UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Visión Electrónica
Más que un estado sólido
https://revistas.udistrital.edu.co/index.php/visele/index

Visión Electrónica
A RESEARCH VISION

Abstraction of linked data's world

Abstracción del mundo de linked data

Jhon Francined Herrera-Cubides[1]; Paulo Alonso Gaona-García[2];
Carlos Enrique Montenegro-Marín[3]; Salvador Sánchez-Alonso[4]; David Martin-Moncunill[5]

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Enviado: 05/12/2018
Recibido: 10/12/2018
Aceptado: 18/12/2018

Keywords:
Language
Linked Data
Metadata
Model
UML
Vocabulary

ABSTRACT:

Linked Data, as a strategy of the Semantic Web, is based on application of some basic principles that contribute to the growth of the Web, thus allowing the transit of the Web of Documents to the Web of Data. Developed process by Linked Data is supported in different scenarios, which interact in order to carry out the linking of resources on the Web. Some of these scenarios present a solid technological background, while others propose challenges when they are implemented. This paper aims to identify and expose a generic abstraction of Linked Data, in order to identify problem situations that restrict Linked Data process.

Palabras clave:
Lenguaje
Linked Data
Metadatos
Modelo
UML
Vocabulario

RESUMEN

Linked Data, como estrategia de la Web Semántica se fundamenta en la aplicación de unos principios básicos que contribuyen al crecimiento de la Web permitiendo así el tránsito de la Web de los Documentos a la Web de los Datos. El proceso desarrollado por Linked Data se soporta en diferentes escenarios que interactúan con el fin de llevar a cabo la vinculación de recursos en la Web. Algunos de estos escenarios presentan un background tecnológico bastante sólido, mientras que otros plantean desafíos al momento de ser implementados. El presente artículo se orienta en identificar y exponer una abstracción genérica de Linked Data, con el fin de identificar situaciones problema que restringen el proceso de vinculación de los datos.

1 Ph.D. (c) In Engineering, Universidad Distrital Francisco José de Caldas, Colombia. Current position: Professor at Universidad Distrital Francisco José de Caldas, Colombia. E-mail: jfherrerac@udistrital.edu.co. ORCID: https://orcid.org/0000-0003-1615-4656.
2 Ph.D. In Information and Knowledge Engineering, Universidad de Alcalá, España. Current position: Professor at Universidad Distrital Francisco José de Caldas, Colombia. E-mail: pagaonag@udistrital.edu.co. ORCID: https://orcid.org/0000-0002-8758-1412.
3 Ph.D. In Systems and informatics services for Internet, Universidad de Oviedo, España. Current position: Professor at Universidad Distrital Francisco José de Caldas, Colombia. E-mail: cemontenegrom@udistrital.edu.co. ORCID: https://orcid.org/0000-0002-3608-7158.
4 Ph.D. In Informatics, Universidad Politécnica de Madrid, España. Current position: Professor at Universidad de Alcalá, España. E-mail: salvador.sanchez@uah.es. ORCID: https://orcid.org/0000-0002-9949-4797.
5 Ph.D. In Information and Knowledge Engineering, Universidad de Alcalá, España. Current position: Professor, Universidad de Alcalá, España. E-mail: david.martin@uah.es. ORCID: https://orcid.org/0000-0003-2422-9005.

Cite this article as: J. F. Herrera-Cubides; P. A. Gaona-García; C. E. Montenegro-Marín; S. Sánchez-Alonso; D. Martin-Moncunill, "Abstraction of Linked Data's world", Visión electrónica, algo más que un estado sólido, vol. 13, no. 1, january-june 2019, pp. 57-74, DOI: https://doi.org/10.14483/22484728.14397

1. Introduction

Linked Data, as a set of design principles for sharing interconnected data, readable by machines, generally focuses on tools that provide meaning to data such as microdata, RDFa or RDF, and ontologies that provide meaning to the terms. To achieve its purpose, it uses a 5-star scheme, which defines requirements to be met in order to advance data publication process [1]. With evolution of Linked Data, an increasing number of repositories publishing Dataset and metadata have been generated [2, 3, 4], as well as providing services for their exploitation.

To carry out this evolution, Linked Data involves a process framed by a set of good practices, among which are detailed [5]:

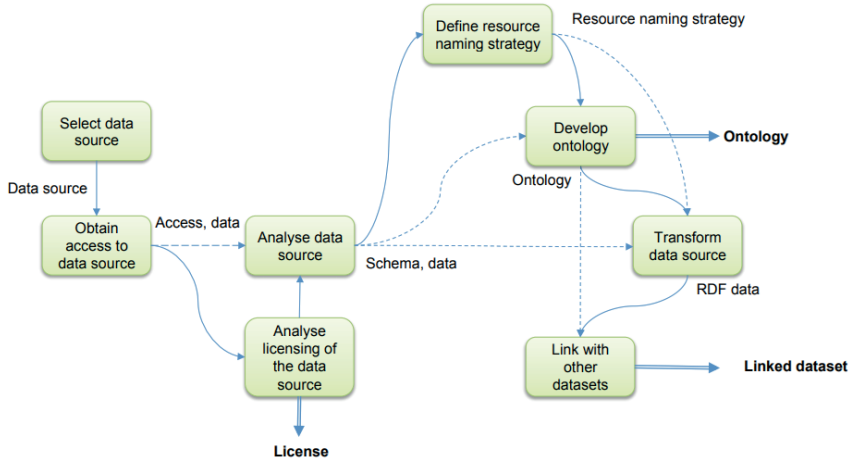
- Prepare the stakeholders: Creation and maintenance process of Linked Data is explained.
- Select a Dataset: Select a dataset that provides benefits to others for reuse.
- Data Model: Data model involves representing data objects, and how they are related in a way that is independent of the application.
- Specify an appropriate license: Data reuse is more likely to occur when there is a specific license about the origin, ownership and terms related to use of published data.
- URI for Linked Data: Consideration for naming objects, multilingual support, and change of data over time and persistence strategy are basic components of useful Linked Data.
- Use a standard vocabulary: Objects are described with previously defined vocabularies whenever possible. Standard vocabularies are expanded when necessary and create vocabularies (only when necessary) that follow best practices whenever possible.
- Convert data: Data is converted to a Linked Data representation.
- Provide access to the machine: Provide several ways for search engines and other automated processes to data access through standard Web mechanisms.
- Announce new Dataset: Remember to announce new datasets in an authorized domain.

For the application of these good practices it requires knowledge about the process and the context of how Linked Data operates (Figure 1):

Regarding this, previous background allows identifying and contextualizing the interaction of different components that participate in the process of publishing resources on the Web. In order to carry out the identification of these interactions, a meta-model abstraction, which allows identifying components and constructing points of view related to their behavior, is considered.

In general, the research questions proposed for this study are: i) How is the behavior of different points of view in the generic context of Linked Data, and ii) What are the challenges under the points of view identified in Linked Data Context?. This research attempts to answer these questions, under a descriptive documentary approach, in order

Figure 1. Linked Data Generation Process [6].



to provide a better understanding of Linked Data process, in addition to identifying shortcomings with respect to the points of view analyzed.

In order to achieve this purpose, this paper is organized as follows: In section 2 background is described, and references of the proposed theme are reviewed. Subsequently, in section 3 methodology and methodological design used to explore the Linked Data context, are exposed. In section 4, methodological development is exposed. Section 5 discussions and conclusions are presented.

2. Background

In order to contextualize the concepts and vocabularies abstraction, proposed model in this paper is defined, a general outline of key elements in Linked Data domain is presented in this section.

