Computer Science and Information Systems 17(2):599–617 https://doi.org/10.2298/CSIS190420001M

Production of linked government datasets using enhanced LIRE architecture

Nataša Veljković[1], Petar Milić[2], Leonid Stoimenov[1], and Kristijan Kuk[3]

[1] Faculty of Electronic Engineering, Aleksandra Medvedeva 14,
18000 Niš, Serbia
{natasa.veljkovic, leonid.stoimenov}@elfak.ni.ac.rs
[2] Faculty of Technical Sciences, Knjaza Miloša 7,
38220 Kosovska Mitrovica, Serbia
petar.milic@pr.ac.rs
[3] University of Criminal Investigation and Police Studies, Cara Dušana 196,
11000 Belgrade, Serbia
kristijan.kuk@kpu.edu.rs

Abstract. This paper describes the enhanced LIRE (LInked RElations) architecture for creating relations between datasets available on open government portals. The architecture is improved to be applicable on different open government data platforms using minimal configuration at the Data processing layer. We evaluated the applicability of enhanced LIRE, its advantages and disadvantages, which resulted in necessary recommendations for the publication of dataset's metadata to obtain better associations between datasets. Moreover, we introduced a LINDAT indicator that reflects the percentage of linked data in the total possible number of linked data on open government data portals.

Keywords: open government data, linked data, datasets, metadata, dataset relations.

1. Introduction

WEB of data represents related structured data connected through links [1]. Offering open government data (OGD) in the Web of data will ease their discovery and usage. That implies interlinking already published data, which can be a tedious and expensive task for governments, but we argue that benefits prevail. Government moves to the higher degree of openness implicitly by progressing in the domain of open data [2]. Not to forget the final users, who appreciate better accessibility, reusability and easy processing of open data [3].

The real value of OGD is revealed with linking which provides unexpected and unexplored insights into different domains and problem areas [4]. OGD interconnected with Web of Data becomes Linked Open Government Data (LOGD). Processing dataset's metadata to create links makes datasets more comprehensible and leads to uncovering potential faults in metadata definition. Such process contributes to the quality of datasets thus raising the openness and transparency in government.

Government itself is a complex engine running on highly distributed set of institutions. Linking data within and between those institutions enables government to publish data in a modular way, benefiting from a ‘small pieces of loosely joined’ approach to government data [5]. However creating relations between OGD from the government perspective is a costly process, it requires trained and equipped staff for OGD processing, which governments do not have. In that sense, sharing insights on dataset relations between government institutions could help to avoid duplicate work and efforts [6].

User behavior while searching for OGD may help in revealing practical approaches and patterns for linking OGD. OGD users usually browse multiple datasets, while doing so they discover new datasets that may be related to the ones they have already checked. If open data portals provided a way for users to propose relations between datasets, this would create more linked data on the portals without much effort from the government side.

When users build applications with raw OGD they need to inspect government datasets to see if they correspond to their needs. Keeping this in mind, we can ask ourselves: What can be helpful in driving the users toward the dataset they are interested into? We believe the answer is dataset’s metadata. Dataset’s metadata, such as description, format, tags and etc. enable users to inspect if a dataset contains information they need. Metadata helps in generating linkable dataset profiles which is important when relating datasets [7].

This paper analyses how dataset’s metadata can be used for the production and utilization of linked government data. Using the enhanced LIRE (Linked Relations) architecture we demonstrate the process of linking datasets based on the metadata keys and their values. The architecture is flexible and generalized for application on different OGD portals. Moreover, we are contributing to the linked datasets representation on OGD portals by proposing a model for semantic representation of dataset’s relations. Based on the user involvement in the process of consumption of OGD and creating useful applications that depend on OGD, we got an idea to include users in the process of relating datasets. Our approach goes towards the interlinking and integration of OGD. As a proof of concept, we have developed a prototype that enables the users of OGD portals to create linked datasets.

Throughout this article we explained the enhanced LIRE architecture, processes which occur inside the architecture, the specification of data necessary to relate datasets and the semantics of relations. We introduced and evaluated a new measure for reflecting the status of the linking of datasets on OGD portal named LINDAT. Furthermore, we analyzed applicability of the architecture, its advantages and disadvantages, and gave necessary recommendations for publication of dataset’s metadata to obtain better assessment of dataset which we want to link.

The rest of the paper is organized as follows. In Section 2 we analyze current approaches in the area of linking OGD, focusing on those that require metadata utilization or user involvement. Section 3 describes the enhanced LIRE architecture for linking government datasets based on metadata structure. In Section 4 we analyze LINDAT indicator and apply it on CKAN powered OGD platforms to check if open government data portals have linked data. The final section concludes the work done so far and presents ideas for the future work.

2. Related Work

In the following subsections, we present available methods and tools for linking open data in the government domain (LOGD) and differentiation of those approaches to our proposal. We talk about the structure of linked datasets (linksets) on OGD portals, and how our approach contributes to that. To lower the costs associated to datasets exploration and linking, users are often involved in the process of linksets creation, and it was interesting to give an overview of tools which assist in the process and talk about our contribution in this domain.

2.1. Linking government datasets

When publishing OGD, data producers strive to fulfil openness criteria, but they are not so concerned with having the linked criteria fulfilled [3]. LOGD is the practice initially adopted by researchers and third parties who have reused existing open data to create linked open data, by exploring datasets, creating RDFs and publishing them on the Web as new data. Nowadays it happens more often that data producers are interested in publishing linked data as well. An interesting analysis of provision of LOGD data is given by Kalampokis et al. [8]. They show that linked data can be provided in two ways: by publishing on the central open government portal, or on the public agency portal. In the second case, central government portal collects metadata about LOGD datasets published by a public agency in order to make that datasets available on its portal. That is the so-called ‘indirect provision of linked data’. This claim is in line with [9] who proposed the architecture for integrating datasets from public agencies via activities and components essential for discovery. Their architecture has two phases, in the first phase, work on downloading and transforming datasets into a common schema language format is conducted, while the second phase addresses semantic heterogeneity with schema matching and statistical analysis of ontology structures. This paper does not take into account the metadata of open datasets, but deals solely with their semantic versions. In relation to our solution, which we will describe in this paper and it is based on the application of metadata, the authors do not extract useful information from "raw" open datasets that can serve to relate them.

On the other hand, some authors claim that successfully linking government datasets requires understanding the data context’s sensitive meaning [10]. This is coupled with enabling the semantic interoperability of OGD and provision of metadata that describes them, which is important for creation of value added applications. It is important that there is a standardized level of metadata, because that will allow the harmonization of specific concepts and terminology, interoperability and multilinguality in government open data portals. For that purpose, Assaf, Troncy and Senart [11] have identified a need for a definition of a harmonized model of OGD dataset’s metadata which contains sufficient information so that consumers can easily understand and process datasets. That information contains general information, access information, ownership information, provenance information, geospatial information, temporal information, statistical information and quality information with mappings between appropriate data.

