Linked Relations Architecture for Production and Consumption of Linksets in Open Government Data

Petar Milić(✉), Nataša Veljković, and Leonid Stoimenov

Faculty of Electronic Engineering, University of Niš,
Aleksandra Medvedeva 14, Niš, Serbia
milicpetar86@gmail.com,
{natasa.veljkovic,leonid.stoimenov}@elfak.ni.ac.rs

Abstract. Linking open data in government domain, can lead to creation of new services and information as well as discovery of new ways to perform queries and get results in accessible, machine processable and structured manner. To reach the full potential of open government data more relations between data should be discovered. The interconnection of open government data and semantic description of their relations can bring new aspect of producing and consuming the data. In this paper we investigate issues for producing and utilizing open government data with special focus on dataset relations. We have proposed the Linked Relations (LIRE) architecture for relations creation between datasets and a basic RDF model of relation between two datasets. The architecture contains different modules that perform analysis of datasets attributes and suggest the type of relation between the datasets. It can be utilized by open data portals for creating relations between datasets belonging to different public agencies and government sectors. An idea presented in this paper is made available as CKAN plugin.

Keywords: Open government data · Linked data · Linkset · Dataset relations · CKAN open data portal

1 Introduction

Publication of open government data (OGD) leads to more openness, transparency and efficiency of public administration. It also brings benefits for citizens by influencing development of government services for society, hence producing better public service outcomes. Open data philosophy suggests that data should be published in open formats and in ways that make them accessible, readily, reusable and available to the public, business and government sector [1]. Following this approach, data can be easily consumed by both web developers and common users. Having in mind that most of the published data comes in original (raw) format, beneficiary contribution is not negligible, we can even say that it is immense. Every reuse of data adds new value and creates new knowledge enabling data lifecycle to expand and evolve.

Web of Data represents decentralized and heterogeneous sources of information interlinked through typed links. This is achieved by publishing structured data in

M. Janssen et al. (Eds.): I3E 2015, LNCS 9373, pp. 212–222, 2015.
DOI: 10.1007/978-3-319-25013-7_17

Resource Description Format (RDF) using URIs. RDF is the format on which is based Semantic Web, as its use of URIs allows data to be identified by reference and linked with other relevant data by subject, predicate or object. Making OGD available in the Web of data, makes them publicly available, and practically expand Web of data space, allowing their discovery and usage. Kalampokis et al. [2] claim that real value of OGD is revealed with linking data which provides unexpected and unexplored insights into different domains and problem areas. Linked Government Data (LGD) are actually OGD that runs on Semantic Web's kerosene – metadata. Metadata provide documentation, context and necessary background information.

According to Sheridan and Tennison [3] the Semantic Web standards are mature and powerful, but there is still a lack of practical approaches and patterns for publishing OGD. Nevertheless, over the last few years there is increasing number of governments that are publishing their OGD data as linked data. The adoption of the LGD has led to the extension of the government open data space, connecting data from diverse domains such as economy, finance, medicine, statistics and others to enable new types of applications. With LGD paradigm users can browse one data source and then navigate along links to related data sources. Most promising implementation of LGD is based on Semantic Web philosophy and technologies. Tim Berners-Lee [4] noted this and he gave instructions how to link data and to include government data into the Web of Data.

By following the Open Government movement [5] in the world, many governments have published their open data through the open data platforms [6]. One of the most utilized open-source solutions for publishing open datasets is the CKAN's (Comprehensive Knowledge Archive Network) open data platform [7]. This platform enables both back-end and front-end interface, used respectively for publishing/modifying and searching/reviewing open datasets. What this platform doesn't offer is the possibility to link datasets between each-other, creation of meaningful relations between them and publishing this data as linked data. This tackled our minds and we wanted to generalize this problem and create a common architecture that could be applied to other open data platforms as well and to suggest possible relations between datasets and enable their linking.

In this paper we explore area of dataset relations for producing and utilizing linked government data. We describe architecture that determines type of relation between datasets, in order to take advantage of the relations for dataset linking. The architecture contributes to defining a model of semantic representation of dataset relations and their automatic production. This leads towards production of quality LGD in line with their interlinking and integration. The remaining of the paper is organized as follows. In Sect. 2 we review related work. In Sect. 3 we propose an architecture for modelling and linking dataset relations. Section 4 presents the visual tool for creating relations between datasets, developed on the basis of architecture described in Sect. 3 that shows benefits of its use. Finally, in Sect. 5 conclusions are drawn along with the future work.