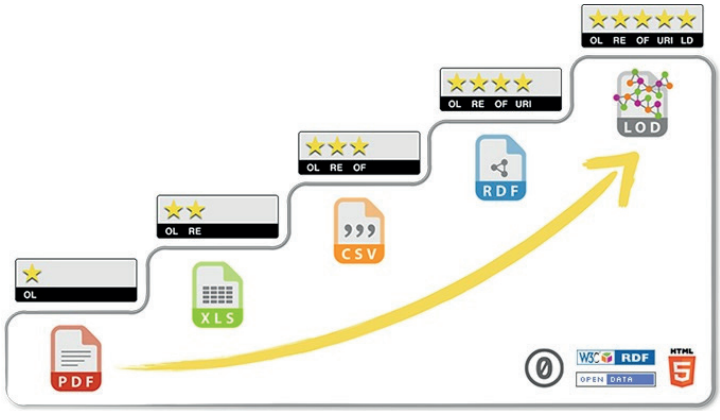
2.1.Linked Data

Semantic Web [7], within its technologies makes use of Linked Data [8], as a strategy to link, relate and query data on the Web, which can be located in different sources; improving its level of utility through semantic queries. This strategy is based on Web technologies such as HTTP–Hypertext Transfer Protocol [9], RDF–Resource Description Framework [10] and URI–Uniform Resource Identifier [11], which share information in a comprehensible way by machines [12]. In 2009, Tim Berners-Lee presented Linked Data Principles [13], among which are:

- **Use URIs to name things or conceptual objects:** If you cannot identify a thing, you cannot talk about it. The underlying technology used for this process are URIs [11].
- **Use HTTP URIs that can be interpreted by humans and machines**: Just as an international standard number (ISBN) commonly identifies books, a URI could be homologated to an ISBN. Now, when writing that URI in the browser, it will inform you that it does not know how to handle the situation. Hence, it is necessary that URIs are resolvable in the Web, therefore HTTP URI is used.
- **Provide useful information about each URI in some standard of the Web (e.g. RDF)**: Any HTTP URI can be written in a Web browser and the browser will know what to do with it (for example: determine the host number, the port to use, etc.). If the remote server responds affirmatively, it will return a representation of the resource in different formats such as RDF [10], among others. Either way, it would be desirable for URIs to be able to resolve some useful descriptions about what you have named.
- **Create links between URIs**: In the same way that Web pages are more useful if they contain links to related information, data is more useful if related data, documents and descriptions are linked. Since HTTP URI was used to publish your data, other people can link your data. The ability to follow these links allows people to browse through the Web of Data just as they can navigate the Web of Documents.

When Linked Data principles are implemented, a 5-level scheme is parameterized [14] (Figure 2). Published resources have a set of restrictions, which place them in a specific category, until they are linked to the LOD Cloud [2]:

Figure 2. Five levels schema of Linked Data [1].



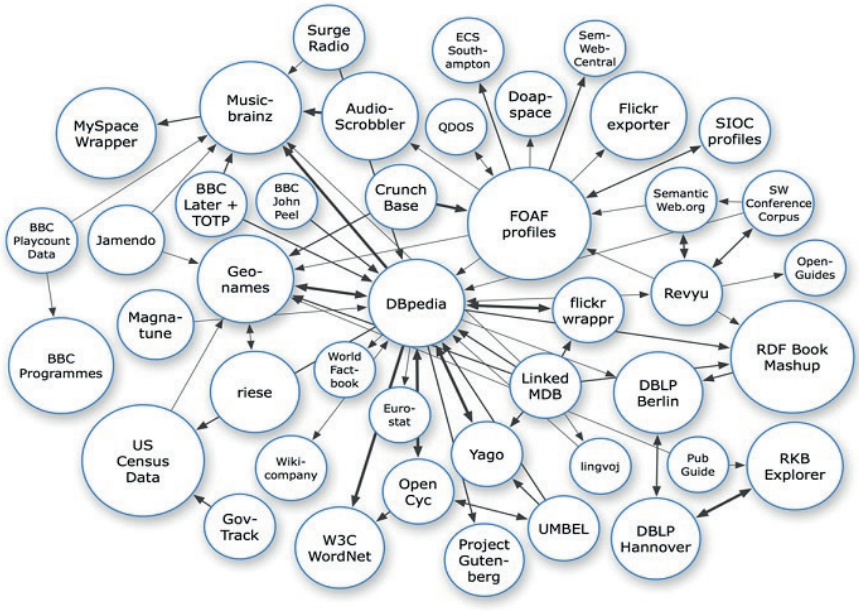
- Level 1: Publish data on the Web, described in some type of format, under an open license [15, 16]. Publish data in any format, even if they are difficult to manipulate or disaggregate, such as a pdf or a scanned image.
- Level 2: Data is published as structured data. Some formats that can be processed automatically are Excel, CSV–comma-separated values [17], XML–eXtensible Markup Language [18], JSON–JavaScript Object Notation [19]. The use of non-editable formats such as PDF should be avoided.
- Level 3: Non-proprietary formats are used for publication. Open formats are suitable for tools and applications that any user can use directly: CSV [17], RSS [20], XML [18], JSON [19], RDF [21].
- Level 4: URI to identify things and properties are used; so that you can link data. For this process, use of RDF standard is required [21].
- Level 5: Data is linked to other data. As a result, .sparql files are attached. In practice, it allows information delivered to also link other data sources. For example, if a country is placed within the data, a reference URI can be added that links data of that country in Dbpedia [22, 23] or to Geonames [24].

Changing from one level to another in Linked Data hierarchy requires a set of processes and technologies, such as modeling and exporting data, understanding and filling out data format used, identifying and using patterns to data representation, among others. In this context, the process developed by Linked Data has had a significant growth, as observed in different knowledge graphs published on the Web. Figure 3, shows one of those graphs.

Dataset are connected to each other, thus composing the "Linked Data Cloud". To become part of this graph, you must [2]:

- URIs http: // (or https: //) that can be resolved must be used.
- Should be resolved with or without "content negotiation", to RDF data in any of its formats (RDFa, RDF / XML, Turtle, N-Triples).
- Dataset must contain at least 1000 triples. (It is likely that your FOAF file does not classify)
- Dataset must be connected through RDF links to another Dataset that is already in the graph. This means that a Dataset must use URI from another Dataset. Arbitrarily, at least 50 links are required.
- Access to the Dataset must be possible through RDF crawling, through an RDF dump or through a SPARQL Endpoint.

Figure 3. Knowledge graphs [2].



2.2.Linked Open Data–LOD

Taking in account that knowledge graphs are built on Open Data, such data can be used, reused and redistributed freely by any person, and are subject to the requirement of attribution and of sharing in the same way they appear [14, 25]. As stated [25], the most important opening aspects are summarized in the following details:

- Availability and Access: information must be available as a whole and at a reasonable cost of reproduction, preferably by downloading it from the Internet. In addition, information must be available in a convenient and modifiable form.
- Reuse and redistribution: data must be provided under terms that allow reuse and redistribution, and even integrate with other datasets.
- Universal participation: everyone must be able to use, reuse and redistribute information. There must be no discrimination in terms of effort, people or groups. "Non-commercial" restrictions that would prevent commercial use of data; or restrictions of use for certain purposes (for example only for education) are not allowed.

Interoperability is one of the main aspects of open data, which denotes ability of diverse systems and organizations to work together. In this case, it is the ability to interoperate or integrate different dataset. Taking this into account, LOD responds that open data is in RDF, allowing users to link data from various sources, institutions or organizations, explore and combine this data freely and without copyright restrictions for new web developments [26].

Additionally, knowledge graphs offer the possibility of classifying published data according to their possible uses and applications, in knowledge domains. As main reference to specify these domains, ones expressed in the Open Data Handbook [25] are took, which raises the following potential applications:

- Culture: data from cultural works and artifacts–for example, titles and authors–and in general, data obtained and maintained by galleries, libraries, archives and museums.
- Science: data produced as part of scientific research from astronomy to zoology.
- Finance: data such as government accounts (income and expenses) and information on financial markets (stocks, bonds, among others).
- Statistics: data produced by statistical offices such as census and socioeconomic indicators.
- Time: types of information used to understand and predict weather and climate.
- Environment: information related to the environment such as the presence and level of pollutants, air quality, among others.
- Transportation: data such as schedules, routes, weather statistics, among others.

2.3.Dataset Repositories

Supported in the Web of Data Model, data sources such as Europeana [27, 28] or Wikipedia offer information abstracted from linked dataset based on a common data model [29], as shown in Figure 4.