Similarly, Kashyap and Sheth’s work on the semantic heterogeneity is based on representation of metadata, context and ontologies. They claim that for the proper linking of OGD, underlying data and capturing of the meaning of domain specific metadata must be considered. This is due to the information overload which arises as a consequence of the heterogeneity of the digital data [12]. In relation to our work, this approach collects different types of metadata independent of type, representation and location, while we utilize metadata from OGD which is not previously curated.

Schmachtenberg et al. [13] presented an overview of relationships between linked datasets in the form of linked open data cloud diagram. Analyzing that, we can notice that two datasets can be linked if there exists at least one RDF link between resources belonging to the datasets. The authors emphasize the use of dataset's metadata because it reveals the origin of datasets and can be used to analyze dataset's quality.

For the integration of OGD datasets into the Web of Data, a group of authors [14] introduce a solution of six stages of dataset integration. Five of these stages named Name, Retrieve, Adjust, Convert, Enhance and Publish are designed to implement an approach for minimization of human effort to incorporate new dataset as linked data, while in the remaining stage the structured and connected descriptions of the initial representations of datasets are added by data modelers. With this approach data structural relations are covered, which supports integration of LOGD from multiple sources. Cverdelj-Fogaraši at al. provides alignment of domain specific metadata ontologies, without modelling relations between OGD datasets [15].

In the work of Scharffe et al. [16] on methods for automated dataset interlinking, different data linking systems are discussed. The authors showed that with an appropriate mapping between dataset's descriptions proper links can be established by denoting the single correspondence between object descriptions. This claim is confirmed by the research of Ellefi et al. [17] who introduced an approach for linking datasets by overlapping dataset's metadata schema definition. For a given dataset, their framework allows identifying the datasets sharing schema with other datasets, which is a useful input for the data linking step. Furthermore, Ngomo and Sherif offer a solution for link discovery between different datasets [18]. Their solution addresses time-efficiency and accuracy challenges which is of central importance when a tool is faced with small amounts of RAM or when is faced with streaming or complex data (e.g., 5D geospatial data). As data alone has little value, to unleash its full potential, it needs to be linked with other referenced data. Datalift tool provided by [19] searches for relationships between local government data and existing public RDF data and enriches that data with links, making them discoverable on the Web of data. Compared to the approach we describe in this paper, the authors integrate ready-made semantic versions of OGD datasets with the rest of the Web of data, while our solution aims to integrate OGD datasets across portals first and then integrate them into the Web of data.

2.2. Representation of linked data on OGD portals

The full potential of OGD has not been realized, and one of the main reasons for that is related to the provision of metadata. Metadata provide documentation, context and necessary background information for interpreting OGD [20], and as such it is often considered as the key enabler for the effective use of linked open data in government domain [21]. To enable the interoperability of OGD and their linking within the Web of Data, they need to be modelled in a semantic way which ensures that data will be reached by linked data applications. In this respect, short summarization of most used dataset vocabularies to describe linked datasets follows.

The analysis of dataset's metadata structure, consistency and availability conducted by [22], have resulted in development of dcat (data catalog) vocabulary that allows the expression of datasets in the RDF data model. Dcat enables datasets to be queried in RDF, supports the reuse and extension of existing metadata standards such as Dublin Core [23] and compatibility with linked data. It avoids the use of ontologies, because the main purpose of the vocabulary is interoperable data exchange. According to the authors, the goal of this vocabulary is to increase dataset's discoverability enabling applications to easily consume them.

For describing public sector datasets in Europe, a dcat-ap (data catalog - application profile) vocabulary is designed [24]. This vocabulary is intended to represent a selected set of properties from those in dcat and its imported vocabularies, also enables cross-data portal searching and enhances discoverability. It is recommended by the Open Data Support to be the standard for describing linked datasets in Europe.

The VoID vocabulary differentiates linked data publishers, persons or organizations exposing linked data and on the other hand linked data consumers, which may be humans or machines [25] with defining "appropriateness" by following criteria: the content of the datasets, interlinking to other datasets and vocabularies used in datasets. [26]. The designed term 'linkset' deals with interlinking between datasets, especially those that are published on the same portal. With RDF properties: 'void:subset', 'void:target' and 'void:linkPredicate', VoID gives opportunity for dataset interlinking. By modelling dataset's links with VoID we actually connect one or more topics that are originating from a certain source or process and that are accessible on the Web. Fiorelli et al. [27] developed LIME (Linguistic Metadata), an extension of VoID with a vocabulary of metadata about the ontology-lexicon interface. LIME provides the lexicalization of VoID relations in a natural language aligning it with a set of lexical concepts. With LIME, dataset relations intend to be more discoverable, understandable and exploitable.

All of the presented vocabularies describe linked datasets and how they can be accessed through the linked data applications. Nevertheless, the question remains: what about relations between OGD datasets on portals? How to utilize them and reach more related data? In this regard, our solution proposes which elements of vocabularies should be exploited in order to resolve this issues and to make OGD dataset relations available in linked data applications.

2.3. User involvement in linking datasets

As pioneers in creating LOGD, UK and USA government initiatives of linking data have showed that availability of LOGD is costly, and that there is a need for solving this problem or at least to mitigate it. One of the means to reduce the costs is to allow datasets consumers to participate voluntarily in linking OGD. Self-service approach introduced by [28] shifts the burden of interlinking datasets to the data consumers. It encourages them to interlink government datasets without waiting for the government to do so. With the application of Google Refine tool [6] users interlink datasets to the Web of Data space, previously providing Google Refine with SPARQL endpoint or dump file for reconciliation of this datasets against any RDF data available through it. Similar to this, Li Ding et al. describe linked data ecosystem in which users manage and consume LOGD in connection with online tools, services and societies [3]. LOGD ecosystem is based on converting raw OGD datasets into linked data and their integration with other resources. The work of Li Ding et al. presumes the existence of linked data from OGD data as a base for their further linkage, but does not consider the possibility for the integration of the datasets on the same platform. Our work goes toward this, including users for achieving this goal due to the fact that successfully linking of datasets requires understanding the data's context sensitive meaning [10].

User feedback and application queries can be used to determine whether two datasets can be interlinked [29]. Application queries help filter datasets that are potentially strong candidates for interlinking, whereas user feedback is used as a way to assess the relevance of the candidate datasets. Furthermore, [30] argued that the visual exploration of dataset content in linked open data exploitation can help in identifying links between datasets. This approach is similar to the one we will present in the following sections, but does not take into account the fact that datasets housed on the same OGD platform can also be related mutually. Specifically, it deals with the semantics of linked data without analyzing the attributes of the datasets and getting users attention to the interaction and information use on OGD platform.

