Enabling visual analysis of tags hierarchy and usage within open data portals

Milena Frtunić Gligorijević*, Miloš Bogdanović*, Nataša Veljković *, Leonid Stoimenov*
* Faculty of Electronic Engineering, University of Niš, Niš, Serbia
milena.frtunic.gligorijevic@elfak.ni.ac.rs, milos.bogdanovic@elfak.ni.ac.rs, natasa.veljkovic@elfak.ni.ac.rs, leonid.stoimenov@elfak.ni.ac.rs

***Abstract*— Data openness and transparency initiatives have led to a large amount of data being published on open data portals. These portals are focused on making the published data both accessible and discoverable. However, it is not a rare situation that metadata is incomplete, making it difficult for the users to obtain the desired information. To improve the discoverability, it is important to categorize datasets with missing category based on the available information. To do so, it is crucial to understand the usage and hierarchy of datasets' metadata elements, particularly tags. In this paper we want to address this issue by introducing a tool for interactive visual analysis of the tags' usage as well as the hierarchy of datasets metadata tags that describe different datasets in a single category of datasets. The tool we present relies on Formal Concept Analysis for creating a hierarchy of the usage of tags and combinations of tags used to describe different datasets. Furthermore, within the tool, we provide visualization of concept lattice that enables analysis of the links between the tags and determining the importance of a tag and a combination of tags within a category.**

Keywords: open data, formal concept analysis, visualization

I. INTRODUCTION

In the past decade, there have been numerous open data movements and initiatives towards data transparency and reuse [1]. This growing demand for openness of public and private organizations has increased the importance of open data [2]. All these initiatives resulted in the creation of open data portals (ODP), where a significant amount of data has been published over the years, with the intention of making it available for anyone to use, exploit in innovative ways, and generate added value out of it [3]. Today, there are numerous portals that store data from different data sources and available data within them is increasing every year. For example, over 260 open data portals with 1.1M datasets that describe 2.1M unique resources were used in 2016, by Neumaier, Umbrich, and Polleres for conducting a research on open data quality [3]. Government data portals are particularly interesting due to their scale and breadth of data they hold and a wide range of subjects they cover [4]. This broad range of subjects covers topics like budgeting, environment, transportation, licenses, statistics, health etc.

Open data portals are usually organized as catalogues where a dataset aggregates a group of data files (resources) that can be accessed or downloaded, and each dataset is accompanied with a metadata record that contains descriptive information about the dataset. Dataset's metadata is organized as key-value pairs, where the key holds a label for a specific data property and the value is a numerical or textual value that corresponds to that label. Different open data portals use different metadata scheme that specifies the data elements of which a metadata instance is composed. Although the schemes between the portals differ in some elements [5], they all have some common keys and keys that can be used for representation of common ODP features.

To facilitate the process of discovery, selection and finding desired datasets, open data portals provide multiple search criteria via Web application to the common users, and via Application Programming Interface (API) for the advanced users. Usual search options on the open data portals include category, format, publisher and tags. Categories are usually a pre-defined list of topics and ODP users use them as a filter for finding datasets from one specific topic and/or to bulk download data for further analysis. For narrower browsing, users take advantage of tags as textual descriptions containing keywords that describe a specific dataset.

However, often the cause of difficulties when browsing datasets is the fact that the metadata is missing. Not only does the missing metadata directly affects search and discovery services to locate relevant datasets based on user's needs, but it also requires a substantial amount of time to manually scan the portal for finding all relevant information [3]. This problem becomes even greater as the number of available datasets increases. From the perspective of data findability, the absence of information on the dataset's category presents a problem. In a research published in 2019, the authors have determined that some portals have significant number of uncategorized datasets [6]. Therefore, to improve findability, as main feature of open data portals, it is important to improve the quality of dataset's metadata by filling out the missing value for dataset's category. This should be done based on existing dataset's metadata and tags present an adequate choice for this process since they represent a list of keywords and phrases that contain descriptive knowledge of dataset's content and structure.