Figure 4. The Statue of Liberty [30].

On the Web, there are LOD initiatives such as data management platforms, including DataHub (https://datahub.io/) [31], Junar (http://junar.com/) [32],

Socrata (https://www.socrata.com/) [32]. These, among other platforms, have allowed a greater visibility of shared data, and have facilitated the participation of initiatives such as LOD Cloud [2] to visualize organizations and content providers that have released and linked their data, [33].

2.4.Metadata

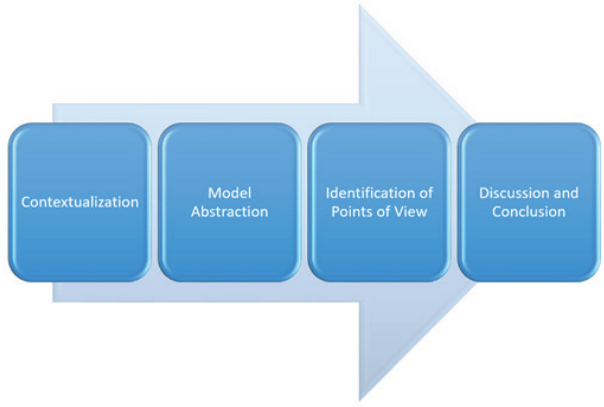
Metadata is an abbreviated representation of data to which it refers [34]. Analogously, it can be thought that Dataset published in a repository, in addition to resources it contains, have a metadata that presents an abbreviated representation of said Dataset. As described in [29], metadata is ubiquitous in information systems. When you listen to music through Spotify, post photos on Instagram, find a video on YouTube, connect with others through email, send text messages and connect to social networks, store contact lists on your devices mobile phones; all this content has attached metadata: information about creation, name, theme, characteristics and other aspects inherent in published data. Metadata is fundamental for the functionality of systems, because they allow users to find elements of interest, record essential information about them and share that information with others. Within metadata are different types [29], such as those described in Table 1:

This metadata can be stored as fields in relational database tables, or as XML documents.

3. Methodology

In order to carry out the proposed work, a descriptive research is proposed, based on bibliographic sources, which will allow exploring valuable elements related to the key components of Linked Data. For this purpose, following methodological design is proposed (Figure 5):

Figure 5. Methodological Design.



Source: own.

Methodological design consists of following stages:

- Contextualization: survey and tabulation of bibliographical references.
- Model Abstraction: Identification of vocabulary, categories and relationships, which will allow building the proposed model for the paper.
- Identification of Points of View: Analysis and identification of points of view, in the context of Linked Data.
- Discussion and Conclusion: Construction of discussion elements proposed, according to the analysis of the identified points of view.

4. Methodological development

In order to model and graphically visualize the interaction of concepts present in Linked Data, a generalization of the vocabulary, categories and relationships identified in the study context is presented below. This model seeks to provide tools to identify scenarios of points of view identified in the context of Linked Data.

Table 1. Metadata Types [29].

Type	Use	Properties of example
Descriptive metadata	Discovery Display Interoperability	Title, Author, subject, Genre, Publication Date.
Metadata Administrative a) Technical Metadata b) Preservation Metadata c) Rights Metadata	Information necessary to manage a resource: a) information about digital files necessary to decode and render them, such as file type b) support long-term management and future migration or emulation of digital files, for example, a checksum or hash; c) as a Creative Commons license [13, 15], which details property rights	Technical: File Type, File Size, Creation date/time Preservation: checksum, preservation event Rights: Copyright status, License terms, Right holder
Structural Metadata	Describe relationships of parts of resources to each other: pages in a sequence, a table of contents with pointers to the beginnings of the milestone sections, etc.	Sequence, Place in hierarchy

a. Contextualization: This research raises the following questions that allow guiding the process carried out:

i) How is the behavior of the points of view identified in the generic context of Linked Data?

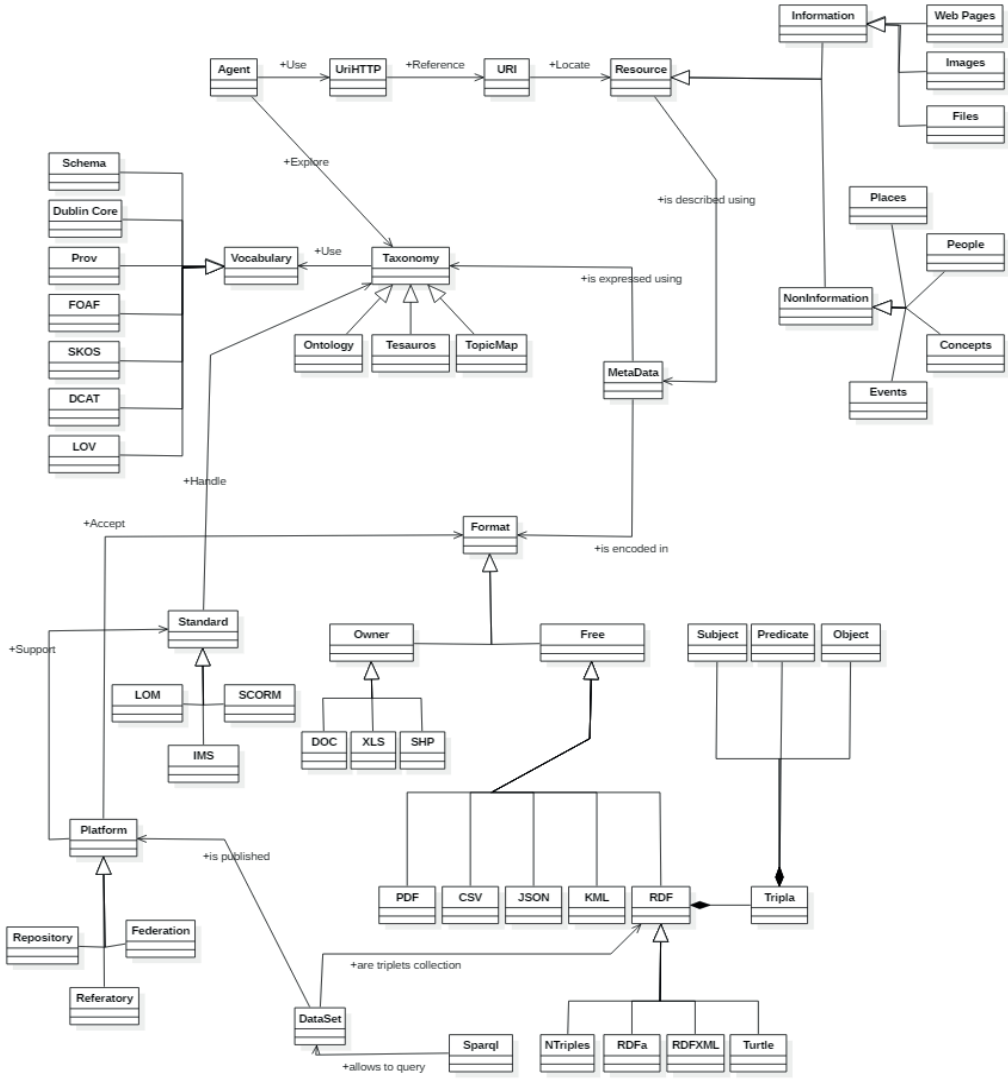
ii) What are the challenges that are determined under the points of view identified in the context of Linked Data?

To formalize the knowledge abstraction process about Linked Data, and taking into account the suggested research questions, concepts, categories and relationships are identified:

- Vocabulary: key concepts in Linked Data, such as: URI HTTP, URI, Schema, DublinCore, FOAF, among others.
- Categories, which group Vocabulary properties, which include Resource, Vocabulary, Taxonomy, Tool, among others.
- Relationships: links identified vocabulary, which include Use, Reference, Locate, among others.

These components describe the generic abstraction of the concrete elements of the Linked Data domain. Figure 6 shows the graph of the constructed metamodel, which describes the

Figure 6. Linked Data Metamodel proposed.



Source: own.