To the best of our knowledge, there is no similar work to what we propose, that deals with the production of interlinked government datasets based on its metadata. Some authors [8, 10, 13, 16, 20] discuss linking datasets based on their semantic description without going deeper into the metadata. That does not tackle information hidden in published OGD, which by our opinion can help in linking OGD datasets. While the semantic version of OGD can be linked to the Web of Data space, the semantic interlinking of OGD on same portals remains unexplored. Leme et al. [31] utilize existing dataset relations on open government portals to produce recommendation for dataset linking, but do not take into account metadata describing the dataset such as description, tags, formats, organization and etc., in order to search for possible hidden information that can assist in their connection and consequently their interlinking. Comparing to the [32], the version of architecture that will be presented in following section contains improved model for determining the type of relations between two datasets, applicability on different types of OGD platforms which consequently means that it is interoperable. Moreover, we define and describe a novel measure here, LINDAT indicator, for the purpose of monitoring the linking status of datasets.

3. Enhanced LIRE architecture

Based on the aforementioned approaches for linking OGD, we got the idea to explore dataset's metadata to check whether they can be used to relate to other OGD, and then to model these relations semantically in order to produce LOGD. For that purpose, we have developed the architecture for managing relations between datasets, and their linking based on user interaction with OGD portal. LIRE architecture enables the interlinking of datasets [32], and can enrich OGD portals with novel functionality. Our architecture assumes the existence of the semantic version of OGD datasets, as the availability of such OGD datasets is embedded in most OGD platforms, to mention few: CKAN[1], DKAN[2], Opendatasoft[3] and Socrata[4]. The prototype exists in a form of a plugin, currently available only for the CKAN platform[5]. The enhanced LIRE architecture is outlined in Fig. 1.

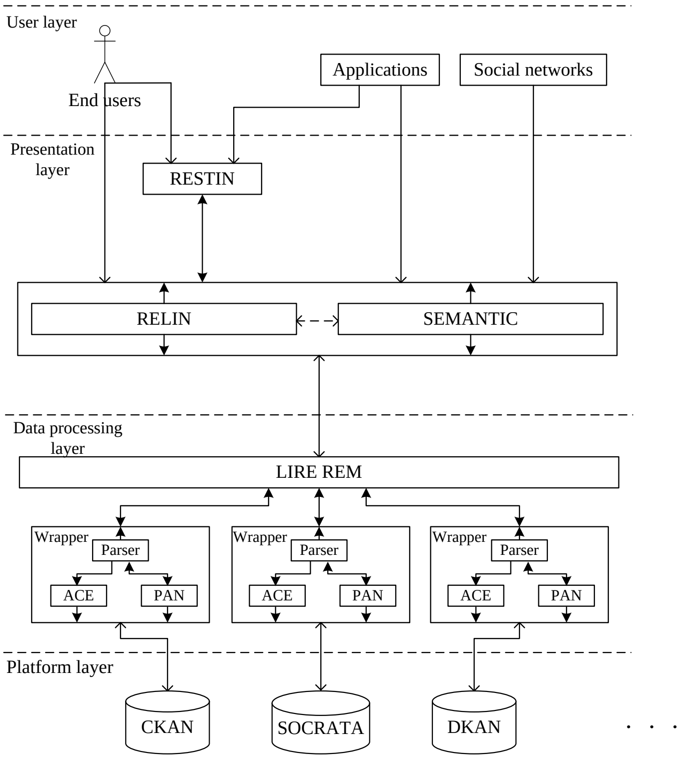


Fig. 1. Enhanced Linked Relations Architecture

[1] http://docs.ckan.org/en/latest/maintaining/linked-data-and-rdf
[2] https://docs.getdkan.com/en/latest/introduction/catalog-basics
[3] https://docs.opendatasoft.com/api/explore/v2.html
[4] https://dev.socrata.com/docs/formats/
[5] https://extensions.ckan.org/extension/lire/

The improvements are reflected in the possibility to apply LIRE on different OGD platforms and in the more precisely defined internal processes for linking datasets. We have grouped some components in order to enable the interoperability of LIRE with different OGD portals.

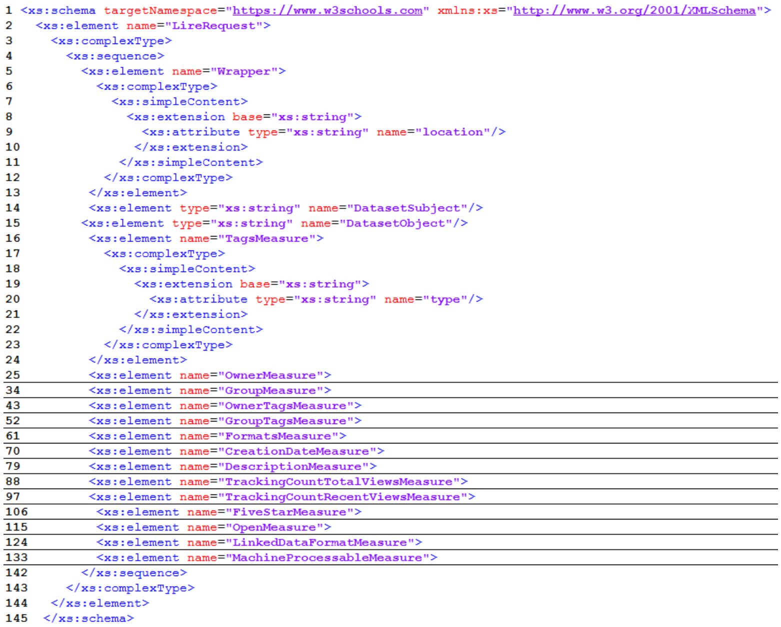
The enhanced LIRE architecture contains different components (modules) that deal with specific tasks in order to ensure data transformation, data interlinking and creation of dataset's relations:

- *WRAPPER* collects data from OGD portals. It contains three submodules:
  - *PARSER* deals with requests from REM component. It parses requests and issues relevant commands to PAN or ACE. It works vice versa as well by collecting responses from PAN or ACE and returning the appropriate responses to REM. The specification of information that PARSER sends/receives is given in Section 3.1.
  - *PAN – Portal Analyzer* is a submodule that creates requests to OGD portals to retrieve datasets. Then it checks specific dataset's metadata fields, according to specification received from the PARSER and returns gathered values.
  - *ACE – Action Executor* module executes actions for creating, updating or deleting dataset relations according to PARSER specifications.
- *REM (Relations Manager)* is responsible for determining the type of relations between datasets and creating and managing dataset's relations. REM examines dataset's metadata received through the PARSER and checks for similarity between datasets for the possible creation of relation. Within this architecture component, a model for relation of OGD is implemented. Given that the proposed model for relation OGD uses the types of relationships that are mutually related, this architecture component does not impose restrictions on the need to use these types of relationships, but leaves the final choice of the type of relationship to the user. The user can choose another type of relationship based on the experience gained in working with OGD on OGD platforms. REM can receive requests and send responses to the components contained in the upper and lower layers of the architecture. REM needs to identify WRAPPER in order to know where to send the applicable requirements and this is done by registering WRAPPER within REM. Moreover, REM creates the specification of request which is send to the WRAPPER, in order to execute proper actions on OGD platform.
- *RELIN (Relations Information)* prepares and creates necessary data for the visual preview of datasets. RELIN incorporates specific libraries for visualizing dataset relations and their graphical management. Every dataset is represented with a graphical element that contains information from dataset metadata, for example the description, tag, formats and others. Also, RELIN triggers REM by forwarding him a request for dataset's metadata examination in order to obtain the possible type of relation between datasets.
- *SEMANTICS (Semantics of Relations)* component performs dataset interlinking by updating the semantic version of OGD datasets i.e. adding links about relations between datasets. Dataset interlinking is modelled by using an RDF graph, described in detail in Section 3.4. The implemented RDF graph model is based on VoID vocabulary, where we exploit linkset feature, because it is naturally intended for modelling of links between datasets.

- *RESTIN (Relations Statistics and Indicators)* offers statistic about linked datasets, such as user created linksets per total linksets on OGD portal, government created linksets per total linksets on OGD portal, linksets per OGD portal datasets, where they all together constitute LINDAT, linked datasets indicator as the indicator of the status of dataset interlinking on a government portal.

3.1. Internal metadata specification

Information provision between REM and WRAPPER occurs by exchanging XML documents. REM issues request to the WRAPPER by sending the XML specification of the information it needs. XML schema specification is given in Fig.2.

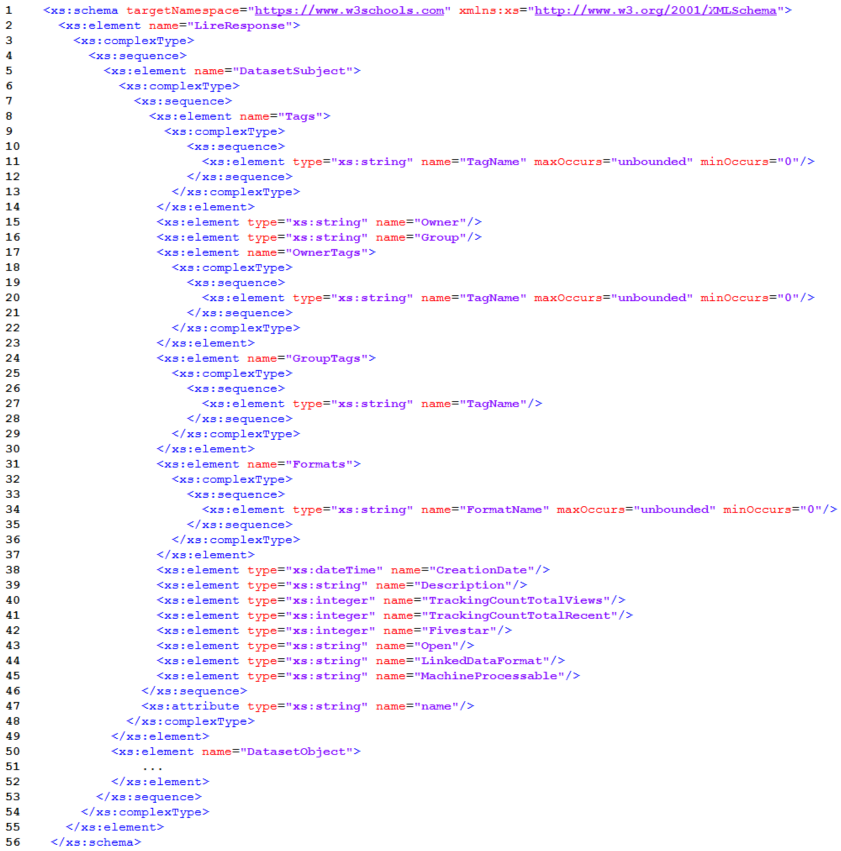


```
1 <xs:schema targetNamespace="https://www.w3schools.com" xmlns:xs="http://www.w3.org/2001/XMLSchema">
2   <xs:element name="LireRequest">
3     <xs:complexType>
4       <xs:sequence>
5         <xs:element name="Wrapper">
6           <xs:complexType>
7             <xs:simpleContent>
8               <xs:extension base="xs:string">
9                 <xs:attribute type="xs:string" name="location"/>
10               </xs:extension>
11             </xs:simpleContent>
12           </xs:complexType>
13          </xs:element>
14          <xs:element type="xs:string" name="DatasetSubject"/>
15         <xs:element type="xs:string" name="DatasetObject"/>
16          <xs:element name="TagsMeasure">
17            <xs:complexType>
18              <xs:simpleContent>
19                <xs:extension base="xs:string">
20                  <xs:attribute type="xs:string" name="type"/>
21                </xs:extension>
22              </xs:simpleContent>
23            </xs:complexType>
24          </xs:element>
25          <xs:element name="OwnerMeasure">
34          <xs:element name="GroupMeasure">
43          <xs:element name="OwnerTagsMeasure">
52          <xs:element name="GroupTagsMeasure">
61          <xs:element name="FormatsMeasure">
70          <xs:element name="CreationDateMeasure">
79          <xs:element name="DescriptionMeasure">
88          <xs:element name="TrackingCountTotalViewsMeasure">
97          <xs:element name="TrackingCountRecentViewsMeasure">
106          <xs:element name="FiveStarMeasure">
115          <xs:element name="OpenMeasure">
124          <xs:element name="LinkedDataFormatMeasure">
133          <xs:element name="MachineProcessableMeasure">
142        </xs:sequence>
143      </xs:complexType>
144    </xs:element>
145   </xs:schema>
```

Fig. 2. XML schema for REM’s request to WRAPPER

On line 5, REM specifies which WRAPPER will be used, since there can be multiple WRAPPER components for different OGD portals. Further on, lines 14 and 15 REM specify two datasets, the so called subject and object of relation. Lines 16 – 142 specify different schema elements suffixed with *Measure* that will be used for creating measure for linking datasets.

