An Approach for Generating and Semantically Enriching Dataset Profiles

Natacha Targino[1], Damires Souza[2] and Ana Carolina Salgado[1]
[1]Centro de Informática (CIn) - Federal University of Pernambuco (UFPE), Recife, Pernambuco, Brazil
[2]Federal Institute of Paraíba (IFPB), João Pessoa, Paraíba, Brazil

Keywords: Datasets on the Web, Dataset Profile, Metadata, Semantic Enrichment, Data Publication and Consumption.

Abstract: The identification of appropriate datasets on the Web for a given task is still a challenge. To help matters, metadata describing datasets can be provided. It is possible to make these metadata available through a Dataset Profile (DSP). In this light, this work presents an approach which generates a DSP composed by descriptive, structural and quality metadata. The DSP is enriched by semantically referencing the provided metadata and by means of some new metadata, such as the dataset domain and some quality metadata, e.g. comprehensibility and processability. In order to evaluate the proposed approach, a prototype has been developed and some experiments have been accomplished.

1 INTRODUCTION

The large amount of datasets available on the Web enables their usage in diverse scenarios, and combinations of these datasets can bring important findings. However, publishers and consumers of datasets on the web usually do not know each other. This implies in a fundamental need for a common understanding between dataset publishers and dataset consumers. Without this comprehension, dataset publishers' efforts may be incompatible with dataset consumers' needs (Lóscio et al., 2017).

In order to facilitate that communication, metadata regarding the datasets are usually made available (Clarke et al., 2014). Dataset metadata can improve the understanding and processing of the data, both by humans and by machines. Metadata are usually composed by descriptive information about the content, structure, quality, and other characteristics of the datasets. Nevertheless, it is not yet a common practice for publishers to provide metadata that completely represent the content of published datasets (Abele, 2016). Recommended as a best practice by the World Wide Web Consortium (W3C) (Lóscio et al., 2017), providing metadata is a fundamental requirement when publishing datasets on the Web. It may help humans and machines not only to understand the data but also important aspects that describe their distributions (i.e., physical forms of a dataset).

To help matters, data enrichment processes can be performed on the metadata in order to generate new metadata and help in assigning meaning to them. In this latter case, the metadata can be semantically referenced by terms of recommended vocabularies available on the Web (Heath et al., 2011). For example, considering a dataset metadata "date of publication", it can be replaced by dct:issued, a term which belongs to the Dublin Core Metadata Initiative (DCTerms) vocabulary. In addition, representing the metadata in semantic data formats, such as the Resource Description Framework (RDF) model, can facilitate their processing and understanding by data consumers.

In order to enable the structuring of dataset metadata, some authors propose the creation of dataset profiles (Abele, 2016; Assaf et al., 2015). However, most of existing dataset profile examples only provide descriptive metadata. Thus, the need of an approach that may provide more detailed information about datasets on the web by means of enriched profiles is evidenced.

This work presents an approach, called DSPro+ (DataSet Profile with Enrichment), which describes datasets published on the Web by means of the creation of enriched Dataset Profiles (DSP). It extends the DSP generation approach introduced in Targino et al., (2017). The main objective of this proposal is to facilitate the understanding between dataset publishers and dataset consumers. The DSP

Targino, N., Souza, D. and Salgado, A.
An Approach for Generating and Semantically Enriching Dataset Profiles.
DOI: 10.5220/0007187400900101
In *Proceedings of the 14th International Conference on Web Information Systems and Technologies (WEBIST 2018)*, pages 90-101
ISBN: 978-989-758-324-7

is composed by: (i) descriptive metadata, which includes keywords, title, knowledge domain and a recommendation of domain vocabularies; (ii) structural metadata, which describes the internal structure of a dataset; and (iii) quality metadata, which regards quality criteria concerned with dataset comprehensibility and processability. The generated dataset profile is represented in machine readable format (RDF), facilitating its manipulation.

Our contributions are summarized as follows:

(i) We present an approach to generate enriched dataset profiles;
(ii) We extract most descriptive metadata which are already available in a dataset;
(iii) We include some new descriptive metadata and also present structural ones;
(iv) We propose the definition of quality criteria associated with a dataset which are also included in a DSP;
(v) We provide a prototype that implements the proposed approach; and
(vi) We describe some experiments that evaluate the proposed approach and, particularly, the semantic enrichment process.

This paper is organized as follows: Section 2 introduces some concepts; Section 3 proposes the approach; Section 4 presents the obtained results; Section 5 discusses related work, and Section 6 exposes some conclusions and future work.

2 FUNDAMENTAL CONCEPTS

Nowadays, the web may be considered as an appropriate ecosystem for production and consumption of datasets. In open data portals, such as the European Union Open Data Portal (https://open-data.europa.eu/) and the Ireland's Open Data Portal (https://data.gov.ie/), some metadata are added to the published datasets. These metadata usually include dataset creation date, defined usage license, and dataset formats or distributions. Some open data portals provide data APIs as well.

Regarding the common provided metadata, we argue that they can be enriched to generate better descriptions of the datasets. A data enrichment process, in general, refers to a set of tasks that can be used to enhance, refine or improve raw or previously processed data (Lóscio et al., 2017). In this sense, it is possible to include, for instance, the identification of the knowledge domain of a dataset (e.g., Health, Music, Education) or even some related quality criteria in terms of metadata.