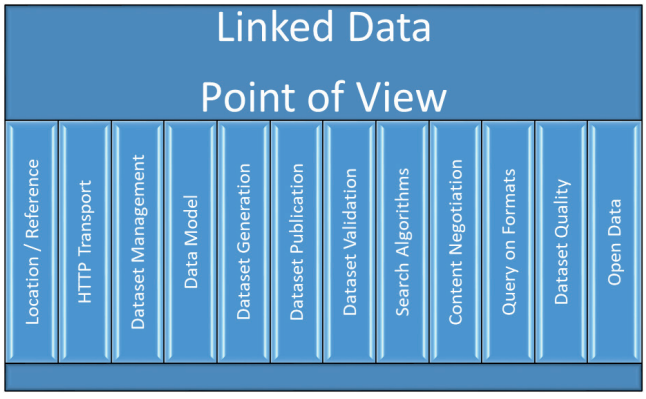
functional relationships of the knowledge domain, expressed in a data model. This meta-model categorizes and relates the Vocabulary set, through the identified relationships. Vocabulary describes primary concepts in the knowledge domain, which correspond to concrete elements of reality. Those concepts are represented in the metamodel as nodes of the graph. On the other hand, each arc corresponds to a relation, simple or compound, of knowledge domain, which links two elements of the vocabulary.

b. Points of View in Linked Data: Points of view configure well-formed graphs. These graphs constitute an important semantic factor of language, [35]. A point of view usually includes a graphic representation (diagram) of the structure being analyzed. Meta model diagram (Figure 6) represents the framework structure of the Linked Data knowledge domain. To document each point of view:

- An extension of a section of the metamodel is taken,
- Diagram is accompanied by additional information explaining identified situations in that segment,
- This additional information provides more detail about elements shown in the diagram and relationships between them.

Identified points of view are shown in Figure 7:

Figure 7. Points of view identified.



Source: own.

Identified points of view can be analyzed from two scenarios:

- Technologies, which support its execution,
- Processes, which require an even greater background, which goes beyond mere use of a tool.

Each of these points of view are described below.

i. Technological points of view: Points of view involve technologies that have proven their functionality, and have allowed developments in different areas. Within this scenario, following points of view are identified:

a. Point of View No. 1 Reference–Location: A resource is a localizable unit of information or service, such as files, images, documents, programs and search results. A URI reference [36] is a generic means to identify entities or concepts in the world (Figure 8). URIs are short strings that identify resources on the Web such as documents, images, downloadable files, services, e-mail boxes, and other resources [37]. URIs make resources available under different access methods such as HTTP or FTP [38]. The use of HTTP allows resources to be consulted over the Internet, and decisions about "namespaces" for resources are managed through the Domain Name System (DNS).

Figure 8. Point of View No. 1 Reference – Location.

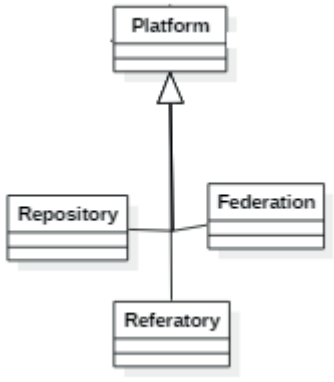
Source: own.

In this scenario, the referencing used may face problems such as broken links, incomplete links label, and empty link labels, among others. Problems that will not allow data to be referenced and located properly, or failing that, be located on the Web.

b. Point of View No. 2 HTTP Transport: The HTTP protocol is a universal mechanism for recovering resources, [39]. In Linked Data, data is found in HTTP servers (Figure 7), which establish a communication with each other in order to exchange data through the Internet, [40]. In this point of view, it is found that the HTTP protocol operates in LOD in an appropriate way, processes that can be checked when launching a URI that identifies a thing [1], for example: http://5stardata.info/en/examples/gtd-4/. On the other hand, in the current knowledge graphs, existing Dataset have managed to link properly.

c. Point of View No. 3 Dataset Management: Dataset are combined in platforms (Figure 9), to create services that allow published data consumption.

Figure 9. Point of View No. 3 Dataset Management.



Source: own.

Data management platforms define a set of rules for HTTP operations on web resources, based on RDF. They also provide an architecture and some type of exploitation interface, either for interaction with humans or with other software systems, [41, 42]. Within identified platforms are a) Repositories, which store resources and metadata, b) Referatory, in which metadata is managed and a link to resource, and c) Federations, which correspond to groupings of repositories with a common access point, such as: Ariadne, Globe, among others, [43]. On the other hand, these platforms use standards to normalize published resources. In addition, these platforms provide a common language with which objects can communicate with different learning management platforms (LMS), databases and web applications. Some of supported standards are [44]:

- Metadata standards such as Dublin Core, IEEE LOM.
- Standards for creation of platforms, such as ADL SCORM, IMS.

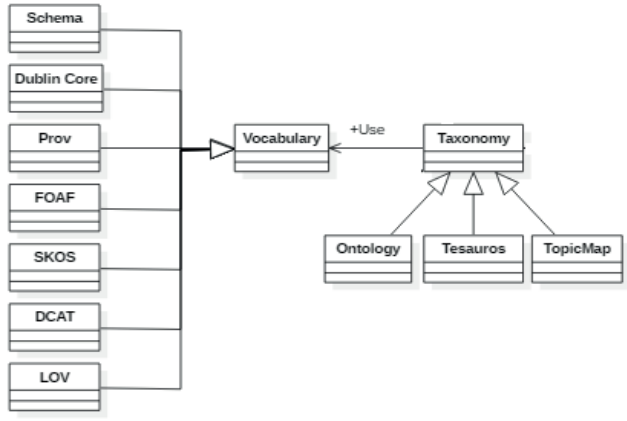
In addition, these platforms make use of protocols such as Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [45], Simple Publishing Interface (SPI) [46], or Simple Query Interface (SQI), [47].

In general, data management platforms allow storing, searching, retrieving, querying and accessing resources from all areas of knowledge, through either Sparql query services or harvesting processes, [48].

d. Point of View No.4 Modeling the Dataset: After analyzing data domain, data model, that will support the dataset instances, is constructed (Figure 10). This work is based on reuse of available vocabularies, in order to make search processes more efficient, and accelerate the model development. This activity consists of the following tasks [49]:

- Search for suitable vocabularies. There are some useful repositories to find available vocabularies, such as Schema, SchemaCache, Swoogle and LOV. For the election of the most suitable vocabularies, it is recommended to follow the guidelines proposed in the Open Data Commons Open Database License (ODbL).
- In case no suitable vocabulary can be found, it should be created trying to reuse a large part of possible existing resources.
- Finally, if an available vocabulary or resources for ontology construction are not found, the ontology must be created from scratch.

Figure 10. Point of View No 4 Data Modeling of Datasets.



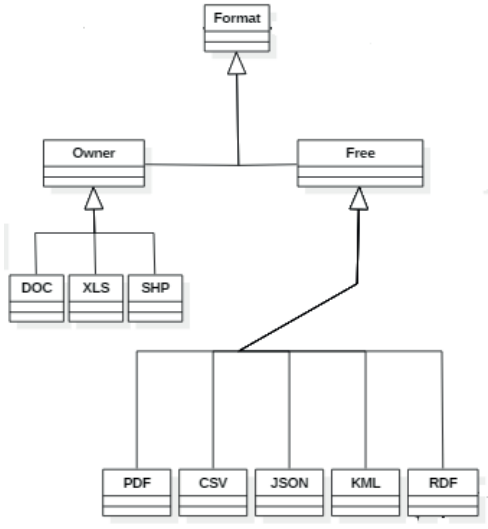
Source: own.

In this scenario, problems such as existence and relevance of the vocabulary that required modeling data of the specific knowledge area are identified, [50].

e. Point of View No. 5 Dataset Generation: In this scenario (Figure 11), instances generation of data models is carried out, with technologies

such as RDF. RDF offers a standard model for exchanging of data on the Web, with features that facilitate data fusion even if underlying schemes are different. In addition, RDF supports the evolution of schemes over time without the need for all data consumers to be changed, [10].

Figure 11. Point of View No 5 Dataset Generation.