Depicted in Fig. 3., we can see the XML schema for the WRAPPER’s response. Both *DatasetSubject* and *DatasetObject* elemets are complex and contain a set of child elements. Lines 7 – 49 contain neccessary child elements for DatasetSubject that correspond to requested Measure elements. Same elements must be present for DatasetObject as well (lines from 50 upwards).



```
<xs:schema targetNamespace="https://www.w3schools.com" xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="LireResponse">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="DatasetSubject">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="Tags">
                <xs:complexType>
                   <xs:sequence>
                     <xs:element type="xs:string" name="TagName" maxOccurs="unbounded" minOccurs="0"/>
                   </xs:sequence>
                 </xs:complexType>
               </xs:element>
               <xs:element type="xs:string" name="Owner"/>
               <xs:element type="xs:string" name="Group"/>
               <xs:element name="OwnerTags">
                 <xs:complexType>
                   <xs:sequence>
                     <xs:element type="xs:string" name="TagName" maxOccurs="unbounded" minOccurs="0"/>
                   </xs:sequence>
                 </xs:complexType>
               </xs:element>
               <xs:element name="GroupTags">
                 <xs:complexType>
                   <xs:sequence>
                     <xs:element type="xs:string" name="TagName"/>
                   </xs:sequence>
                 </xs:complexType>
               </xs:element>
               <xs:element name="Formats">
                 <xs:complexType>
                   <xs:sequence>
                     <xs:element type="xs:string" name="FormatName" maxOccurs="unbounded" minOccurs="0"/>
                   </xs:sequence>
                 </xs:complexType>
               </xs:element>
               <xs:element type="xs:dateTime" name="CreationDate"/>
               <xs:element type="xs:string" name="Description"/>
               <xs:element type="xs:integer" name="TrackingCountTotalViews"/>
               <xs:element type="xs:integer" name="TrackingCountTotalRecent"/>
               <xs:element type="xs:integer" name="Fivestar"/>
               <xs:element type="xs:string" name="Open"/>
               <xs:element type="xs:string" name="LinkedDataFormat"/>
               <xs:element type="xs:string" name="MachineProcessable"/>
             </xs:sequence>
             <xs:attribute type="xs:string" name="name"/>
           </xs:complexType>
         </xs:element>
         <xs:element name="DatasetObject">
             ...
         </xs:element>
       </xs:sequence>
     </xs:complexType>
   </xs:element>
 </xs:schema>
```

Fig. 3. XML schema for WRAPPER’s response to REM

3.2. Workflow: creating linksets

In Fig. 4 we illustrate the basic workflow in the proposed architecture for the case when a user wants to relate two datasets.

When users access an OGD portal via browsers, they make exploration and preview of available datasets and related information depending on their need. During that process, they may find out that some datasets can be related. LIRE architecture is designed to suggest a type of relation to a user based on the examination of dataset’s metadata. This examination is taking place when a user tries to relate two datasets. For that purpose, RELIN triggers REM by forwarding a request for dataset’s metadata examination. REM then creates the specification of necessary data, where it includes dataset names which will be examined and forwards the information to WRAPPER with a request to gather proper information from the OGD portal. In a few iterations which WRAPPER performs with the OGD portal, a set of information is obtained, which is returned to REM in an appropriate form. When REM receives the requested information, it performs the analysis of the obtained information, as per model explained in [32], in order to make a suggestion for a possible type of relation for those datasets to a user. The user is given a choice, to apply the recommendation or chose another more appropriate type of relation. As the last iteration in the process, user commits the creation of relation between selected datasets.

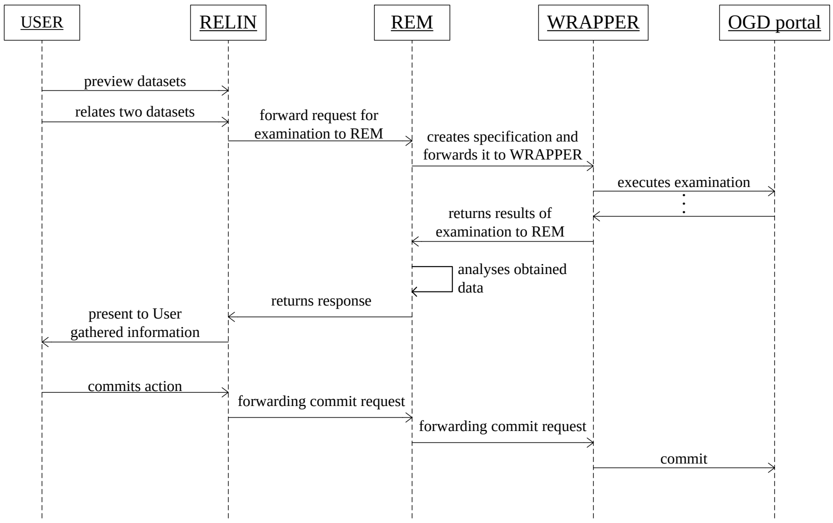


Fig. 4. Workflow for relating two datasets

When users access an OGD portal via browsers, they make exploration and preview of available datasets and related information depending on their need. During that process, they may find out that some datasets can be related. LIRE architecture is designed to suggest a type of relation to a user based on the examination of dataset's metadata. This examination is taking place when a user tries to relate two datasets. For that purpose, RELIN triggers REM by forwarding a request for dataset's metadata examination. REM then creates the specification of necessary data, where it includes dataset names which will be examined and forwards the information to WRAPPER with a request to gather proper information from the OGD portal. In a few iterations which WRAPPER performs with the OGD portal, a set of information is obtained, which is returned to REM in an appropriate form. When REM receives the requested information, it performs the analysis of the obtained information, as per model explained in [32], in order to make a suggestion for a possible type of relation for those datasets to a user. The user is given a choice, to apply the recommendation or chose another more appropriate type of relation. As the last iteration in the process, user commits the creation of relation between selected datasets.

The relation itself is added to the dataset's metadata. Relation metadata keys can be one of the following:

- subject – the first dataset in relation
- object – the second dataset in relation
- type – the type of relation between two datasets
- label – a string that indicates whether the linkset was created by government or user

Relation metadata is nested in dataset's metadata structure under a "relationships" key. This key is initially enabled in CKAN while other OGD platforms require an extension or modification of their existing metadata models to support this (DKAN, Socrata, Opendatasoft) [21].

3.3. The semantics of related dataset

Accessing related OGD datasets by linked data applications goes toward extracting context and meaning of OGD, thus enabling their proper use for the retrieval of information in conjunction with other linked datasets and the Web of Data space. For that purpose, relations between OGD datasets need to be modelled semantically. That is achieved through the RDF description of dataset relations. As already most of OGD portals ensure semantic availability of their datasets [21], it will be only necessary to model dataset relations semantically.