The quality of a dataset can have a big impact on the quality of applications that use it (Lóscio et al., 2017). Thus, Information Quality (IQ) criteria may be defined on datasets in order to enrich their suitability for specific usages. The notion of IQ has emerged during the past years and shows a steadily increasing interest. IQ is based on a set of dimensions or criteria. The role of each one is to assess and measure a specific quality issue (Naumann et al., 2000; Wang et al., 1996; Pipino et al., 2002). Examples of IQ criteria concerned with datasets on the web are timeliness, consistency, verifiability and comprehensibility (Zaveri et al., 2013; Naumann et al., 2000; Flemming, 2011). Each quality criterion has a set of indicators, which allows the evaluation of the quality of a data source (Flemming, 2011). In this work, we consider two IQ criteria named as comprehensibility and processability of a dataset, which are defined in Section 3.3.

Given the need to provide better descriptions of datasets, some works have used dataset profiles for this purpose. Abele (2016) defines data profiling as the process of creating descriptive information and collecting statistics about the dataset. For Ellefi et al., (2014) a dataset profile is a set of characteristics, both semantic and statistical, that allow to describe in the best possible way a dataset.

3 THE DSPro+ APPROACH

The DSPro+ approach has been defined as a means to generate a dataset profile with some kinds of metadata enrichment. Considering the related works and according to the indications of good practices for data publication on the Web from W3C (Lóscio et al. 2017), this work defines some concepts. In this work, a dataset is defined as follows.

Definition 1. Dataset (d) - A dataset d represents a collection of data published on the Web that is available for access through a distribution.

A distribution of a dataset represents a specific way (e.g., CSV, API) in which a dataset d is made available to consumers. The same dataset d may be available in one or more distributions.

The main idea underlying this work is that datasets are published and searched for consumption. To facilitate this task, we associate a dataset with a profile, which is defined as follows.

Definition 2. Dataset Profile (DSP (d)) - A Dataset Profile DSP(d) consists of a body of semantically-enriched metadata regarding a dataset d and composed by descriptive, structural and quality information on d.

The DSP generation aims to provide a better understanding of a dataset and its content, making it more accessible and understandable by both people and machines. The proposed approach is composed of some steps, such as data extraction, metadata identification, generation of new metadata, and formation of DSP. Figure 1 illustrates the process that underlies the approach. As depicted in Figure 1, it receives as input a dataset published on the web, which can be accessed through its URL, and extracts its data. Then, it identifies some metadata, which are originally included in a dataset. These metadata may be available in different forms, such as a specific file, in HTML or through a JSON-LD script. Thus, usually, it is possible to directly obtain some descriptive metadata, such as a dataset title, its description and date of publication.

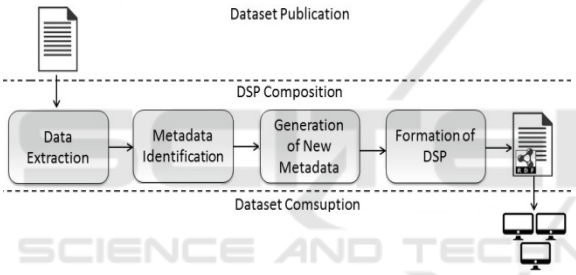

Figure 1: Process for the generation of a DSP(d).

Next, new descriptive metadata (domain and vocabulary recommendation) as well as structural and IQ metadata are generated (steps explained in the following sections). Then a DSP is built in accordance with the generated and identified metadata. The resulting metadata are semantically referenced by means of recommended vocabulary terms. To this end we use vocabularies, as the Data Catalog Vocabulary (DCAT) and Data Quality Vocabulary (DQV), which are recommendations provided by the W3C (Lóscio et al., 2017).The resulting DSP is serialized in an RDF distribution.

3.1 Descriptive Metadata

In this work the concept of descriptive metadata is stated as follows.

Definition 3. Descriptive Metadata (DM(d)) - Descriptive metadata DM(d) represent information about the overall features of a dataset d.

In Table 1, the stated DM(d) for a DSP(d) and their corresponding vocabulary terms are presented. Some of these metadata regard the scheme proposed by the W3C Best Practices (Lóscio et al., 2017), which are the following: title, description, keywords, date of last modification, date of publication, publisher, domain (theme) and distribution. The other ones are new and part of our proposal.

Table 1: DM(d).

Metadata	Vocabulary Term
Identifier	dcterms:identifier
Title	dcterms:title
Description	dcterms:description
Keyword	dcat:keyword
Domain	dcat:theme
Domain.name	rdfs:label
Domain.URI	void:uriSpace
Domain.vocabulary	void:vocabulary
URL Address	dcat:landingPage
Date of last Modification	dcterms:modified
Date of publication	dcterms:issued
Publisher	dcterms:publisher
Version	owl:versionInfo
Distribution	dcat:distribution
Distribution.format	dcterms:format
Distribution.size	dcat:byteSize
Distribution.URL	dcat:downloadURL
Distribution.type	dcat:mediaType
Date of DSP creation	dcterms:created

The values of some descriptive metadata (identifier, title, description, keywords, URL address, date of last modification, date of publication, publisher, version, and distribution) are derived of the information extracted directly from datasets. However, it is not always possible to find keywords. Since they are used in other phases of the metadata generation process, it is possible to find the most frequent terms of a dataset using the TF-IDF metric (Targino et al., 2017). As a result, a set of dataset keywords is obtained.

Information about recommended domain vocabularies and the knowledge domain of a dataset are also not usually found. Thereby these two specific metadata are contributions of this work. Their definitions are established as follows.

Definition 4. Domain Identification (DI(d)) - DI(d) refers to the identification of a knowledge domain to which a dataset d belongs.