2 Related Work

A group of authors [8] proposed architecture for integrating Public Sector Information (PSI) catalogs via the activities and components essential for discovery, allowing the presentation of catalogs in standardized form, facilitating search and retrieval across

resources. This architecture requires downloading and transforming catalogs with retrievable records into a common schema language format along with addressing semantic heterogeneity with schema matching and statistical analysis of ontology structures. Pioneers in LGD, UK and USA have shown that creating high quality Linked Data from raw data files requires considerable investment into reverse-engineering, documenting data elements, data cleanup, schema mapping and instance matching [9].

If we want that government datasets become linked government datasets, semantics must be added to them. Appropriate rules tell us how to describe and how to establish links between them. For linking datasets, Alexander et al. [10] propose Vocabulary of Interlinked Datasets (VOID), a vocabulary that allows to formally describe linked RDF datasets. It defines classic LOD and 3rd-party case. In the LOD case the linkset is a subset of one of the two involved datasets, while in 3rd-party case a third dataset is involved that actually contains the linkset.

Interoperability between government datasets and bringing them closer to the Web of Data is also discussed in [11]. The authors designed DCAT vocabulary, based on exploration of seven existing open data portals to allow expression of datasets in the RDF data model. They conduct feasibility study to prove their claim that different catalogues can be rendered in proposed vocabulary.

Many authors in literature deals with proposing systems for production and consumption of LGD. Ding et al. [12] suggest the use of LGD ecosystem for LGD data production and consumption as a Linked Data – based system where users manage and consume open government data in connection with online tools, services and societies. It supports large-scale LGD production, promote LGD consumption and grow of the LGD community. TWC LGD ecosystem from [12] is based on converting raw OGD datasets into linked data and their integration with other resources. Kalampokis et al. [13] give classification scheme for OGD, where they showed technological and organizational approach for provision of linked data based on relevant literature. The proposed architecture links decentralized data with maintaining a list of available resources in the area and assigning a URI to each of them. This solution is intended to use in single government cases to link data in different datasets belonging to different public agencies and government sectors.

Schmachtenberg et al. [14] present an overview of the linkage relationships between datasets in the form of an updated LOD cloud diagram based on data that can actually be retrieved by a Linked Data crawler. They consider that two datasets can be linked if there exists at least one RDF link between resources belonging to the datasets. Using metadata with appropriate metadata architecture can bring benefits for LGD publication and use, along with improving the ability for finding and interpreting of LGD data, creating order within datasets, comparing, correct interpretation, accessibility, visualization and other benefits, as discussed in [15].

Janssen, Estevez and Janowski state in [16] that successfully linking of datasets requires understanding the data's context sensitive meaning. They claim that collected data from organizations which do not always anticipate its full potential use, might not sufficiently align with other datasets or possible relationships among datasets are unknown.

To the best of our knowledge, there are not work that deals with automatic production of linked datasets. Some of them [8, 9, 12] requires intervention of users which is a time-consuming process and does not specifies rules for their interlinking. Kalampokis [13] and Schmachtenberg [14] discuss on linking datasets based on their semantic description without going deeper in the OGD datasets. This work does not tackle informations hidden in published OGD datasets, which by our opinion can help in linking OGD datasets. Zuiderwijk et al. [15] claims that metadata can give potential in realization of all benefits from linked data that exists in literature. Similarly, Janssen et al. [16] points to the fact that the relations between datasets are unknown and that they can contribute to better linking of datasets.

Based on the approaches for linking OGD datasets mentioned in previous paragraphs, we got the idea to explore relations between datasets to check whether they can be used to produce linked datasets. According to that, we have developed an architecture for visual creation of relations between datasets, and their automatic linking. The so called LIRE (Linked Relations) architecture was created at first to enrich CKAN open data portal with a tool that enables automatic creation and deletion of linksets. In the following section we will explain more thoroughly the proposed architecture.