Within the SEMANTICS component of enhanced LIRE architecture, we carry out the semantic modelling of dataset relations by using VoID vocabulary. We exploit the feature "linkset", which is a collection of RDF links, where an RDF triple has a subject and object described in different dataset [26].

In this way we are enabling access to linksets from semantic web applications, but also from OGD portals. By exploring linksets, users can find related data and see more on the topic they searched for.

3.4. Architecture applicability

Although it requires the customization of the WRAPPER component, the main strength of LIRE architecture lies in its application to different OGD portals. OGD portals may be developed in different technologies, and it's up to the portal manager/developer to setup the WRAPPER to communicate with the portal. This is a compromise they need to make for applying the LIRE. This problem can also be solved by encouraging the community of developers to create a custom WRAPPER for the OGD portal, and later on share their effort with others. In this way we will stop accumulating time and costs for this task.

As the application of our architecture requires availability of metadata fields for the examination of datasets and storing dataset relations, OGD portals need to make them available. We can see from Table 1, which gives metadata support on some OGD platforms, that only CKAN supports relations between datasets, while in the SOCRATA some minor modifications are needed. Other platforms don't support relations, because they offer metadata that describe resource data values, but not the metadata about relations. The incorporation of the metadata that describe relations into the dataset structure will enable the usage of our architecture and contribute to the better definition of datasets.

Table 1. Metadata Support in Open Data Platforms

Platform	Dataset Metadata	Resource Metadata	Custom Metadata	Storing Relations
CKAN	YES	YES	YES	YES
DKAN	YES	YES	YES	NO
SOCRATA	YES	YES	YES	NO
OGDI	NO	YES	NO	NO
OGPL	NO	YES	NO	NO
OPENDATASOFT	YES	YES	YES	NO
PLENARIO	NO	YES	NO	NO
JUNAR	NO	YES	NO	NO

The presence of relevant dataset information in the metadata structure of dataset is given in Table 2. We performed an analysis of 8 popular OGD portals powered by four different platforms: CKAN, DKAN, Socrata and Opendatasoft. This analysis aims to validate the model for proposing relations by examining whether the information of importance is presented in the dataset’s metadata structure. Defined conditions, named C1-C13 [32], examine the following:

- **C1** – Number of same/similar tags between two datasets
- **C2** - Do they belong to the same organization
- **C3** - Do they belong to the same group
- **C4** - Whether the number of the same/similar tags of the first dataset is greater than (or less than) the number of the same/similar tags in the second dataset organization
- **C5** - Whether the number of the same/similar tags of the first dataset is greater than (or less than) the number of the same/similar tags in the second dataset group
- **C6** - Are they linked via links in extra field
- **C7** - Whether the number of the same/similar resource formats of the first dataset is greater than (or less than) the number of the same/similar resource formats in the second dataset
- **C8** - Whether the first dataset was created after the second
- **C9** - Whether the descriptions of two datasets are similar
- **C10** - Whether the number of total views of the first dataset is less than (or greater than) the number of total views of the second dataset
- **C11** - Whether the number of recent views of the first dataset is less than (or greater than) the number of recent views of the second dataset
- **C12** - Whether the five star index of the first dataset is less than (or greater than) the five star index of the second dataset
- **C13** - Whether they are open

Table 2 shows the percentage of defined data for the whole portal for each condition. It is important to notice that C6 has no values. This indicates that C6 is irrelevant for the determination of the type of relation and we have excluded this condition in the revised model. The remaining of the table, shows the presence of mostly all data. It is interesting to note that portals powered by CKAN have a lower presence of some data relative to other portals.

Table 2. Presence of Relevant Dataset Information in Metadata Structure

Portal	Platform	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
catalog.data.gov	CKAN	84	100	0	100	100	0	88	100	100	100	100	88	53
data.gov.uk	CKAN	28	100	0	100	100	0	9	100	100	0	100	9	9
offenedaten-koeln.de	DKAN	99	0	97	0	99	0	99	99	99	0	0	99	99
www3.unog.ch	DKAN	72	0	0	0	90	0	90	100	95	0	0	90	49
data.edmonton.ca	Socrata	97	100	100	100	0	0	100	99	100	100	100	100	100
data.oregon.gov	Socrata	73	100	98	100	0	0	95	100	79	100	100	95	100
opendurham.nc.gov	Opendatasoft	100	91	0	100	0	0	100	100	75	0	0	100	0
opendata.paris.fr	Opendatasoft	99	93	0	100	0	0	99	99	99	0	0	99	0

To improve dataset relations in OGD portals, we can in the future rely more on data consumers. Whoever browses datasets can be the one creating links between the datasets. We should not wait for the OGD portal's management team to link datasets [6]. This is aligned with civic-sourcing, a particular type of "crowd-sourcing" being adopted as a part of Government 2.0 to harness the wisdom of citizens [33]. As a full automatic approach does not always guarantee good results, the inclusion of consumers improves obtaining linked datasets, which imposes our solution as swift, less demanding and approachable by non-expert users.

4. LINDAT indicator

We introduced an indicator named LINDAT (LINkedDATaset), calculated by RESTIN component of the architecture, which indicates how much linked data is created vs. total possible linked data on the portal. LINDAT is calculated by dividing two measures: Created Linksets (CL) and Total Possible Linksets (TPL), as given in equation (1), and it is expressed in percentage. CL represents a sum of all created linksets on the OGD portal, both by OGD data publishers and data consumers. Before making the data available to stakeholders, government representatives perform internal data linking (GCL – Government Created Linksets). The process probably happens at data import, but can also be done at later stage if required. Data consumers can participate in the linking of open data (UCL – User Created Linksets), because their experience in using open government data is significant and should be taken into account.

$$\text{LINDAT} = \text{CL} / \text{TPL}. \tag{1}$$

Total Possible Linksets is calculated as given in the equation (2), where "n" is the number of datasets that reside on the OGD portal.

$$\text{TPL} = n * (n - 1). \tag{2}$$

LINDAT calculation excludes the possibility of an erroneous data, in a sense that some member of linkset can have missing links. This can happen while adding relations manually, but by using LIRE architecture it is impossible to make this mistake since relations are created automatically. Knowing that linksets are expressed through directed multigraphs, they are completely equivalent to their inverse ones. For example, a linkset with the triple: A skos:broader B is semantically equivalent to a linkset with the triple: B skos:narrower A [34]. From here it becomes clear that each direction of relation between datasets has a linkset with its own inverse linkset.