Examples of data knowledge domains are Music and Education. For the identification of a given domain, a semantic background knowledge can be used. In this work, we use the DBpedia ontology

since it provides the classification of a huge amount of knowledge domains. To this end, by using the dataset keywords, corresponding classes or properties of the ontology are identified. When a property is returned, its corresponding class is then identified. Thereby, a domain term is returned based on the most frequent class among all the obtained class results (Targino et al., 2017).

Definition 5. Recommendation of Domain Vocabularies (RDV(ɗ)) - RDV(ɗ) refers to a suggestion of domain vocabularies which are considered as appropriate for a dataset ɗ.

The vocabulary recommendation can, for example, help data conversion processes when transforming source data (e.g., in CSV formats) into target RDF ones. This recommendation makes use of an open vocabulary repository as background knowledge. During this process, vocabularies related to the identified dataset keywords are identified. Among the identified vocabularies, the ones that have the highest number of occurrences and which are preferably active are prioritized. As a result, the best ranked vocabulary is recommended as RDV(ɗ). If there are more than one vocabulary presenting the higher and same number of occurrences, all of them are provided as recommendations (Targino et al., 2017).

3.2 Structural Metadata

Structural metadata are defined as follows.

Definition 6. Structural Metadata (SM(ɗ)) - Structural metadata SM(ɗ) describe the internal structure of a dataset ɗ in terms of its properties.

In a DSP(d), the SM(ɗ) are organized through a specific scheme, as described in Table 2. To this end, each element of the scheme is semantically referenced by a recommended vocabulary term.

Table 2: SM(ɗ).

Metadata	Vocabulary Term
Number of properties	void:properties
Property	void:property
Property.name	rdfs:label
Property.type	dcterms:type

3.3 Quality Metadata

Since IQ is based on a set of criteria, we need do identify and define ways of assessing specific quality indicators related with datasets on the web. We define a Quality Indicator as follows.

Definition 7. Quality Indicator ($\mathbf{QI}(ɗ)_n$) - A quality indicator $QI(ɗ)_n$ represents a measurable characteristic of a dataset ɗ that is related with the quality of its data or metadata.

Quality indicators may provide information regarding data content, data meta-information, and human ratings that give indications about the suitability of datasets for some intended usage. We argue that a number of specific quality indicators may be assessed in order to produce IQ measures. IQ measures are defined in this work as Quality Metadata, as follows.

Definition 8. Quality Metadata ($\mathbf{QM}(ɗ)$) – Quality Metadata $QM(ɗ)$ correspond to specific quality criteria related with a given dataset ɗ, which are assessed from a set of quality indicators $QI(ɗ)$.

Considering issues related with datasets publication and consumption and also with benefits suggested by the W3C Best Practices (Lóscio et al., 2017), two IQ criteria have been proposed in this work, namely: (i) dataset comprehensibility, and (ii) dataset processability. These IQ criteria are explained in the following sections.

3.3.1 Dataset Comprehensibility

Dataset comprehensibility is an IQ criterion which takes into account quality indicators related with the understanding of a given dataset by humans. It is defined as follows.

Definition 9. Comprehensibility ($\boldsymbol{C}(ɗ)$) – The comprehensibility of a dataset ɗ, denoted by C(ɗ), is stated as the degree to which ɗ presents information that promotes or facilitates its understanding by human users. C(ɗ) is measured from a set of quality indicators $QI(ɗ)_n$ in such a way that:

$$C(ɗ) = \frac{\sum_{n=1}^{6} QI(ɗ)_{Cn}}{\#QI(ɗ)} \quad (1)$$

Where,

$QI(ɗ)_{Cn}$ is the value of a quality indicator related with $C(ɗ)$.

$\#QI(ɗ)$ is the number of quality indicators associated with $C(ɗ)$.

The idea is that humans can have a better understanding about a given dataset on the web if some quality indicators are provided. Information regarding the dataset structure and its descriptive metadata are examples of quality indicators which may help such comprehension.

The comprehensibility criterion C(ɗ) is measured from a set of six quality indicators $QI(ɗ)_{Cn}$, namely: (i) Descriptive metadata; (ii) Structural metadata; (iii) Descriptive metadata referenced semantically; (iv) Dataset in an RDF distribution; (v) Metadata in an RDF distribution; and (vi) Contact point. They are described in the following.

$QI(ɗ)_{C1}$: Descriptive metadata

$$QI(ɗ)_{C1} = \frac{\#\mathrm{DM}(ɗ)_D}{\#\mathrm{DM}_D} \quad (2)$$

Where,

$\#\mathrm{DM}(ɗ)_D$ is the amount of desirable descriptive metadata (DM_D) found in dataset ɗ;

$\#\mathrm{DM}_D$ is the amount of desirable descriptive metadata that can describe the dataset ɗ.

For a dataset ɗ, a certain quantity of descriptive metadata (DM) should be made available. The idea is to describe a dataset overall features, in such a way that its description may improve its understanding. Based on this idea and in accordance with suggestions provided by the W3C (Lóscio et al., 2017), the following descriptive metadata are considered as desirable to a dataset, namely: title, keywords, URL address, date of publication, date of last modification, publisher, and its distribution.

$QI(ɗ)_{C2}$: Structural metadata

$$QI(ɗ)_{C2} = \frac{\#\mathrm{SM}(ɗ)}{\#ƥ(ɗ)} \quad (3)$$

Where,

$\#\mathrm{SM}(ɗ)$ is the amount of structural metadata (SM) provided by the description of dataset ɗ;

$\#ƥ(ɗ)$ is the number of properties which exists in ɗ structure .