3 Linked Relations Architecture – LIRE

LIRE system architecture, outlined in Fig. 1, consists of different interconnected components, each with specially assigned tasks. Implemented functionalities are available through a single workbench. LIRE enables users to find, manage, integrate, publish and reuse relations between datasets. It promotes production and consumption of linksets, semantic data that describe relations between datasets.

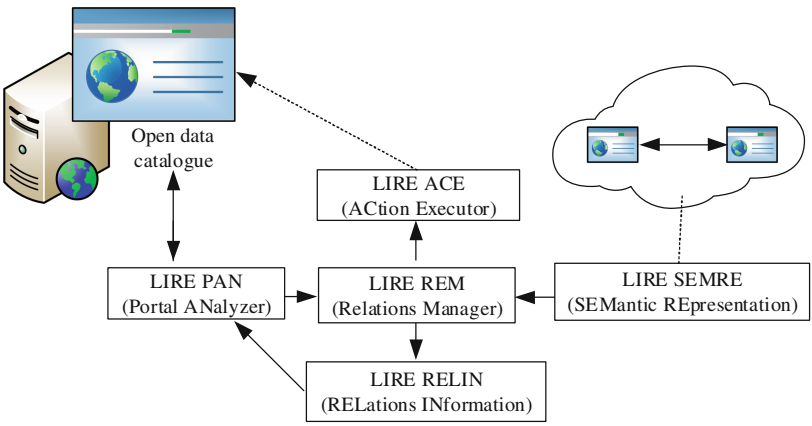

Fig. 1. LIRE system architecture

LIRE consists of the following modules:

- Portal Analyzer – PAN. This module is the entry point of LIRE. PAN enables filtering of datasets on different parameters (by tag, organization, group, all datasets or random datasets). It prepares datasets for processing for REM module and gives necessary information about datasets to RELIN module for their visual display.
- Relations Manager – REM. Creating, managing and determining type of relations between datasets based on their metadata are carried out in REM. This module implements a model on which is based determination of the type of relation, described in detail in the next section. REM examines datasets metadata to determine similarity between them for possible relation creation. Module does not limit the user in choosing the type of relations, so that the user does not have to apply the type of relation that is suggested by REM, but can apply selected, if it differs from the proposed one. It calls ACE module to execute pending actions to the OGD portal and RELIN module to refresh display after editing.
- Action Executor ACE – This module executes actions created by REM to store results of editing in the OGD platform. Supported actions are "CREATE" and "DELETE".
- Relations Information – RELIN is module used for visualization of relations. It enables short preview of information of datasets in REM module to enable user to decide whether to relate datasets or not, and user interface for managing datasets relations. Also incorporates jQuery and CSS libraries for visualizing datasets relations and their graphical management. Every dataset is represented with graphical element that contains information on dataset's description, tags, formats and existing relations.
- Semantic Representation – SEMRE module creates semantic representation of any existing relations. This semantics is created based on the model of RDF graph, described in detail in Sect. 3.2. Implemented RDF graph model is based on voID (vocabulary of Interlinked Datasets) vocabulary, because of its simplicity for describing linked datasets.

3.1 Creating Relations with LIRE

Relations Manager deals with managing of relations between datasets. It examines data that describe datasets to determine type of relation. After examination, determined type is suggested to user who can accept suggested solution or choose another one. Using developed relation suggestion models, described in Table 1, that are based on presence/absence of selected datasets metadata, REM module of LIRE architecture can determine whether datasets can have one of the following relations: parent_of, child_of, links_from, links_to. The child_of relation model consists of thirteen conditions, listed as C1-C13, where each condition examines certain dataset property or combination of properties on a true/false basis. If all conditions are met then relation between two datasets is of type child_of. The conditions for child_of relation can be used also for determining whether the relation between two dataset is of type parent_of, but with following modifications: conditions C4, C5, C7 and C8 should be less than, while C10-C12 should be greater than.

