12), which will be consumed by the Front-End tool, where they are ordered and structured.

Data consumed representation allows them to be readable and descriptive, based on raw data (Json). Data visualization process in a structured and orderly manner allows information of both metadata and resources attached to dataset to be useful for tool user. Dataset after use of its data in the visual analytics module of VACIT tool is shown in Table 3.

According to the review carried out, tools that implement LOD and Machine Learning, have different objectives to the data validation in management data platforms such as CKAN. These tools generally focus on determining certain types of dataset behaviors, such as tool [25]. This tool allows classifying the dataset in a conformance scheme between levels 1-4 for the Datahub instance. This process limits the tool performance to determine the compliance level to a single instance of Ckan, while VACIT tool is able to consume resources from all organizations joined to Ckan.

Similarly, tool [26] has a dataset status analysis (active or inactive) and subsequently allows display of labels and the number of resources attached to this dataset along with their links. In the same way, tool [25] only use Datahub instance, therefore it is limited considerably, taking into account that currently organizations attached to Ckan exceed 140. Otherwise, VACIT tool offers more options, except the relationship between datasets based on their labels, but considerably exceeding the number of instances, metadata description and resources of each dataset.

Figure 12: JSON segment with the metadata of a Ckan instance.


```
"packages": [
    {
        "author": "",
        "license_title": "Creative Commons Attribution",
        "metadata_created": "2016-11-28T15:09:35.337776",
        "metadata_modified": "2017-05-29T14:57:56.805841",
        "organization_description": "",
        "relationships_as_object": [],
        "relationships_as_subject": [],
        "resources": [
            {
                "format": "XLS",
                "hash": "",
                "last_modified": "2017-05-29T14:57:56.781763",
                "name": "Estructura organigrama funcional - julio 2015
                "state": "active",
                "url": "http://catalogo.datosabiertos.gob.ec/dataset/6
            },
```


Source: own.

Table 3: Structured information of an instance.

#	Title	Author	License	Creation Date	Modification Date	Organization Resources
1	Estructura orgánica funcional - Servicio Integrado de Seguridad ECU 911		Creative Commons Attribution	2016-11-28T15:09:35.337776	2017-05-29T14:57:56.805841	3
2	Regulaciones y procedimientos internos - Servicio Integrado de Seguridad ECU 911		Creative Commons Attribution	2016-11-28T20:30:21.984211	2017-05-29T14:56:38.054147	3
3	Directorio de la institución – Servicio Integrado de Seguridad ECU 911		Creative Commons Attribution	2016-11-28T20:46:14.253485	2017-05-29T14:54:22.689295	3
4	Distributivo del personal - Servicio Integrado de Seguridad ECU 911		Creative Commons Attribution	2016-11-28T20:55:06.480719	2017-05-29T14:53:29.995924	3
5	Puntos de Atención - Servicio Integrado de Seguridad ECU 911		Creative Commons Attribution	2016-11-28T21:11:45.369360	2017-05-29T14:51:33.351593	3
6	Presupuesto Anual - Servicio Integrado de Seguridad ECU 911		Creative Commons Attribution	2016-11-28T21:31:07.342435	2017-05-29T14:45:42.789255	3
7	Planes y programas en Ejecución - Servicio Integrado de Seguridad ECU 911		Creative Commons Attribution	2016-11-25T19:41:05.065710	2017-05-29T14:40:52.121435	7

Source: own.

On the other hand, few tools are transversal to Linked Open Data and Machine Learning. There is only one tool [29], which intends to work with LOD resources, which are uploaded in CKAN. This tool implements data mining and semi-supervised learning to obtain dataset that relate to each other. However, it focuses on the LOD-CLOUD platform for the dataset consumption. This platform is linked to Datahub instance and therefore only makes queries about this instance, which is why VACIT tool continues to be more productive because of its use for many instances of Ckan.