4.1. Datasets evaluation using LINDAT

Evaluation of datasets on OGD portals using LINDAT, started with accessing the list of CKAN powered OGD portals around the world[6]. We used custom generated application to search all listed portals and find whether *relationships* metadata key is defined, which would reveal linked relations between datasets. The results of the evaluation are listed in Table 3. Unfortunately only 4 out of 197 portals have used relationships metadata key to link datasets. This extremely low number tells us that not many portals are linking datasets.

Table 3. LINDAT indicator calculation with CKAN powered OGD portals

Portal	Number of datasets	TPL	CL	LINDAT
datahub.io	11462	131365982	340	0.000259%
dados.gov.br	6458	41699306	9	0.000022%
data.gov.ie	8823	77836506	9	0.000012%
dartportal.leeds.ac.uk	25	600	2	0.333333%
data.gov.uk	47139	2222038182	0	0.000000%

Table 3 shows weak interlinking between datasets on every portal, because there is a low number of created linksets. Portal datahub.io has the highest number of defined relationships but very low LINDAT index, because the number of total possible linksets is very high. Portal dartportal.leeds.ac.uk has the best score for LINDAT index, because of the small number of datasets available on the portal. Each of the analyzed OGD portals shows diversity in the number of created relationships with comparing of number of datasets that reside on these portals. The relationships feature is not fully utilized and there should be more attention paid to this issue. There are several approaches [3, 27, 29, 30] that can help to achieve better interlinking between datasets, and lead to higher number of linked datasets on the portal. Using enhanced IRE architecture and applicable prototype application, this number could increase even more since it includes data consumers in datasets interlinking. Their feedback can be used to assess candidate datasets for interlinking and to offer the best suited workflow and best practice to support achieving this aim in collaborative and participatory manner.

[6] https://ckan.org/about/instances

5. Discussion and future remarks

The enhanced LIRE architecture is created for the purpose of managing and creating relations between datasets on OGD portals. It utilizes dataset's metadata to propose relations and to interlink datasets using custom refined model. Model uses dataset's metadata keys in a way it has not been used before, to semantically link datasets and enable their availability in linked data domain. This paper makes a step forward in this direction and utilizes valuable information hidden in metadata with the aim to expose OGD for processing in semantic applications.

Involving data consumers in the process of datasets interlinking helps to increase the number of linksets and contributes to the better accessibility of government data on the portal itself but also in various software applications via publicly available APIs. This is in line with the raised awareness of the user involvement in the process of creation of linked data. In LIRE, users are allowed to identify and reveal possible relations between datasets and to validate them through the proposed model. We like to believe that in this way our architecture contributes to better participation and collaboration between data consumers and data holders (i.e. government). To track the status of dataset's interlinking we have exposed a measure called LINDAT. LINDAT gives the information on how much linksets are added compared to the total number of possible linksets. The total number of possible linksets is directly affected by the total number of datasets on the portal. The value of this indicator is available at any time as it is calculated automatically.

The proposed architecture is applicable on different OGD platforms. Interoperability is ensured by creating custom WRAPPER for each new platform. Connectivity between WRAPPER and the rest of the architecture is achieved via the specification of necessary actions that are issued from components located above the WRAPPER, enabling in that manner the interoperability of the whole architecture. The question that arises is: Should there be a framework for the development of WRAPPER component? This is something that we plan to address in the future, in order to achieve better integrity of the architecture. The future improvements of the architecture could also go in the direction of the examination of the semantic similarity of datasets, based on their names and description. As there are many techniques and tools available for semantic similarity, it will be first necessary to evaluate the applicability of those techniques in short and long texts which exists in dataset metadata description. The research in this area can potentially improve the operability of LIRE application.

While doing the research on dataset's metadata we stumbled upon a question: Should there be a standard for permissible names of metadata's tags, for achieving their unique interpretation? Different OGD platforms differently name the metadata tags [35, 36], giving a task to the developers to develop a technique for their processing based on their notation. The existence of a standard for naming and structure of dataset's metadata will contribute to better processing of this information and speed up the development of end-users applications. A solution to this problem can be found in the development of a mediator, which will process uniquely dataset's metadata from different OGD platforms. Nevertheless, it must be kept in mind that there is no guarantee that metadata will be available on each portal.

References

1. Jacksi, K., Zeebaree, S. R. M., Dimililer, M.: LOD Explor-er: Presenting the Web of Data. International Journal of Advanced Computer Science and Applications, Vol. 9, No. 1, pp. 45 – 51. (2018)
2. Veljković, N., Bogdanović-Dinić, S., Stoimenov, L.: Benchmarking open government: An open data perspective. Government Information Quarterly, Vol. 31, No. 2, pp. 278-290. (2014)
3. Ding, L., Lebo, T., Erickson, J. S., DiFranzo, D., Williams, G. T., Li, X., Michaelis, J., Graves, A., Zheng, J. G., Shangguan, Z., Flores, J., McGuinness, D. L., Hendler, J. A.: TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 9, No. 3, pp. 325-333. (2010)
4. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked Open Government Data Analytics. In Proceedings of the International Conference on Electronic Government. Koblenz, Germany, pp. 99-110, 2013.
5. Sheridan, J., Tennison, J.: Linking UK Government Data. In Proceedings of the Workshop on Linked Data on the Web. Raleigh, North Carolina, USA. (2010)
6. Maali, F., Cyganiak, R., Peristeras, V.: A Publishing Pipeline for Linked Government. In Proceedings of the 9th Semantic Web Conference. Heraklion, Crete, Greece, 778-792. (2012)
7. Assaf, A., Senart, A., Troncy, R.: Roomba: Automatic Validation, Correction and Generation of Dataset Metadata. In Proceedings of the 24th International Conference on World Wide Web Companion. Sophia Antipolis, France, 159 – 162. (2015)
8. Kalampokis, E., Tambouris, E., Tarabanis, K.: A classification scheme for open government data: Towards linking decentralised data. International Journal of Web Engineering and Technology, Vol. 6, No. 3, pp. 266-285. (2010)
9. Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., Schraefel, M. C.: Linked open government data: Lessons from data.gov.uk. IEEE Intelligent Systems, Vol. 27, No. 3, pp. 16-24. (2012)
10. Janssen, M., Estevez, E., Janowski, T.: Interoperability in Big, Open, and Linked Data Organizational Maturity, Capabilities, and Data Portfolios. Computer, Vol. 47, No. 10, pp. 44-49. (2014)
11. Assaf, A., Troncy, R., Senart, A.: HDL - Towards a Harmonized Dataset Model. In Proceedings of the 2nd International Workshop on Dataset Profiling & Federated Search for Linked Data. Sophia Antipolis, France, 159 - 162. (2015)
12. Kashyap, V., Sheth, A.: Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. In: M. Papazoglou and G. Schlageter (eds.), Cooperative Information Systems: Tends and Directions 1998, Academic Press: London, UK. pp. 139-178. (1996)
13. Schmachtenberg, M., Christian, B., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In Proceedings of the 13th International Semantic Web Conference. Riva del Garda, Italy, 245-260. (2014)
14. Lebo, T., Erickson, J. S., Ding, L., Graves, A., Williams, G. T., DiFranzo, D., Li, X., Michaelis, J., Zheng, Z., Flores, J., Shangguan, J. G., McGuinness, D. L., Hendler, J. A.: Producing and Using Linked Open Government Data in the TWC LOGD Portal. In Proceedings of the Linking Government Data Conference, 51-72. (2011)16. Scharffe, F., Fan, Z., Ferrara, A., Khrouf, H., Nikolov, A.: Methods for automated dataset interlinking. Datalift Deliverable D4.1, pp. 34 - 73. (2013)
15. Cverderlj-Fogaraši, I., Sladić, G., Gostojić, S., Segedinac, M., Milosavljević, B.: Semantic integration of enterprise information systems using meta-metadata ontology. Information Systems and e-Business Management, Vol 15, No. 2, pp. 257 - 304. (2010)
16. Scharffe, F., Fan, Z., Ferrara, A., Khrouf, H., Nikolov, A.: Methods for automated dataset interlinking. Datalift Deliverable D4.1, pp. 34 - 73. (2013)