Table 1. Models for relations child_of and links_from

CHILD_OF		LINKS_FROM	
C1. Number of same/similar tags between two datasets	>0	C1. Number of same/similar tags between two datasets	>0
C2. Do they belong to the same organization	true	C2. Whether they are open	true
C3. Do they belong to the same group	true	C3. Whether the number of the same/similar resource formats of the first dataset is greater than the number of the same/similar resource formats in the second dataset	>
C4. Whether the number of the same/similar tags of the first dataset is greater than the number of the same/similar tags in the second dataset organization	>	C4. Whether the five star index of the both datasets is higher than 3	>3
C5. Whether the number of the same/similar tags of the first dataset is greater than the number of the same/similar tags in the second dataset group	>	C5. Whether they have at least one linked format in its resources	true
C6. Are they linked via links in extra field	true	C6. Whether they have at least one machine processable format	true
C7. Whether the number of the same/similar resource formats of the first dataset is greater than the number of the same/similar resource formats in the second dataset	>	C7. Whether the first dataset was created before the second	<
C8. Whether the first dataset was created after the second	>	C8. Whether the descriptions of two datasets are similar	>n
C9. Whether the descriptions of two datasets are similar	>n		
C10. Whether the number of total views of the first dataset is less than the number of total views of the second dataset	<		
C11. Whether the number of recent views of the first dataset is less than the number of recent views of the second dataset	<		
C12. Whether the five star index of the first dataset is less than the five star index of the second dataset	<		
C13. Whether they are open	true		

For links_from relation there are eight conditions, listed as C1-C8. If all conditions are met then relation between two datasets is of type links_from. These conditions can be used for the determining whether the relation between two datasets is of type links_to with following modifications: condition C3 should be less than and C7 greater than.

3.2 Creating Semantics of Relations with LIRE

Modelling relations between OGD datasets using linked data principles and techniques can add more semantics to government data, enabling thus easier search and retrieval of information by using semantic tools. Adding semantics to OGD is achieved through RDF description of datasets with help of Dublin Core and DCAT vocabularies [17, 18]. Dublin Core expresses metadata that describe dataset in RDF for direct machine processing through most well-known and basic terms, while DCAT facilitates interoperability and increases discoverability for easy consume of LGD.

LIRE architecture has SEMRE component which carries out modelling of relations with use of voID (vocabulary of Interlinked Datasets) vocabulary, because voID is one of the most widespread vocabularies for LGD and has a feature called linkset. A linkset is a collection of RDF links where an RDF triple has subject and object described in different datasets [19]. Vocabulary voID is convenient for use in our case because it is naturally intended for describing linked datasets. Knowing linkset structure, we can define a basic RDF model of relation between two OGD datasets implemented in SEMRE (Fig. 2).

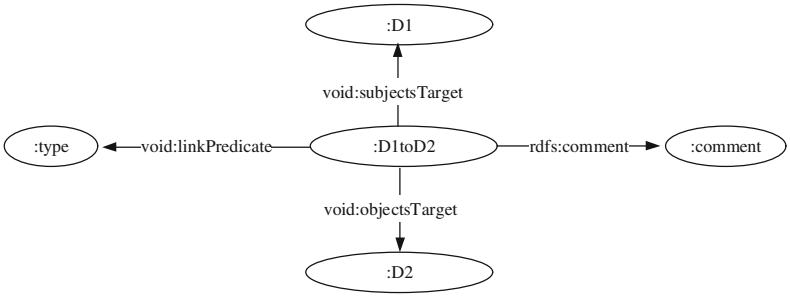

Fig. 2. RDF graph of relation between two datasets implemented in LIRE SEMRE

Implemented RDF graph model can be described using two semantic web data formats Turtle and RDF + XML. In Turtle, it would be:

```
@prefix void: <http://rdfs.org/ns/void#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
:D1toD2 a void:Linkset;
void:subjectsTarget :D1;
void:objectstarget :D2;
void:linkPredicate :type;
```

```
rdfs:comment :comment;
```

Represented with RDF+XML, it would look like:

```
<rdf:RDF>
<void:linkSet>
<void:subjectsTarget>D1</void:subjectsTarget>
<void:objectsTarget>D2</void:objectsTarget>
<void:linkPredicate>type</void:linkPredicate>
<rdfs:comment>comment</rdfs:comment>
</void:linkSet>
</rdf:RDF>
```

In Turtle code D1toD2 represents the name of linkset, i.e. the name of the relation, and it is identified by void:linkset statement. It is also a part of URI of appropriate RDF/XML syntax. Terms D1 and D2 are dataset's names represented by void:subjectsTarget and void:objectsTarget statements respectively. In both data formats type is the type of relation (parent_of, child_of, etc.) represented by void:linkPredicate. Relation description is contained in comment element identified by rdfs:comment.