Structural metadata (SM(ɗ)) should be made available to describe the properties that compose a dataset ɗ. Information about all the properties of a dataset should be provided.

$QI(ɗ)_{C3}$: Descriptive metadata referenced semantically

$$QI(ɗ)_{C3} = \begin{cases} 1 \\ 0 \end{cases} \quad (4)$$

Where,

1 means that the desired descriptive metadata (DM_D) provided by dataset ɗ are semantically referenced by recommended vocabularies;

0 means that the desired descriptive metadata (DM_D) provided by dataset ɗ are not semantically referenced by recommended vocabularies.

Desirable descriptive metadata (DM_D) provided by a dataset ɗ should be referenced by recommended vocabularies. Such provided semantics increases the metadata discovery and consumption capacity. One of the current recommendations regards the use of the DCAT vocabulary.

$QI(ɗ)_{C4}$: Dataset in an RDF distribution

$$QI(ɗ)_{C4} = \begin{cases} 1 \\ 0 \end{cases} \quad (5)$$

Where,

1 Dataset ɗ is available in an RDF distribution;

0 Dataset ɗ is not available in an RDF distribution.

A dataset should be also available in a distribution format with an RDF serialization. In RDF, each resource and vocabulary are identified by URIs, which eliminate ambiguities. Also, data in RDF are semantically described and have usually named links. This practice provides not only the right meaning of a resource but also possible relationships with other ones.

$QI(ɗ)_{C5}$: Metadata in an RDF distribution

$$QI(ɗ)_{C5} = \begin{cases} 1 \\ 0 \end{cases} \quad (6)$$

Where,

1 Metadata is available in an RDF distribution;

0 Metadata is not available in an RDF distribution.

The metadata corresponding to dataset ɗ are available in an RDF format. The RDF model is indicated for the representation of metadata, since it allows to formally define its semantics. It also facilitates the location and access of datasets.

$QI(ɗ)_{C6}$: Contact point

$$QI(ɗ)_{C6} = \begin{cases} 1 \\ 0 \end{cases} \quad (7)$$

Where,

1 A contact point is provided by the dataset publisher, thus enabling data consumers to get in touch;

0 A contact point is not provided by the publisher.

The data publisher should provide means to easy communication with dataset consumers. As an illustration, emails or contact forms may be provided.

3.3.2 Dataset Processability

In order to enable machines to automatically process data within a dataset, it is important that a dataset publisher applies some good practices (Lóscio et al., 2017). Examples of good practices are the following: providing structural and descriptive metadata; making use of recommended vocabularies; providing data in more than one machine readable format; and making data available through APIs. In accordance with these recommendations, the dataset processability criterion has been established as follows.

Definition 10. Processability (P(ɗ)) - The processability of a dataset ɗ, denoted by P(ɗ), measures the degree to which ɗ is processable by machines or software agents. P(ɗ) is assessed from a set of quality indicators $QI(ɗ)_n$ in such a way that:

$$P(ɗ) = \frac{\sum_{n=1}^{5} QI(ɗ)_{Pn}}{\#QI(ɗ)} \quad (8)$$

Where,

$QI(ɗ)_{Pn}$ is the value of a quality indicator of P(ɗ).

$\#QI(ɗ)$ is the number of quality indicators of P(ɗ).

The processability criterion P(ɗ) is measured from a set of five quality indicators, namely: (i) Data API; (ii) Structural metadata; (iii) Distributions in machine readable formats; (iv) Dataset download; and (v) Dataset in more than one distribution. They are explained in the following:

$QI(ɗ)_{P1}$: Data API

$$QI(ɗ)_{P1} = \begin{cases} 1 \\ 0 \end{cases} \quad (9)$$

Where,

1 A data API is available for access to dataset ɗ;

0 A data API is not available for access to dataset ɗ.

Among the distributions of a dataset, an API should be made available. This makes data processing and accessibility more feasible. It also provides means to use real-time data.

$QI(ɗ)_{P2}$: Structural metadata

$$QI(ɗ)_{P2} = \frac{\#SM(ɗ)}{\#þ(ɗ)} \quad (10)$$

Where,

$\#SM(ɗ)$ is the amount of structural metadata (SM(ɗ)) made available by dataset ɗ;

$\#þ(ɗ)$ is the number of properties found in ɗ structure.

As explained in $QI(ɗ)_{C2}$, Structural metadata (SM(ɗ)) are relevant information when trying to easy comprehension and processability of datasets.

$QI(ɗ)_{P3}$: Distributions in machine readable formats

$$QI(ɗ)_{P3} = \begin{cases} 1 \\ 0 \end{cases} \quad (11)$$

Where,

1 The dataset ɗ is available in distributions with machine readable file formats;

0 The dataset ɗ is not available in distributions with machine readable file formats.

A dataset ɗ is usually made available through distributions. Distributions must make data available in file formats that are more easily processed by machines, such as JSON, RDF, and CSV. This good practice may help applications to process data.

$QI(ɗ)_{P4}$: Dataset download

$$QI(ɗ)_{P4} = \begin{cases} 1 \\ 0 \end{cases} \quad (12)$$

Where,

1 Allows the download of dataset ɗ;

0 Does not allow the download of dataset ɗ.

Dataset ɗ is available for download, increasing the possibilities of processing and using its data.

$QI(ɗ)_{P5}$: Dataset in more than one distribution

$$QI(ɗ)_{P5} = \begin{cases} 1 \\ 0 \end{cases} \quad (13)$$

Where,

1 The dataset ɗ is available in more than one distribution;

0 The dataset d is not available in more than one distribution.

Dataset d should be available in more than one distribution. For example, it may be provided in open formats such as XML, RDF, JSON and/or CSV. This practice increases the chances of data consumption, since data consumers may have preferences on the data formats.