Source: own.

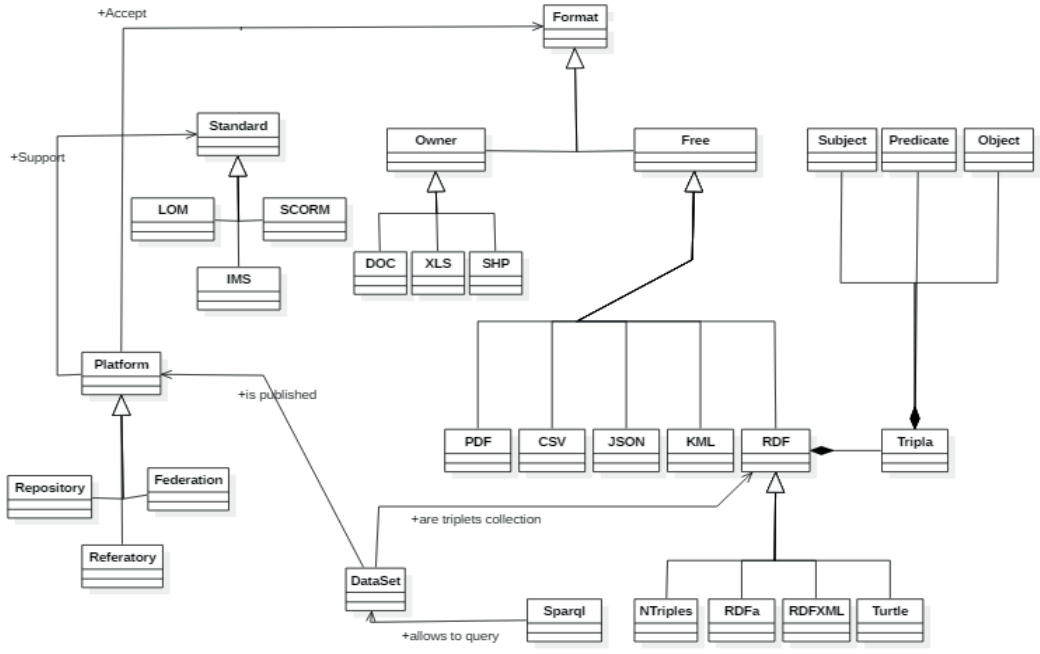
After data models design, Dataset instances generation process carries out following [49]:

- Transformation: Your requirements are (1) the complete conversion, which implies that all possible queries in the original source are also possible in the RDF version; and (2) generated RDF instances should reflect the structure of the selected ontology.
- Data cleaning: activity focused on cleaning of the identified noise in the Dataset, such as broken links.
- Link: create links between the Dataset and the external Dataset. This task involves discovering relationships between data elements. These links can be created manually or with automatic tools.

f. Point of View No. 6 Dataset Publication: As shown in Figure 12, after Dataset instances are generated in RDF, these instances are published in a data management platform.

Tools such as [51]: D2R server, Triplyfy, Talis Platform, Pubby, Paget, Linked Media Framework, Virtuous Universal Server Open Link, Virtuoso Sponger, Sesame, Jena TDB,

Figure 12. Point of View No. 6 Dataset Publication.



Source: own.

3Store, are used in Dataset publishing process. Publication task consists of following sub tasks:

- Data publication: RDF instances are stored and published in a repository. This is done using tools such as Jena, Sesame. Some of them already include the SPARQL endpoint and Linked Data frontend.
- Metadata Publication: Metadata information must be included. For this task, vocabularies such as VoID are used. These vocabularies allow expressing Dataset RDF metadata, and it includes general metadata, access metadata, structural metadata and links between Dataset.
- Enable effective detection: Site map is defined to allow efficient discovery and synchronization of Dataset. This includes the Dataset in the LOD cloud diagram and in the available open catalogs.

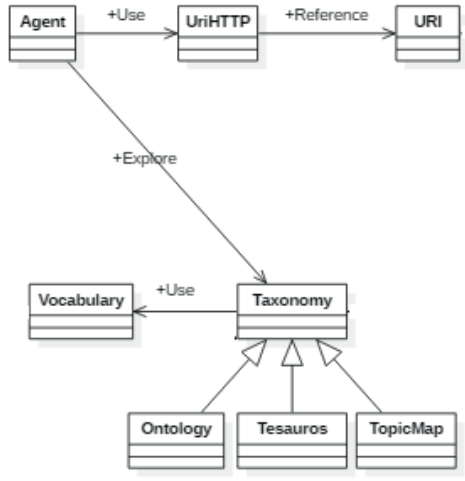
g. Point of View No. 7 Dataset Validation: Publication process in Linked Data is parameterized under a 5-level score scheme [8]. DataSet must comply with a series of validations. These validations place it in a specific category, until it is linked to the LOD Cloud [2]. To carry out Dataset validation, there are Linked Data editors and validators, such as [51]:

- Hyena: RDF Editor, a mixture of a wiki and a database, available as a desktop application and as a web application.
- Vapor: Linked Data Validator.
- DrifftR: Linked Data editor and browser
- Validation Service, [52].

Tools that have functions such as determining if a URI identifies a real-world object; check the publication method and check the validity of the redirection chains between the URIs of the real-world object and the URI document, in order to avoid redirection problems.

h. Point of View No. 8 Search Algorithms used by Agents: Semantic Web technology combined with Machine Learning techniques allows the creation of intelligent agents (Figure 13). These agents can expand their knowledge base to all types of semantic content linked to the Internet, as well as learn through interaction with their trainer and other people who interact with them, [53].

Figure 13. Point of View No. 7. Intelligent agents.

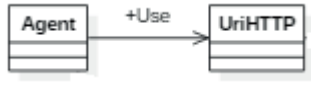


Source: own.

LOD has an extensive amount of information that intelligent agents, supported by semantic technologies, can reuse and take advantage of to create solutions with high benefit for user, [54].

i. Point of View No. 9 Content Negotiation: Derreferenble URIs allows the use of the HTTP protocol in content negotiation (Figure 14), that is, data transmission between the client (which indicates the format) and the server (which supplies data in the indicated format), [55].

Figure 14. 2 Point of View No. 8. Content Negotiation.



Source: own.

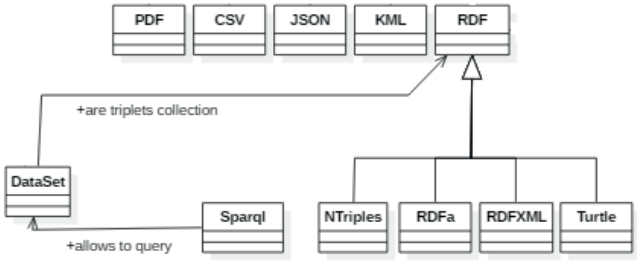
In content negotiation, if a web client requests an object in a given format through the neutral URI, the server returns HTTP status code "303 See others", along with the URI where data is located in the format requested. If the client connects with a specific URI and format, it retrieves data in that format, [55].

In general, RDF is based on triplets formed by three elements: subject, predicate and object. In the example, the movie "Way of the Dragon" is the subject described in "about" attribute of the rdf: Description. The predicate dbpprop: starring

indicates the meaning of relationship: "acts on that movie". The "Chuck Norris" object appears in rdf: resource attribute. It is important to keep in mind that URIs identify concepts. For example, the URI http://dbpedia.org/resource/Chuck_Norris, identifies Chuck Norris, not the document that contains information about him. If the URI refers to Chuck Norris, when writing it in the browser, the same Chuck Norris should appear. As this is not possible, a content negotiation is carried out, which returns the page http://dbpedia.org/page/Chuck_Norris, which contains an HTML document with information about it, [56].

j. Point of View No. 10 Queries on Formats: Tools such as the SPARQL endpoints (Figure 15) are used for data exploitation. Endpoints are set up as web services, which allow querying semantic repositories using the SPARQL query language [57]. SPARQL is a semantic query language based on graphs search. The query describes the RDF graph pattern to be searched in the RDF graph of the repository, [58].

Figure 15. Point of View No. 9. Queries on Formats.