The appropriate mapping between dataset relation elements and voID vocabulary is illustrated on Fig. 3.

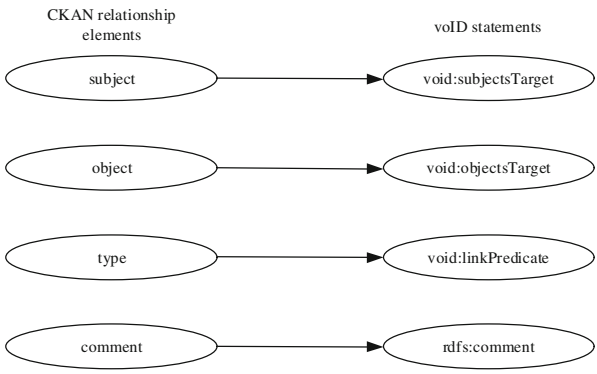

Fig. 3. Mapping between CKAN relations properties and voID statements conducted in SEMRE

SEMRE uses voID description of dataset relations with aim to offer users more data related with it and to enable easy access, search and retrieval of information. In this manner SEMRE enables access to LGD from semantic web applications. It also offers a mechanism to implement a semantic description of dataset relations into the open data catalog.

The RELIN component of architecture enables user interface for creating related datasets. Every dataset is represented with graphical element that contains information on dataset's description, tags, formats and existing relations.

4 LIRE Architecture Deployment for CKAN Data Catalogue

To demonstrate the value of LIRE architecture, we have deployed the architecture as plugin for CKAN open data portal. The plugin is in beta phase now, but it will be soon available from the CKAN's online plugin repository. To present the use case of using LIRE as plugin we have installed CKAN platform on local computer, and uploaded to it few datasets from datahub.io. To see existing relations between datasets, user filters the display by using one of the following parameters: datasets per tag, per group, per organization or random number of datasets. If he skips filtering, all datasets from portal and their relations will be loaded. RELIN component gives output depicted on Fig. 4. All datasets matching user filters and their related datasets are presented in the page. With the given datasets, user can: create, update and delete relations between the datasets and to select datasets for which want to obtain a semantic description of the relations in selected format.

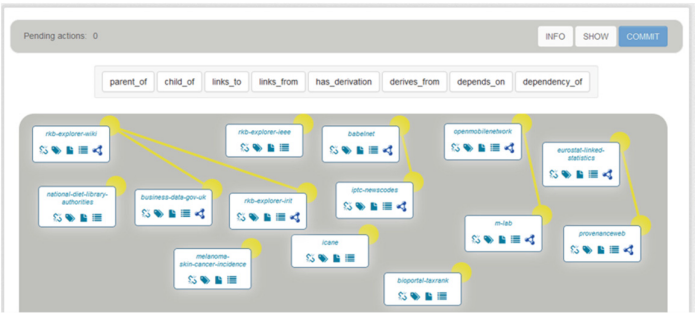

Fig. 4. Visual representation of dataset relations

User can perform following operations: remove or create relation between datasets. Every user action is saved in the list of actions pending to be executed in CKAN. After user finishes changes on the datasets, he needs to commit actions to CKAN. Removing existing relation is enabled through the user interface by clicking on the relation and choosing option delete from the context menu. When a user chooses to create relation between two datasets, for example *rkb-explorer-ieee* and *rkb-explorer-wiki* with child_of relation, firstly he needs to choose relation type from the application menu. Direction of the connection determines subject and object of the dataset relation, starting dataset is subject and ending dataset is object. Application examines meta-data of these datasets based on model described in Table 1 and returns results to the user. In these concrete example, user pick was child_of, and this type is not matched with the one proposed by the model links_from (Fig. 5). User can choose to proceed with his action and relate datasets in child_of relation or take the suggestion from LIRE and go for links_from relation.