4 IMPLEMENTATION AND EXPERIMENTS RESULTS

This section presents some results regarding the implementation of the proposed approach and accomplished experiments.

4.1 The DSPro+ Tool

The DSPro+ tool allows the user to automatically generate, view and download a DSP. The main functionalities available through the tool are summarily described as follows.

- **Search Dataset:** It provides a keyword-based method for searching one or more datasets that have previously had their DSP generated.
- **Query Dataset Information:** The tool can retrieve information about a selected dataset.
- **Validate Dataset URL:** Before starting the DSP generation process, the user must provide the URL of a dataset. The tool then verifies if it is a valid URL which can enable the extraction of its underlying information.
- **Generate DSP:** After validating a dataset URL, the user can start the process of generating a DSP. Information regarding the dataset and the results will be depicted at the end of the process.
- **View DSP:** The user can view a DSP generated by the tool, which contains descriptive, structural and quality metadata of a given dataset.
- **Download DSP:** The user can download a DSP in an RDF format.

As an example of usage, consider a dataset called "Rolling Stone's 500 Greatest Albums of All Time" (Kaggle, 2018), which is available in CSV. Most descriptive metadata can be directly extracted from the dataset at hand. Thus, the dataset title, identifier, description, keywords, URL address, date of last modification, date of publication, publisher, version, and distribution are obtained and saved for later use.

The recommendation of domain vocabularies to that dataset is then performed. To this end, SPARQL queries are built and executed in order to find corresponding vocabularies properties to the dataset properties or keywords. These queries are executed on the Linked Open Vocabularies endpoint (LOV, 2018), i.e., a web service provided by that open vocabulary catalog. In this current example, the vocabulary that presented the highest number of occurrences among the obtained queries results was "The Music Ontology". Thereby, this vocabulary is the one to be recommended to the example dataset.

For the knowledge domain identification, dataset keywords are also used to compose SPARQL queries. These queries are now accomplished on the public DBPedia endpoint. As a result, a class named "Musical Work" is identified as the knowledge domain to the dataset, which had the highest number of occurrences among the obtained class results.

According to the extracted dataset information, the properties that compose its structure are identified. For a dataset in CSV format, the header line of the file is identified and its columns names are extracted. Regarding the example, six properties have been identified: number (numeric), year (numeric), album (string), artist (string), genre (string), subgenre (string). Thus, the dataset structural metadata are produced.

Quality metadata are also generated in accordance with the information collected from the dataset. Each quality criterion is assessed according to its set of quality indicators. Each quality indicator is measured based on the information provided by or together with the dataset (e.g., when there is a profile). Then, corresponding results from each one are put together in Formula 1 and Formula 8. Regarding the IQ criterion Comprehensibility, among the quality indicators for the dataset at hand, the great majority of them were completely met. Only the indicator concerned with the existence of the dataset as an RDF distribution was not found. Thereby C(d) was measured as 0.83. Referring to Processability, quality indicators with respect to providing a data API and more than one distribution were not met. The other indicators were successfully served. Consequently, P(d) was measured as 0.6.

The DSP metadata generated are made available in RDF format. Figure 2 shows a fragment of the generated DSP for the example at hand. This fragment corresponds to some descriptive metadata (identifier, dataset title, description, keywords,

identified domain, recommended vocabulary) and the defined quality metadata (i.e., comprehensibility and processability criteria), where each criterion is instantiated and specified with its label, value and definition. The obtained DSP uses recommended vocabularies to refer the metadata, thus enabling its consumption or publication along with the dataset. This feature may be provided in Open Data portals, for example.

```
:dataset.ID629.Kaggle
   a       dcat:Dataset ;
   dcterms:identifier "629" ;
   dcterms:title "Rolling Stone's 500 Greatest Albums
                  of All Time" ;
   dcterms:description "From ... try to fix them." ;
   dcat:keyword "Music" , "Humanity" , "Performing Arts" ,
                "Critical Theory" , "Culture" ;
   dcat:theme :theme-Musical_Work ;
   void:vocabulary "http://purl.org/ontology/mo/" ;
   dqv:hasQualityMetadata  :dimensionComprehensibility ,
                           :dimensionProcessability ;

:theme-Musical_Work
   a       dcat:theme ;
   rdfs:label "Musical_Work"@en ;
   void:uriSpace "http://dbpedia.org/ontology/MusicalWork" .

:dimensionComprehensibility
   a       dqv:Dimension ;
   rdfs:label "Comprehensibility"@en ;
   dqv:value "0.83"^^xsd:float ;
   skos:definition "Represents the degree to which a
                    dataset presents information that
                    promotes or facilitates its
                    understanding by human users."@en .

:dimensionProcessability
   a       dqv:Dimension ;
   rdfs:label "Processability"@en ;
   dqv:value "0.6"^^xsd:float ;
   skos:definition "Represents the degree to which a
                    dataset is processable by machines
                    or software agents."@en .
```

Figure 2: Fragment of a generated DSP for the example, corresponding to some DM(d) and QM(d).

4.2 Experiments

We have conducted some experiments to verify the effectiveness of our approach. To this end, we have used 30 CSV datasets provided in the English language. They have been divided into three groups of 10 datasets, where each one belongs to the following knowledge domains: “Video Games”, “Automobiles”, and “Music”. These datasets are made publicly available on Kaggle (Kaggle, 2018). The data and metadata from these datasets can be accessed through their page. Some descriptive metadata are represented in JSON-LD scripts, which uses the vocabulary Schema.org.