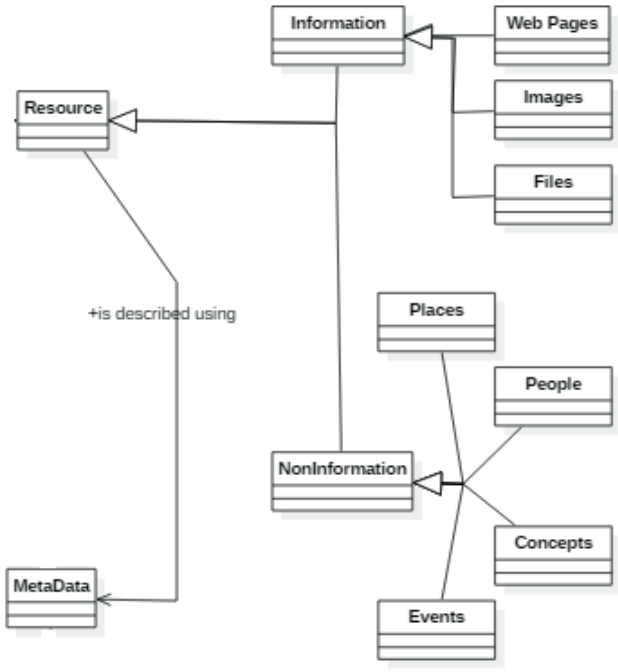


Source: own.

k. Point of View No. 11 Data Quality: According to Pons et al. [59], metadata quality is a prerequisite for metadata to be useful. The metadata must be properly defined so that it can be filled correctly and can be effectively exploited. The metadata quality reflects the degree to which they perform their essential functions of search, location, use, origin, authentication and administration.

In response to this, different quality assurance models have been formulated, based on Linked Data quality metrics, which address criteria to evaluate both data and metadata of the linked resources (Figure 16).

Figure 16. Metadata Quality.



Source: own.

As described in BizkaiLab [60], one of the fundamental principles of Linked Data is that data sources contain links to information in other data sources. RDF-based data instances have mechanisms to link information from different data sources into a single global graph, based on world abstraction. However, this is not always possible since data sources providers do not carry out adequate data model, do not publish their models openly, or need much more specific vocabularies than those available. Therefore, many of the Linked Data quality models evaluate and try to correct problems in the published data instances, which allows identifying the need to improve the quality in the abstraction and design of open data models.

l. Point of View No. 12 Open Data: Governments and institutions put data they administer, freely available to individuals and institutions–without copyright, patent or other restrictions–in formats that allow reuse for any purpose, [61]. For its part, LOD indicates that they are open data in RDF, so that the user can link data from various sources, institutions or organizations explore and combine this data freely and without copyright restrictions for new web developments, [26].

LOD is often used to refer to Linked Data in general, rather than explicitly Linked Data that is published under an open license. Not all Linked Data will be open, and not all open data will be linked. Therefore, care must be taken when using the appropriate term, depending on the terms of the license of used data, [62].

ii. Scenarios referring to Processes

The points of view involve processes of abstraction. Within this scenario, the following points of view are identified:

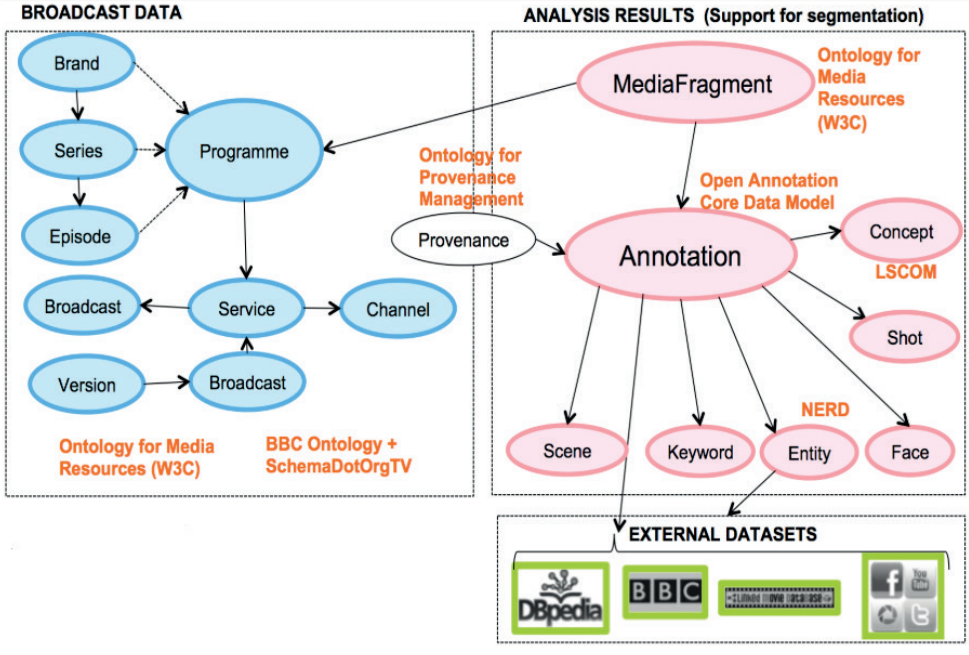
a. Point of View No. 4 Dataset Model: Dataset correspond to RDF instances of data models. These data models are designed for the integrated representation of information that originates from multiple sources, and correspond to real-world abstractions, which are subsequently represented by different schemes (Figure 17).

In this scenario, there are some problems such as:

- The adequate abstraction of the elements of the world, which allow a data model according to the reality that is sought to represent.
- The opening of data model, in order to be known and reused by other developers.

b. Point of View No. 5 Dataset Generation: RDF was designed as a data model for metadata. It has come to be used as a general method for modeling information that is implemented in web resources. This model converts resource declarations into expressions with the subject-predicate-object form (known in RDF terms as triplets). The subject is the resource, that is, what is being described. The predicate is the property or relationship that you want to establish about the resource. Finally, the object is the value of the property or the other resource with which the relationship is established [10]. These triplets can contain poorly constructed, incomplete or empty objects, such as broken or incomplete links; which would mean that the location and transport is not carried out in the correct way.

Figure 17. Linked Data Model [63].



c. Point of View No. 7 Dataset Validation: In some of these validation services (Figure 18), only the triplet construction is validated, but not the content offered by this construction.

This validation of content is delivered to quality assurance models, which are responsible for applying metrics to data instances, in order to generate trust to users.

d. Point of View No. 11 Data Quality: In most cases, authors are responsible for the information contained in the Dataset. In this process, authors may not have knowledge of the model that supports the information that is being completed, or does not know adequately, domain restrictions that these elements preserve, reason why data begin to present quality problems from the moment of its instantiation.

5. Discussion and conclusion

As seen in different points of view, resulting from metamodel abstraction, Linking Data process is based on technological tools such as URI, HTTP and RDF. These tools allow a very solid operational performance such as data transport, derreferencing, triplet construction, consultation processes, etc.

For example, original purpose of RDF is data exchange. RDF was not created with the objective of replacing any technology, on the contrary, it allows complementing and providing added value to existing technologies, improving the representation level of published data on the Web and facilitating communication between systems and agents; all this, regardless of how they are constructed and underlying technology, [64].

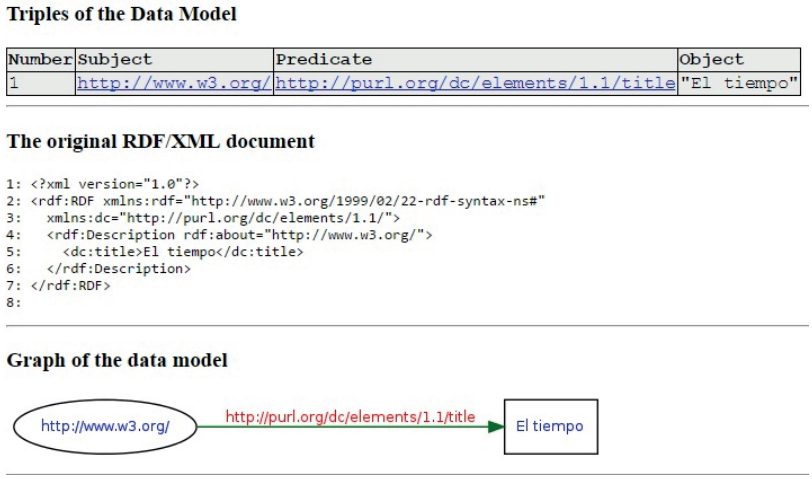
This technological support allows to carry out data publication and exploitation process in an appropriate way, from factors such as:

- Validate the structure of the RDF triplets.
- Perform the loading and storage of data on platforms.
- Offer consulting services through endpoints.
- Referencing resources that you want to link.