Other tools of this style do not focus on the platform analysis, instances, dataset or LOD resources. Those tools were created for commercial purposes such as the tool [27, 28] that are used for recommendation and acquisition systems electronic respectively. For that reason, VACIT turns out to be a practical, complete and transversal tool for LOD resources analysis, which goes beyond analyzing only an instance or a specific field of its metadata such as domain labels, authors, resource format, among others. However, it covers and uses all the metadata offered by Ckan instances, allowing user to perform a thorough analysis of each instance, dataset or resource, determine levels of trust, quality, usability, linkage and other items that may offer indicators of this type for LOD resources.

7. Conclusion and future work

Confidence and data quality are two closely related factors in the consumption of LOD resources. These factors are direct references for those who aspire to consume these resources. CKAN has evolved widely in the last years, achieving advances in linked data consumption and publication, providing interfaces and APIs that facilitate and standardize these processes, for instance. However, there is still a slow process in delivering tools, which allow a graphic evaluation of linked process variables, carried out by the instances.

Determining of the trust level of resources published through LOD is a process that, in its origins, is concentrated and affected by the processes carried out in the metadata modeling and description phases corresponding to each dataset, both in its descriptions and usability richness, and in its link.

Following this approach to generating elements, in order to establish trust criteria, this research provides as a key component for users, a tool with agreed characteristics for determination of mentioned levels. This tool facility the analysis processes at either instance level (dataset), where user who can be supported in the different modules of the VACIT tool, in order to establish

indicators of linked, trust, usability or other aspects that are considered relevant for the consumption of these resources.

Regarding to the obtained indicators, problems about linking process can be evidenced. As a result, only two of the queried instances through the tool handle linking processes, either as subject or as object. This allows us to show that linked resources in these instances does not have an adequate treatment yet, not to allowing advance in the last two levels of the linked data scheme.

On the other hand, when different visual analytics elements are applied to the queried instances, results such as a proliferation of publication formats, little standardization in the use of licensing, etc., are observed. Factors that make it possible to demonstrate that, even though there are recommendations and good practices for the linking of resources, there is still a way to go to achieve the adequate compliance of the proposed characteristics by both Open Data and LOD.

In this context, the VACIT tool is expected to become a focal point for the creation of new modules, which contribute to the insertion of new indicators, as well as linking other platforms for LOD resources publication and consumption. As well as studies and researches on the CKAN platform, in order to increase methods and indicators of different levels of trust, usability, provenance, quality and linked open data, given that they are necessary due to the exponential growth of the semantic web [4].

Machine learning module abstraction, in terms of language, transcending the main language (English) is proposed as future work it. This process allows each dataset can be related to others that do not belong to the same instance, but are under the same domain tag. Additionally, as mentioned in the implementation of the Machine Learning module, the concordance level works on a training base, which was analyzed previously and obtained from a specific instance [29]. This base can be generalized by performing a statistical study on all the domain labels used by the platform, matching these labels with other languages for its simplification, in order to obtain results that are more accurate.

8. Acknowledgments

This research has been developed within the framework of the doctoral research project in Linked Data, at the Universidad Distrital Francisco José de Caldas. In the same way, the issue is being worked as a research line of the GIIRA Research Group.

References

[1] BCN, "Linked Open Data: ¿Qué es?", https://datos.bcn.cl/es/informacion/que-es

[2] Open Knowledge International, "Open Data HandBook", 2018. [Online]. Available at: http://opendatahandbook.org/

[3] C. Caicedo, "Virtualización Organizacional, Web Semántica y Redes Sociales", *Visión Electrónica*, vol. 6, no. 2, 2012, pp. 134-159. https://doi.org/10.14483/22484728.3894

[4] T. Berners, J. Hendler and O. Lassila, "The Semantic Web", 2001. *Scientific American*, vol. 284, no. 5, 2001, pp. 29-37. [Online]. Available at: https://www.scientificamerican.com/article/the-semantic-web/

[5] T. Berners-Lee, C. Bizer and T. Heath, "Linked Data - The Story so Far", *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, 2009, pp. 1-22. https://doi.org/10.4018/jswis.2009081901

[6] C. Bizer and T. Heath, "Linked Data. Evolving the Web into a Global Data Space", Morgan & Claypool Publishers, 2011. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