The proposed DSP+ approach takes into account different aspects (e.g., knowledge domain identification, dataset quality) in order to produce a DSP. Thus, it was not possible to find a single baseline that could be used to accomplish experiments and compare results. Instead, four kinds of evaluations have been defined and accomplished, as follows.

The goal of the first experiment is to verify whether the generated DSP represents a more comprehensive and semantically rich description of a given dataset. To this end, the original descriptive set of metadata (made available through JSON-LD scripts in datasets) is compared with the produced metadata of the DSPro+ approach. As shown in Figure 3, a larger amount of metadata is provided from the generated DSP. Furthermore, these generated metadata consider more aspects of a dataset, generating new descriptive metadata (e.g., domain and recommended vocabulary), as well as structural and quality ones. In the available JSON-LD scripts, only descriptive metadata are provided. Usually they regard to the dataset title, description, identifier, version, and some information, such as a set of comments and number of downloads.

For the second and third experiments, gold standards have been manually produced by domain users of our group. In order to verify the results, we have used the traditional Precision, Recall and F-Measure metrics (Baeza-Yates et al. 1999).

The second experiment aims to check whether the automatic identification of the knowledge domain of a dataset presents a similar result in comparison with its manual identification by a human. It has been observed that obtained keywords belonging to datasets of a same knowledge domain group can have differences among them and this can directly affect the results. During the comparative analysis, it was found that among the 30 datasets used in the experiment, only four of them did not receive the recommendation according to the domain recommended by the specialists. As depicted in Figure 4, considering the domain identification for datasets on “Video Games”, all of them corresponded to the gold standard, resulting in a value of 1 for Precision, Recall and F-measure. For the “Music” group, two datasets have not received the expert-defined domains, resulting in a value of 0.8 for Precision, Recall, and F-Measure metrics. In the “Automobile” group, eight datasets presented the identified domains in accordance with the gold standards, but for one dataset does not have been identified more than one domain as expected according to the gold standard. The following values were obtained: Precision = 0.8, Recall = 0.67 and F-measure = 0.73.

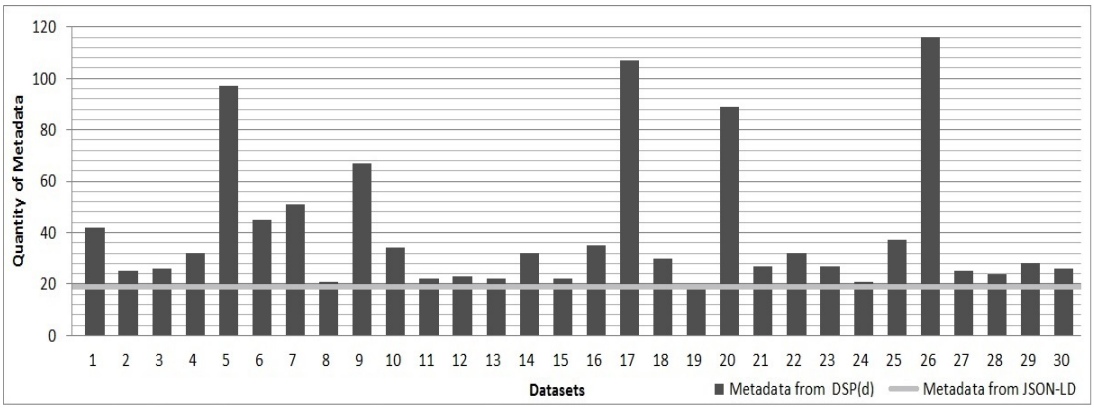

Figure 3: Comparison between the metadata provided by the generated DSPs and the metadata provided by the datasets.

The third experiment verifies whether the automatic recommendation of domain vocabularies for a dataset presents a similar result when compared with the recommendation provided by humans. It was observed that, in many cases, the approach recommends more than one vocabulary. This also occurred among the gold standards generated by the experts. However, due to the number of vocabularies that are related to the same knowledge area, among the results obtained for each dataset, not all the recommended vocabularies sometimes correspond to all the gold standard vocabularies. As shown in Figure 5, among the results of the "Video Games" group, the vocabularies defined as the gold standard were recommended for all datasets, but except for one dataset, only one vocabulary was recommended diverging from the gold standard, resulting in Precision = 1, Recall = 0.91 and F-Measure = 0.95. Considering the datasets of the "Music" group, seven of them obtained the recommendation of vocabularies defined as the gold standard. However, for some of them, none of the expected vocabularies were recommended, resulting in Precision = 0.73, Recall = 0.85 and F-Measure = 0.78. Among the datasets of the "Automobile" group, only four datasets presented some vocabulary corresponding to the one recommended by the specialists, with values of Precision = 0.21, Recall = 0.31 and F-Measure = 0.25. In these obtained results, it has been noted that when the dataset belongs to a more specific knowledge area, its properties and keywords are very specialized. Thus, most suitable vocabularies are found, making it more likely to achieve the expected outcomes, as observed in the light of the "Video Games" group.

The fourth experiment intends to measure the degree of comprehensibility and processability of the datasets. To this end, we have performed two evaluations: (i) considering information originally provided by the datasets and measuring both IQ criteria and (ii) considering the information originally provided by the datasets plus the information provided by the generated DSP and measuring both IQ criteria. In these evaluations, a significant improvement was observed between the values assigned to the quality criteria after the profile generation. Then, we compared the obtained results.