Among other factors that the Web of Data has allowed through the evolution of its tools. However, it is important to consider that linking resources process is not only based on technological tools, it also has an even greater support in the abstraction and modeling of the objects of the world, and in the quality of published data.

Regarding data quality, many quality assurance models of Linked Data have been generated, most of them oriented to data instances already published. In the case of data abstraction, this process allows configuring data models, which will later be

Figure 18. Validation Services.

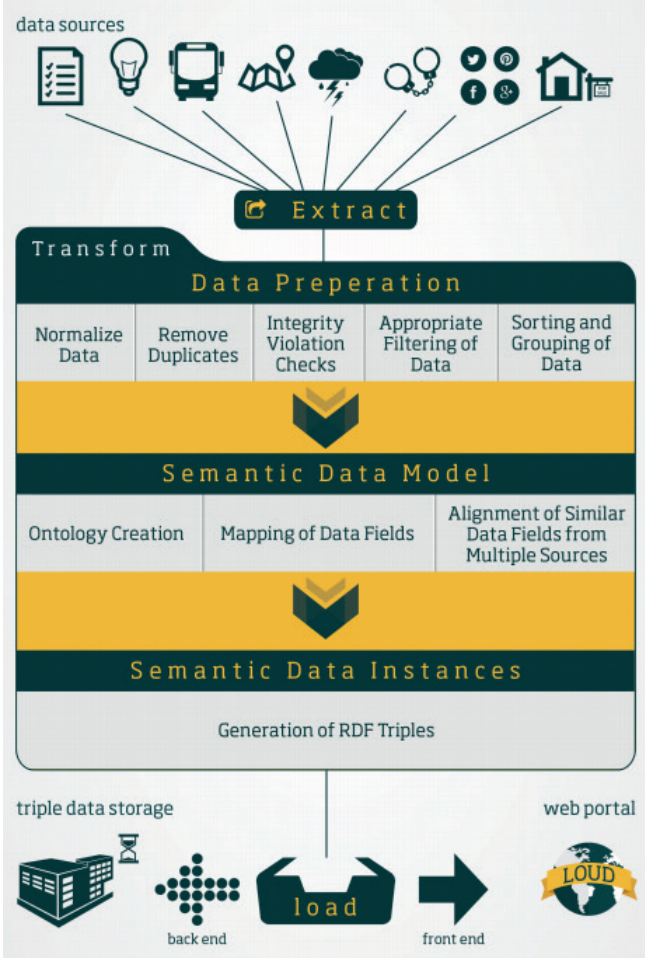


Source: own.

instantiated as Dataset, which will be exploited as linked resources. Data modeling, as a primary factor in linking processes, is faced with circumstances such as:

- Build a mega ontology that covers all existing objects in the world has been impossible.
- Data models for particular domains have been developed by authors, under their own abstraction, and in some cases do not take into account considerations such as calling things or entities in the same way that others do.
- In most cases, data models are not published openly, so that they cannot be shared or reused by other authors.
- As data models are not published openly, it increases the heterogeneity among data sources, which generates the creation of additional layers that allow dialogue between data instances, in addition to reducing the data quality and therefore, of results in generated queries.
- Quality assurance processes of the Linked Data are carried out on data instances [27], applying different quality models that circumscribe particular metrics to evaluate the quality in these instances.
- In order to address these problems, proposals have been developed such as Extract-Transform-Load process (Figure 19), which carries out processes of [65]:
- Data Extraction from appropriate data sources. Data is generally available in plain file formats, such as csv, xls and txt, or is available through a RESTful client.
- Transform: this phase involves data cleaning to comply with the objective scheme. Some of the typical transformation activities include normalizing data, eliminating duplicates, verifying violations of integrity constraints, filtering data based on some regular expressions, classifying and grouping data, applying integrated functions when necessary, etc.
- Load: this phase involves data propagation in a data warehouse or a data warehouse that serves Big Data.

Figure 19. ETL Process [65].



Nevertheless, as evidenced in its description, data come from plain files that require a transformation process. Process for which it is very important to have a data model that represents an adequate abstraction of the world, in a specific data domain, that uses reusable vocabularies, and above all, that is open to all people.

Acknowledgement

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

References

[1] DERIGalway, "Linked Data (and the Web of Data)", 2010. [Online]. Available at: https://www.youtube.com/embed/GKfJ5onP5SQ

[2] W3C, "RSS 2.0 Specification", 2002. [Online]. Available at: https://validator.w3.org/feed/docs/rss2.html

[3] W3C, "CSV on the Web current status", 2015. [Online]. Available at: https://www.w3.org/standards/techs/csv#w3c_all

[4] D. Wood, M. Zaidman, L. Ruth and M. Hausenblas, "Linked Data: Structured Data on the Web", Manning Publications, 2014.

[5] Creative Commons, "Creative Commons Global Summit", [Online]. Available at: http://creativecommons.org/

[6] DBpedia, "Explore DBpedia". [Online]. Available at: http://www.dbpedia.org/

[7] A. Figueredo and P. Wolf, "Assortative pairing and life history strategy–a cross-cultural study", Human Nature, vol. 20, 2009, pp. 317-330, https://doi.org/10.1007/s12110-009-9068-2

[8] W3C, "HTTP Current Status", [Online]. Available at: https://www.w3.org/standards/techs/http#w3c_all

[9] Geonames, "Geographical database". [Online]. Available at: http://www.geonames.org/

[10] N. Korn and C. Oppenheim, "Licensing Open Data: A Practical Guide", 2011. [Online]. Available at: http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf

[11] S. Speicher, J. Arwe and A. Malhotra, "Linked Data Platform 1.0", 2015. [Online]. Available at: https://www.w3.org/TR/ldp/

[12] J. Riley, "Understanding metadata: what is metadata, and what is it for?: A Primer", 2017. [Online]. Available at: https://www.niso.org/publications/understanding-metadata-2017

[13] M. Holley, "What is Metadata and what is it as important as the data itself?", 2016. [Online]. Available at: https://www.opendatasoft.com/2016/08/25/what-is-metadata-and-why-is-it-important-data/

[14] W3, "Validation Service", 2006. [Online]. Available at: https://www.w3.org/RDF/Validator/

[15] Asociación Nacional de Centros de E-Learning y Distancia, "Multimedia en elearning", 2011. [Online]. Available at: https://sites.google.com/site/multimediaenelearning/estandares-de-multimedia

[16] DataHub, "DataHub Project". [Online]. Available at: http://datahub.io/dataset?tags=lod

[17] Portal Biblioteca del Congreso Nacional de Chile, "Linked Open Data: ¿Qué es?", 2014. [Online]. Available at: http://datos.bcn.cl/es/informacion/que-es

[18] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, "Dbpedia: A nucleus for a web of open data", 2007. [Online]. Available at: https://www.cis.upenn.edu/~zives/research/dbpedia.pdf

[19] 5 star data, "Datos Abiertos". [Online]. Available at: http://5stardata.info/es/

[20] F. Bauer and M. Kaltenbock, "Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers", Vienna, Austria: Edition mono/monochrom, 2012.

[21] R. M. Pérez, J. M. Santos, V. M. Alonso, L. M. Álvarez and F. A. Mikic, "Linked Data como herramienta en el ámbito de la nutrición", Nutición Hospitalaria, vol. 27, no. 2, 2012.