[7] M. Schmachtenberg, C. Bizer and H. Paulheim, "State of LOD Cloud", 2014. [Online]. Available at: http://stats.lod2.eu/

[8] LOD2, "Version 1.0 of the LOD2", 2014. [Online]. Available at: http://lod2.stat.gov.rs/lod2statworkbench

[9] M. López-Bonilla, "Semántica para repositorios de Objetos de Aprendizaje", *Scientia Et Technica*, vol. 19, no. 4, 2014, pp. 425-432. http://dx.doi.org/10.22517/23447214.9292

[10] R. Melero, "Repositorios", 2014. [Online]. https://ucrindex.ucr.ac.cr/docs/repositorios_2014.pdf

[11] CKAN, "API Guide", 2018. [Online]. Available at: http://docs.ckan.org/en/latest/api/

[12] J. Winn, "Open Data and the Academy: An Evaluation of CKAN for Research Data Management", 2013. [Online]. Available at: http://eprints.lincoln.ac.uk/9778/1/CKANEvaluation.pdf

[13] J. Herrera-Cubides, P. Gaona-García and S. Sánchez-Alonso, "Linked Data: Qué sucede con la Heterogeneidad y la Interoperabilidad", *Scientia et Technica*, vol. 23, no. 2, 2018, p.p. 230-240. http://dx.doi.org/10.22517/23447214.16791

[14] B. Farias, C. Burle and N. Calegari, "Data on the Web - Best Practices", 2017. [Online] Available at: https://w3c.github.io/dwbp/bp.html

[15] E. Ruckhaus, M. Vidal, S. Castillo, O. Burguillos and O. Baldizan, "Analyzing Linked Data Quality with LiQuate", The Semantic Web: ESWC 2014 Satellite Events. ESWC, 2014. https://doi.org/10.1007/978-3-319-11955-7_72

[16] J. Herrera-Cubides, P. Gaona-García and K. Gordillo-Orjuela, "A View of the Web of Data. Case Study: Use of Services CKAN", *Ingeniería*, vol 22, no. 1, 2017, pp. 111- 124. https://doi.org/10.14483/udistrital.jour.reving.2017.1.a07

[17] E. Rajabi, S. Sanchez-Alonso and 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, 2014, pp. 1721–1727. https://doi.org/10.1002/asi.23109

[18] C. Bizer, "The Emerging Web of Linked Data", *IEEE Intelligent Systems*, vol. 24, no. 5, 2009, pp. 87-92. https://doi.org/10.1109/MIS.2009.102.

[19] P. Gaona-García, J. Herrera-Cubides, J. Alonso-Echeverri, K. Riaño-Vargas and A. Gómez-Acosta, "A Fuzzy Logic System to Evaluate Levels of Trust on Linked Open Data Resources",*Revista Facultad de Ingeniería*, no. 86, 2018, pp. 40-53. http://dx.doi.org/10.17533/udea.redin.n86a06

[20] E. Arias-Caracas, D. Mendoza-López, P. Gaona-García, J. Herrera-Cubides and C. Montenegro-Marín, "Evaluation of the Linked Open Data Quality Based on a Fuzzy Logic Model", Artificial Intelligence Applications and Innovations, 2018. https://doi.org/10.1007/978-3-319-92007-8_47

[21] Lod Cloud, "The Linked Open Data Cloud", 2011. [Online]. Available at: https://lod-cloud.net/.

[22] Linked Science, "Tutorial on Visual Analytics with Linked Open Data", 2014. [Online]. Available at: http://linkedscience.org/events/vislod2014/

[23] T. Berners, "Linked Data", 2006. [Online]. Available at: https://www.w3.org/DesignIssues/LinkedData.html

[24] R. Ávila-Alonso, "Aplicación de los Principios Linked Open Data a la lista de encabezamientos de materia de la Biblioteca de la Universidad Politécnica de Madrid", thesis MSc., Universidad Carlos III de Madrid, Spain, 2014.