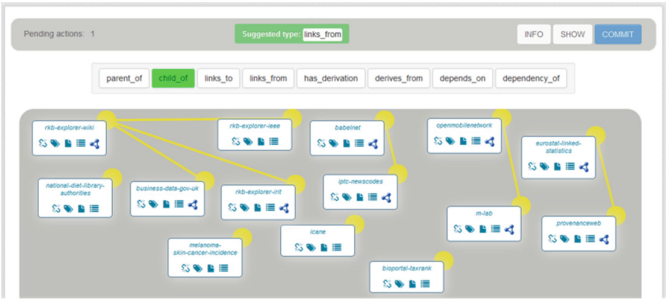

Fig. 5. Relating datasets

5 Conclusion

In this paper we presented an architecture for relating datasets and modelling or managing their relations with linked data principles and techniques along with model for their semantic representation. It reduces effort needed to preview datasets in order to relate them based on characteristics and data that describe them. As we have shown in Sect. 2, so far there is no research in the area of datasets relations and their modelling by using linked data. Dataset relations offer local level of relating, but if we describe them by linked data, dataset can be enriched with new data and information and with added semantics. For that purpose an RDF graph model for describing relations between two datasets is defined using voID vocabulary. Mapping between CKAN relations properties and voID statements shows that there are simple way for producing linked datasets. The modular nature of the proposed architecture makes it applicable to other portals except CKAN, but it requires additional time to review and analyze the data that describe the datasets, which are an essential element of our architecture. Future work includes investigation of the possibility for the development of the model and tool for semantic management of datasets and their relations and platform for accessing linked datasets based on defined RDF model. Also, incorporating proposed linked datasets model into the CKAN will be of great help to the users and developers in creating semantic applications.

References

1. Ayers, D.: Evolving the link. IEEE Internet Comput. **11**(3), 95–96 (2007)
2. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked open government data analytics. In: Wimmer, M.A., Janssen, M., Scholl, H.J. (eds.) EGOV 2013. LNCS, vol. 8074, pp. 99–110. Springer, Heidelberg (2013)
3. Sheridan, J., Tennison, J.: Linking UK government data. In Proceedings of the WWW 2010 Workshop on Linked Data on the Web (2010)
4. Berners-Lee, T.: Linked data. http://www.w3.org/DesignIssues/LinkedData.html
5. Veljković, N., Bogdanović-Dinić, S., Stoimenov, L.: Benchmarking open government: an open data perspective. Gov. Inf. Q. **31**(2), 278–290 (2014). Elsevier

6. Veljković, N., Bogdanović-Dinić, S., Stoimenov, L.: Exploring collaboration between public administrations through the notion of open data (ICIST 2015), Kopaonik, 8–11 March, vol. 1, pp. 122–127 (2015)
7. CKAN Open data portal. http://ckan.org/
8. Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W.: Linked open government data: lessons from data.gov.uk. IEEE Intell. Syst. **27**(3), 16–24 (2012)
9. Government, H.M.: Putting the frontline first: smarter government. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/228889/7753.pdf
10. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: Proceedings of the 2nd Workshop on Linked Data on the Web (2009)
11. Maali, F., Cyganiak, R., Peristeras, V.: Enabling interoperability of government data catalogues. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 339–350. Springer, Heidelberg (2010)
12. Ding, L., Lebo, T., Erickson, S.J., DiFranzo, D., Williams, T.G., Li, X., Michaelis, J., Graves, A., Zheng, G.J., Shangguan, Z., Flores, J., McGuinness, L.D., Hendler, A.J.: TWC LOGD: a portal for linked open government data ecosystems. Web Semant. Sci. Serv. Agents World Wide Web **9**(3), 325–333 (2010)
13. Kalampokis, E., Tambouris, E., Tarabanis, K.: A classification scheme for open government data: towards linking decentralised data. Int. J. Web Eng. Technol. **6**(3), 266–285 (2011)
14. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: Mika, P., et al. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 245–260. Springer, Heidelberg (2014)
15. Zuiderwijk, A., Jeffery, K., Janssen, M.F.: The potential of metadata for linked open data and its value for users and publishers. JeDEM-e-J. e-Democracy Open Gov. **4**(2), 222–244 (2012)
16. Janssen, M., Estevez, E., Janowski, T.: Interoperability in big, open, and linked data-organizational maturity, capabilities, and data portfolios. Computer **47**(10), 44–49 (2014)
17. Dublin Core Metadata Initiative: DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/
18. W3C: Data Catalog Vocabulary (DCAT). http://www.w3.org/TR/vocab-dcat/
19. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Linkset. http://www.w3.org/TR/void/#linkset