[22] LOD cloud diagram, "The Linking Open Data Cloud Diagram. [Online]. Available at: http://lod-cloud.net/

[23] W3C, "URI", 2005. [Online]. Available at: https://www.w3.org/wiki/URI

[24] T. Berners-Lee, "What is linked data? TED2009", 2009. [Online]. Available at: http://data.gov.uk/linked-data

[25] M. Lama, "Web Semántica y Linked Data, Tecnologías semánticas". [Online]. Available at: https://citius.usc.es/sites/default/files/formacion/BD&DS_ManuelLama.pdf

[26] Open Archives initiative, "Protocol for Metadata Harvesting". [Online]. Available at: https://www.openarchives.org/pmh/

[27] C. Bizer and T. Heath, "Linked Data. Evolving the Web into a Global Data Space", Morgan & Claypool Publishers. The Semantic Web: Theory and Technology. 2011.

[28] Ministerio de Hacienda y Administración Pública. Ministerio de Industria, Energía y Turismo, "Plataformas de Publicación de Datos Abiertos", 2015. [Online]. Available at: http://datos.gob.es/sites/default/files/informe-herramientas-publicacion.pdf

[29] D. Wood, "Linking Government Data", New York: Springer, 2011. https://doi.org/10.1007/978-1-4614-1767-5

[30] Wikipedia, "Statue of Liberty", 2007. [Online]. Available at: https://en.wikipedia.org/wiki/Statue_of_Liberty

[31] M. Hallo, M. Martínez and P. de la Fuente, "Las tecnologías de Linked Data y sus aplicaciones en el gobierno electrónico", Scire, vol. 18, no. 1, 2012.

[32] R. García, F. Radulovic, O. Corcho, M. Poveda, V. Rodríguez, A. Gómez and D. Vila, "Linked Data Generation Process", 2015. [Online]. Available at: https://www.slideshare.net/ld4sc/linked-data-generation-process?from_action=save

[33] C. Medina, "Posgrado en Ciencias y Tecnologías de la Información". [Online]. Available at: http://mcyti.izt.uam.mx

[34] W3C, "Naming and Addressing: URIs, URLs", 1993. [Online]. Available at: https://www.w3.org/Addressing/

[35] Colombia Aprende, "Objetos Virtuales de Aprendizaje e Informativos". [Online]. Available at: http://colombiaaprende.edu.co/html/directivos/1598/article-172369.html

[36] T3chFest, "Creación de Agentes Inteligentes aplicando Tecnologías de la Web Semántica y Aprendizaje Automático", 2015. [Online]. Available at: https://t3chfest.uc3m.es

[37] R, Saquete, "El Impredecible Futuro de la Web Semántica", 2013. [Online]. Available at: http://www.humanlevel.com/articulos/desarrollo-web/el-futuro-de-la-web-semantica.html

[38] J. Pastor, F. Martínez, R. López and J. Rodríguez, "Publicación como Linked Open Data de la Nomenclatura Internacional de Ciencia y Tecnología y del Tesauro UNESCO". [Online]. Available at: http://eprints.rclis.org/24272/1/ICongressoISKOEspanhaePortugal000211987.pdf

[39] A. Graves, "Tutorial: Creando Aplicaciones Basadas en Linked Data", 2012. [Online]. Available at: http://manzanamecanica.org/2012/01/tutorial_creando_aplicaciones_basadas_en_linked_data_parte_13.html

[40] D. Pons, J. Hilera and C. Pagés, "La estandarización para la Calidad en los Metadatos de Recursos Educativos Virtuales". [Online]. Available at: http://www.esvial.org/wp-content/files/estandarizacionmetadatosPonsHileraPages.pdf

[41] W3C, "LinkingOpenData", 2017. [Online]. Available at: https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[42] J. F. Herrera-Cubides, P. A. Gaona-García, K. Orjuela, "A View of the Web of Data. Case Study: Use of Services CKAN", Ingeniería, vol. 22, no. 1, 2017, p.p. 46-64. https://doi.org/10.14483/udistrital.jour.reving.2017.1.a07

[43] BizkaiLab, "Estado del arte en confianza y calidad de fuentes de datos enlazadas", 2011. [Online]. Available at: http://www.bizkailab.deusto.es/wp-content/uploads/2012/04/5761.pdf

[44] W3C, "HTTP–Hypertext Transfer Protocol", 2014. [Online]. Available at: https://www.w3.org/Protocols/

[45] J. Bolaños, H. Medina and R. Ferro, "Metaproceso de Desarrollo de Software", Bogotá: Editorial UD Universidad Distrital Francisco José de Caldas, 2016.

[46] J. L. Redondo and R. Troncy, "LinkedTV Ontology", 2013. [Online]. Available at: http://semantics.eurecom.fr/linkedtv/

[47] B. Hyland, G. Atemezing and B. Villazón-Terrazas, "Best Practices for Publishing Linked Data", 2014. [Online]. Available at: https://www.w3.org/TR/ld-bp/

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

[49] Datos España, "Linked Data Como Modelo de Datos", 2017. [Online]. Available at: http://datos.gob.es/es/noticia/linked-data-como-modelo-de-datos

[50] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini and H. Van de Sompel, "El Modelo de Datos de Europeana (EDM)", 2010. [Online]. Available at: http://conference.ifla.org/past-wlic/2010/149-doerr-es.pdf

[51] S. K. Bansal and S. Kagemann, "Integrating Big Data: A Semantic Extract-Transform-Load Framework", Computer, vol. 48, no. 3, 2015, pp. 42-50. https://doi.org/10.1109/MC.2015.76

[52] S. Ternier, D. Massart, M. Totschnig, J. Katholieke and E. Duval, "The Simple Publishing Interface (SPI)", D-Lib Magazine, vol. 16, no. 9/10, 2010. https://doi.org/10.1045/september2010-ternier

[53] J. Canabal and A. Sarasa, "Agrega–plataforma de Objetos Digitales Educativos". [Online]. Available at: http://ceur-ws.org/Vol-318/Canabal.pdf

[54] Linked Data, "Connect Distributed Data across the Web". [Online]. Available at: http://linkeddata.org/home

[55] P. A. Gaona-García, A. Fermoso-García and S. Sánchez-Alonso, "Exploring the Relevance of Europeana Digital Resources: Preliminary Ideas on Europeana Metadata Quality", *Revista Interamericana de Bibliotecología*, vol. 40, no. 1, 2017, pp. 59-69. https://doi.org/10.17533/udea.rib.v40n1a06

[56] IDECA, "Propuesta de Patrón de URI como parte de la Iniciativa de Linked Data, Infraestructura de Datos Espaciales para el Distrital Capital", 2014. [Online]. Available at: http://www.ideca.gov.co

[57] W3C, "Resource Description Framework (RDF)", 2014. [Online]. Available at: https://www.w3.org/RDF/

[58] G. Klyne and J. J. Carroll, "Resource description framework (RDF): Concepts and abstract syntax", 2004. [Online]. Available at: https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[59] OData, "JSON Format (OData Version 2.0)". [Online]. Available at: http://www.odata.org/documentation/odata-version-2-0/json-format/

[60] P. A. Gaona-García, S. Sánchez-Alonso and A. Fermoso, "Visual analytics of Europeana digital library for reuse in learning environments: A premier systematic study", Online Information Review, vol. 41, no. 6, 2017, pp. 840-859. https://doi.org/10.1108/OIR-04-2016-0114

[61] F. Herrera-Cubides, P. A. Gaona-García and S. Sánchez-Alonso, "The Web of Ddata: Past, Present and ¿Future?", XI Latin American Conference on Learning Objects and Technology (LACLO), 2016. https://doi.org/10.1109/LACLO.2016.7751802

[62] W3C, "Guía Breve de Tecnologías XML". [Online]. Available at: http://www.w3c.es/Divulgacion/GuiasBreves/TecnologiasXML

[63] 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

[64] M. Lamarca, "XLL", 2013. [Online]. Available at: http://www.hipertexto.info/documentos/xll.htm

[65] Open Knowledge International, "What is Open?". [Online]. Available at: https://okfn.org/opendata/