17. Ellefi, M. B., Bellahsene, Z., Todorov, K., Dietze, S.: Dataset Recommendation for Data Linking: An Intensional Approach. In Proceedings of the 13th European Semantic Web Conference. Heraklion, Crete, Greece, 36-51. (2016)
18. Ngomo, A. C. N., Sherif, M. A.: LIMES - A Framework for Link Discovery on the Semantic Web. Journal of Web Semantics, (2018)
19. Kepeklian, G., Bihanic, L., Troncy, R.: Datalift: A platform for integrating big and linked data. In Proceedings of the International Conference on Big Data from Space, Frascati, Italy, 370-373. (2014)
20. Milic, P., Veljkovic, N., Stoimenov, L.: Comparative analysis of metadata models on e-government open data platforms. IEEE Transactions on Emerging Topics in Computing, (2018)
21. Zuiderwijk, A., Jeffery, K., Janssen, M.: The potential of metadata for linked open data and its value for users and publishers. JeDEM-e-Journal of e-Democracy and Open Government, Vol. 4, No. 2, pp. 222-244. (2012)
22. Maali, F., Cyganiak, R., Peristeras, V.: Enabling Interoperability of Government Data Catalogues. In Proceedings of the 9th International Conference on Electronic Government. Lausanne, Switzerland, 339-350. (2010)
23. DCMI Usage Board. Dublin Core Metadata Initiative: DCMI Metadata Term, http://dublincore.org/documents/dcmi-terms/. (2018)
24. EC JoinUP, DCAT Application Profile, https://joinup.ec.europa.eu/asset/dcat_application_profile/home. (2018)
25. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with voID Vocabulary," http://www.w3.org/TR/void. (2018)
26. VOID: Vocabulary of Interlinked Datasets, Available: https://www.w3.org/TR/void/, (2018)
27. Fiorelli, M., Stellato, A., McCrae, J. P., Cimiano, P., Pazienza, M. T.: LIME: The Metadata Module for OntoLex. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds.): The Semantic Web, Latest Advances and New Domains, Vol. 9088, 321-336. (2015)
28. Cyganiak, R., Maali, F., Peristeras, V.: Self-Service Linked Government Data with dcat and Gridworks. In Proceedings of the 6th International Conference on Semantic Systems. Graz, Austria, 37:1-37:3. (2010)
29. Oliveira, H. R., Tavares, A. T., Lóscio, B. F.: Feedback-based data set recommendation for building linked data applications. In Proceedings of the 8th International Conference on Semantic Systems. Graz, Austria, 49-55. (2012)
30. De Vocht, L., Dimou, A., Breuer, J., Van Compernolle, M., Verborgh, R., Mannens, E., Mechant, P., Van De Walle, R.: A Visual Exploration Workflow as Enabler for the Exploitation of Linked Open Data. In Proceedings of the 3rd International Conference on Intelligent Exploration of Semantic Data. Riva del Garda, Italy, Vol. 1297, 30-41. (2014)
31. Leme, L. A. P. P., Lopes, G. R., Nunes, B. P., Casanova, M. A., Dietze, S.: Identifying candidate datasets for data interlinking. Lecture Notes in Computer Science, Vol. 7977, 254-266. (2013)
32. Milić, P., Veljković, N., Stoimenov, L.: Linked Relations Architecture for Production and Consumption of Linksets in Open Government Data. In: Janssen, M., Mäntymäki, M., Hidders, J., Klievink, B., Lamersdorf, W. (eds.). Open and Big Data Management and Innovation, Vol. 9373, 221-222, (2015)
33. Nam, T.: The Wisdom of Crowds in Government 2.0: Information Paradigm Evolution toward Wiki-Government. In Proceedings of the 16th Americas Conference on Information Systems. Lima, Peru, 337-348, (2010)34. SKOS: Simple Knowledge Organization System, Available: https://www.w3.org/TR/skos-reference/, (2018)
35. CKAN. CKAN API. http://docs.ckan.org/en/latest/api. (2018)
36. SOCRATA. Socrata API. http://dev.socrata.com/docs/endpoints.html. (2018)

Nataša Veljković received the BSc, MSc and PhD degrees in computer science at the University of Niš, Serbia. She is currently working as a Teaching Assistant at Faculty of Electronic Engineering with the Department of Computer Science. Her area of research includes e-government, e-systems, open data, IoT.

Petar Milić received the BSc and MSc degrees in computer science at the University of Priština temporary settled in Kosovska Mitrovica, Serbia and PhD degrees in computer science at the University of Niš, Serbia. He is currently working as a Teaching Assistant at the University of Priština temporary settled in Kosovska Mitrovica, Serbia. His area of research includes e-government, e-systems and web.

Leonid Stoimenov received the BSc, MSc and PhD degrees in computer science at the University of Niš, Serbia. He is a Professor at Faculty of Electronic Engineering at this University. His research interests in computer science include e-systems, GIS, databases, ontologies and semantic interoperability. He is a member of IEEE, IAENG and representative in AGILE association of GIS laboratories in Europe.

Kristijan Kuk received the PhD degrees in computer science at the University of Niš, Serbia. He is a Professor at the University of Criminalistic and Police Studies in Belgrade, Serbia. His research interests in computer science include machine learning, artificial intelligence, pedagogical agent, natural language processing, IT security.

Received: April 20, 2019; Accepted: December 20, 2019