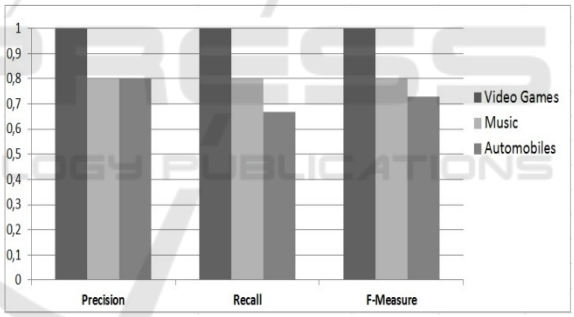

Figure 4: Measures w.r.t. the Domain Identification.

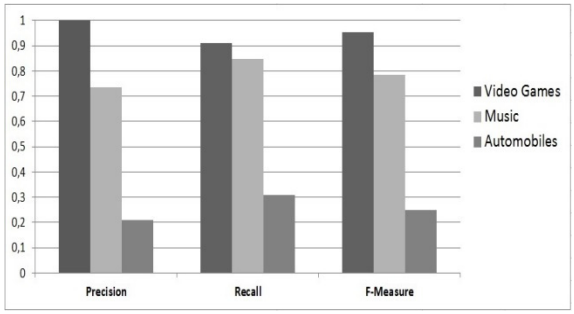

Figure 5: Measures w.r.t. the Vocabulary Recommendation.

As shown in Figure 6, for the comprehensibility criterion of a dataset, an improvement of at least 33% in the value received after the DSP generation

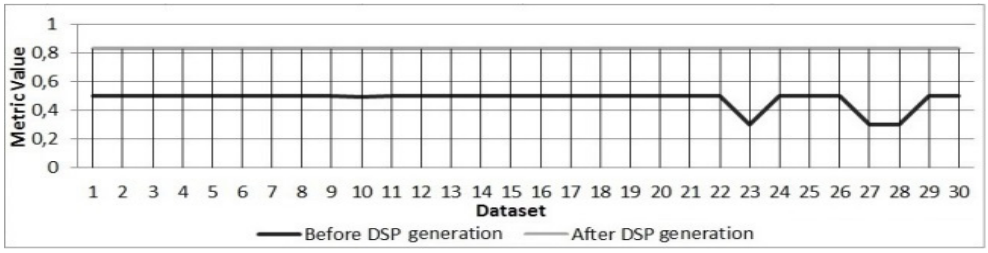

Figure 6: Dataset Comprehensibility Measures before and after the generation of DSPs.

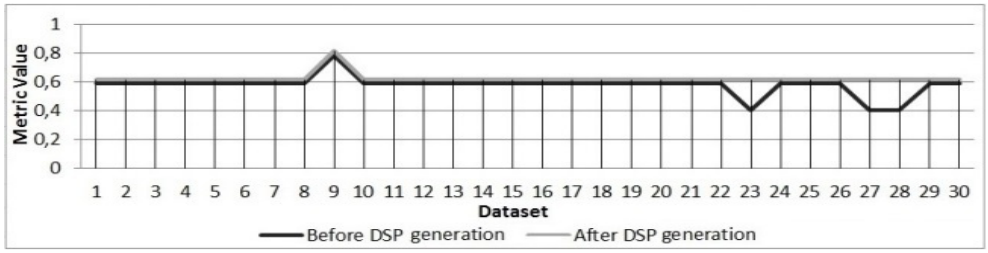

Figure 7: Dataset Processability Measures before and after the generation of DSPs.

has been obtained. This occurred because before the DSP generation there were no metadata available in RDF format and the metadata were not available using recommended vocabularies.

For the processability criterion of a dataset (Figure 7), no significant improvements were observed when analyzing the datasets as a whole, since most datasets already provided structural metadata on their page. For this criterion, improvements were observed when before the DSP generation there were no structural metadata provided by the dataset. Thus, it resulted in a small variation between the outcomes obtained before and after the DSP generation. However, when a specific dataset is observed, significant improvements are identified. For example, dataset 23, that presented a processability of 0.4 before the generation of DSP, improved to a value of 0.6 after the profile production, resulting in 20% of improvement in the processability criterion.

5 RELATED WORKS

The generation of datasets profiles with enriched metadata is an activity composed of different phases, which consider the generation of descriptive, structural and quality metadata. Considering semantic enrichment, more specifically the recommendation of vocabularies, the works of Ellefi et al., (2015) and Schaible et al., (2013) present approaches for recommending vocabularies for each concept of a data source. In our work, vocabularies are identified associated with the properties of the dataset, but what is recommended are domain vocabularies, without the need of user assistance.

Among the works related to domain identification, the works of Ouksili et al. (2014) and Lalithsena et al. (2013) present approaches to identify the domains of datasets. However, in these works it is possible to identify the domain only for datasets in RDF format. Also, metadata about the identified domain that could be made available to users are not generated.

Considering the works that use IQ criteria to evaluate the quality of datasets, in the work of Assaf et al. (2016) a framework was developed to evaluate the quality of connected datasets through quality criteria and indicators. Although the quality framework is related to the generation of a profile, the results obtained related to the dataset quality are not inserted in the profile. Also the processability criterion is not considered.