[25] Loud Cloud, "Data Hub LOD Datasets", 2012. [Online]. Available at: http://validator.lod-cloud.net/index.php

[26] C. Baron-Neto, K. Müller, M. Brümmer, D. Kontokostas and S. Hellmann, "Lodvader: an interface to LOD visualization, analytics and discovery in real-time", 25th WWW Conference, 2016. https://doi.org/10.1145/2872518.2890545

[27] J. Ruhland and L. Wenige, "Scalable Property Aggregation for Linked Data Recommender Systems", 3rd International Conference on Future Internet of Things and Cloud, 2015. https://doi.org/10.1109/FiCloud.2015.30

[28] E. Loza-Mencía, S. Holthausen, A. Schulz and F. Janssen, "Using data mining on Linked Open Data for analyzing e-procurement information", DMoLD'13 Proceedings of the 2013 International Conference on Data Mining on Linked Data, vol. 1082, 2013, pp. 50-57. [Online]. Available at: http://ceur-ws.org/Vol-1082/paper4.pdf

[29] N. Fanizzi, C. d'Amato and F. Esposito, "Mining linked open data through semi- supervised learning methods based on self-training", IEEE Sixth International Conference on Semantic Computing, 2012. pp. 277–284. https://doi.org/10.1109/ICSC.2012.54

[30] Ckan, "What is ckan? User guide", 2018. [Online]. Available at: http://docs.ckan.org/en/latest/user-guide.html#what-is-ckan

[31] Priyadharshini, "Machine Learning: What it is and Why It Matters", 2018. [Online]. Available at: https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters\-article.

[32] A. Varón, J. Herrera-Cubides and P. Gaona-García, "VACIT - Visual Analytics for CKAN Instances Tool (Herramienta de Software)", Universidad Distrital Francisco José de Caldas, Bogotá, Colombia, 2018.

[33] J. Thomas and K. Cook, "Illuminating the Path: Research and Development Agenda for Visual Analytics", 2005. [Online]. Available at: http://vis.pnnl.gov/pdf/RD_Agenda_VisualAnalytics.pdf

[34] IEBS, "¿Que es el Visual Analytics?", 2005. [Online]. Available at: https://comunidad.iebschool.com/visualanalyticsbusinessintelligencebigdata/que-es-el-visual-analytics/.

[35] J. E. Rodríguez and J. Ortiz-Pimiento, "Métodos bayesianos para la clasificación de páginas Web inapropiadas", Visión Electrónica, vol. 11, no. 2, 2017, pp. 179-189. https://doi.org/10.14483/22484728.13135

[36] Minnesota Geospatial, "About the Minnesota Geospatial Commons", 2018. [Online]. Available at: https://gisdata.mn.gov/content/?q=about.

[37] W3, "SPARQL Endpoint", 2011. [Online]. Available at: https://www.w3.org/wiki/SparqlEndpoints

[38] ECMA. "The JSON Data Interchange Syntax", 2017. [Online]. Available at: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf.

[39] R. Thomas-Fielding, "Representational State Transfer (REST)", 2000. [Online] Available at: https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

[40] The Pallets Projects, "Flask", 2018. [Online] Available at: https://www.palletsprojects.com/p/flask/.

[41] Fetch, "Living Standard 2.1 Methods", 2018. [Online] Available at: https://fetch.spec.whatwg.org/

[42] P. Bloem and G. de Vries, "Machine Learning on Linked Data, a Position Paper", Proceedings of the First International Conference on Linked Data for Knowledge Discovery, vol. 1232, 2014, pp. 64-68. https://dl.acm.org/citation.cfm?id=3053834

[43] A. Rettinger, U. Losch, V. Tresp, C. d'Amato and N. Fanizzi, "Mining the Semantic Web statistical learning for next generation knowledge bases", Data Mining and Knowledge Discovery, vol. 24, no. 3, 2014, pp 613-662. https://doi.org/10.1007/s10618-012-0253-2

[44] Tensorflow, "API Documentation", 2018. [Online] Available at: https://www.tensorflow.org/api_docs/.

[45] A. Kariv and R. Pollock, .ᴬbout Datahub", 2018. [Online] Available at: https://datahub.io/docs/about

[46] Angular, "What is angular?", 2018. [Online]. Available at: https://angular.io/docs