In terms of dataset generation, in the work produced by Abele (2016), two approaches are proposed for the generation of metadata representing the content of the datasets and for the identification of connections between the datasets. The work considers only datasets in RDF format and, for the generation of metadata, only descriptive and structural aspects are considered. The work presented by Assaf et al. (2015) proposes an approach for extracting, validating, correcting and generating data profile, which is generated in JSON format. However, structural or quality metadata are not generated. When comparing works related to

DSP generation, such as Abele (2016) and Assaf et al. (2015), an approach which provides more detailed information about datasets, including descriptive, structural, and quality metadata is not found. In addition, some of them do not use vocabulary terms associated to the metadata provided by the profile. This allows to assign more meaning and a representation of the metadata which facilitates its consumption.

6 CONCLUSIONS

In this work, we have presented an approach for the generation of semantically enriched Dataset Profiles. To help matters, a DSP composed of descriptive, structural and quality metadata is proposed. During the DSP generation process, some metadata are extracted from the datasets, and, additionally, the dataset domain is identified and domain vocabularies are suggested. Furthermore, the process includes the generation of structural metadata and quality metadata, which proposes two IQ criteria to be measured as relevant and additional information. The main idea of providing enriched DSPs is to facilitate the communication between dataset publishers and consumers (humans and machines).

In order to evaluate the proposed approach, a prototype has been implemented. It provides an automatic DSP generation process. The tool assists data producers who wish to make DSPs available to certain datasets. Dataset consumers can also generate a DSP, without the need of prior knowledge about the data.

The experiments used datasets from different knowledge domains. They demonstrated that the proposed strategy produces good results, by allowing the generation of new metadata. Improvements were also observed with respect to the quality of the datasets after the DSP generation.

As future works, we consider to include user feedback and other IQ criteria (e.g., completeness, correctness), to link the approach to an existing dataset catalog, and also to include in the DSP the recommendation of vocabularies for each identified structural metadata. New experiments with expert users and datasets belonging to a wider range of domains will also be accomplished.

REFERENCES

Abele, A., 2016. Linked Data Profiling: Identifying the Domain of Datasets Based on Data Content and Metadata, In: *25th International Conference Companion on World Wide Web.* Canada, p. 287-291.

Assaf, A., Senart, A., Troncy, R., 2016. An Objective Assessment Framework & Tool for Linked Data: Enriching Dataset Profiles with Quality Indicators, In: *IJSWIS, International Journal on Semantic Web and Information Systems*, Special Issue on Dataset Profiling and Federated Search for Linked Data, Vol. 12, N°3, 2016, ISSN: 1552-6283

Assaf, A., Troncy, R., Senart, A., 2015. Roomba: An extensible framework to validate and build dataset profiles, In: *24th International Conference on World Wide Web*, Italy, p. 159-162.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. *Modern Information Retrieval.* Addison-Wesley, First Edition.

Clarke, M., Harley, P., 2014. How smart is your content? Using semantic enrichment to improve your user experience and your bottom line, *Science Editor*, Vol. 37, N° 2, p. 40–44.

Ellefi, M. B., Bellahsene, Z., Scharffe, F., Todorov, K., 2014. Towards Semantic Dataset Profiling In: *International Workshop on Dataset Profiling & Federated Search for Linked Data co-located with the 11th Extended Semantic Web Conference.* Greece.

Ellefi, M. B., Bellahsene, Z., Todorov, K., 2015. Datavore: a vocabulary recommender tool assisting Linked Data modeling, In: *14th International Semantic Web Conference*, Posters and Demonstrations Track, United States.

Flemming, A. (2011). *Quality Characteristics of Linked Data Publishing Datasources*. Master's Thesis, Humboldt-Universität zu Berlin, Institut für Informatik.

Heath T., Bizer C., 2011. Linked Data: Evolving the Web into a Global Data Space, 1st edition. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

Kaggle platform, 2018. Available at https://www.kaggle.com. Last access on june, 20th.

Lalithsena, S., Hitzler, P., Sheth, A. P., Jain, P., 2013. Automatic Domain Identification for Linked Open Data. In: *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, United States, p. 205-212.

Lóscio, B. F., Burle, C., Calegari, N., 2017. Data on the web best practices. W3C, Version: https://www.w3.org/TR/2017/REC-dwbp-20170131/ Last Acess: march 20, 2018.

LOV, 2018. Linked Open Vocabulary Repository. Available at https://lov.okfn.org/dataset/lov/. Last access on June 20th.

Naumann, F., Rolker, C., 2000. Assessment methods for information quality criteria In: *IQ, 5th Conference on International Quality.* United States, p. 148-162.

Ouksili, H., Kedad, Z., Lopes, S., 2014. Theme Identification in RDF Graphs, In: *MEDI*, *International Conference on Model and Data Engineering*. Cyprus, p. 321-329.

Pipino, L. L., Lee, Y. W., Wang, R. Y. (2002) Data Quality Assessment. In: *Communications of the ACM*

- *Supporting community and building social capital*, Vol. 45, N°.4, April 2002, p. 211-218.

Schaible, J., Gottron, T., Scheglmann, S., Scherp, A., 2013. LOVER: support for modeling data using linked open vocabularies. In: *Joint EDBT/ICDT*. Italy, p. 89-92.

Targino, N., Souza, D., Salgado, A. C., 2017. Uma Proposta de Perfil de Conjuntos de Dados na Web com Enriquecimento Semântico. In: *SBBD, 32nd Brazilian Symposium on Databases*, Brazil, p. 172-183.

Wang, R. Y., Strong, D. M., 1996. Beyond accuracy: What data quality means to data consumers. In: *Journal of Management Information Systems*, United States, Vol. 12, N° 4, p. 5–33, ISSN 0742-1222.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., 2013. Quality Assessment Methodologies for Linked Open Data. In: *Semantic Web Journal*.