If the usage of tags is foreseen as the base for categorizing uncategorized datasets, it is important to analyze the usage and appearances of different tags in different categories. Same tags may be used as descriptions for different datasets that belong to different categories. Further, one dataset may belong to more than one category which leads to conclusion that certain combinations of tags may appear in more than one category. Therefore, all this should be taken into consideration and this analysis should include answering

the questions such as: How often does the tag appear in the category? How important is one tag in the category? Which combinations of tags are used together for describing datasets in one category? How important is the combination of tags in the category? What is the hierarchal order of the usage of tags and combinations of tags used to describe different datasets in one category? Furthermore, for a better understanding of such analysis, it is important to visualize the usage of tags within categories and hierarchal order among them. This kind of visual analysis is extremely valuable when working with large amount of information. A good comparison can be made with Linked open data where visualization is one of the core features for exploring data [7]. Therefore, since the number of datasets and tags describing them can be very large, it is important to have proper visualization that will support the process of the analysis.

For that reason, within this paper we propose the solution for the visual analysis of the usage of tags and categories within datasets on open data portals. We use Formal Concept Analysis for creating hierarchal order among tags used for describing datasets within one category. The created hierarchy is then used within the presented tool for the visual analysis.

The rest of the paper is organized as follows. Section 2 covers background in the field of open data and open data portals, and related work that covers our research approach regarding Formal Concept Analysis. Section 3 presents the tool for the visual analysis and the final section presents the conclusion of this paper.

II. BACKGROUND AND RELATED WORK

A. Open data portals and metadata usage

The idea of open data refers to data that is available free of charge for the public without any limitations [8]. Open Definition established the principles that define "openness" in relation to data and content. Their most succinctly definition is: Open data and content can be freely used, modified, and shared by anyone for any purpose [9]. In Open Data Handbook three major characteristics of open data are highlighted: availability and access, re-use and redistribution, and universal participation [10]. The eight open government data principles describe basic properties of data that need to be present so that we can label that government data as "open", and those are: complete, primary, timely, accessible, machine processable, non-discriminatory, non-proprietary and license-free [11].

The value of open data is recognized when it is used, thus publishers need to provide easy discoverability [12]. For that reason, governments and public institutions around the world publish their data on open data portals specifically designed to meet that requirement. These portals are built using one of the available portal frameworks. Most of them use open source platforms like Comprehensive Knowledge Archive Network (CKAN), and Drupal Knowledge Archive Network (DKAN) or proprietary software like Socrata and Opendatasoft. Every ODP software framework defines its own structure and limits considering dataset's metadata and attributes that can be used and defined. As given in [5] most of them depend on Data Catalog vocabulary (DCAT) metadata element set [13], but they also have some customized fields for storing dataset's features.

Metadata is data about data [14]. In open data portals, metadata is structured information that describes different datasets features and is used, among other things, for browsing datasets. It is organized as key-value pairs, whereas the key holds a label for a specific data property and the value is information that corresponds to that label. Some of the common dataset's metadata keys represent title, description, category, publisher, tags, resources, etc.

To provide better discoverability, ODPs use several metadata keys (meta-keys) to organize browsing via Web or API. The most important browsing meta-key is the one that corresponds to the dataset's category. Every ODP usually has limited number of pre-defined categories to which one dataset may belong. Depending on the ODP, one dataset may belong to one or more categories and meta-key that corresponds to category may be labeled differently. However, the most common meta-key labels are "groups, "category", and "theme". Another important meta-key for browsing datasets on the ODP is the one that corresponds to tags. Tags represent list of keywords and phrases that closer describe the dataset and can be used for a more focused dataset search. The most common meta-key labels for tags are "tags" and "keywords". Tags are entered by the user in the process of publishing the dataset and represent free-text enter by will. Therefore, as observed by Maali et al. in [15], categories are chosen from a controlled vocabulary which enables intuitive browsing of datasets and gives an instant overview of the available data in a catalogue, while tags lack that capacity due to their low consistency.

However, to ensure that a dataset will be discoverable, an appropriate quality of its metadata should be ensured. Missing metadata directly affects the search and discovery services to locate relevant and related datasets for consumer needs [16]. The problem of metadata quality, as well as consequences it brings in terms of data discoverability, increase together with the increase of number of datasets on the open data portal. The problem of incomplete and inaccurate information about the datasets was reported in multiple papers like [17], [8] and [16]. Knowing the existing problem of metadata quality, authors in [8] identified the need for an automatic quality assessment since evaluating and improving existing metadata is not always feasible and in [3], authors monitored over 260 open data portal regarding quality issues. Furthermore, quality of published data is also very important for e-government assessment with benchmark frameworks [18].

Missing value for meta-key that corresponds to the category represent a major problem for dataset's discoverability. Therefore, we propose an interactive visual analysis of the usage of tags within the categories to assist with the correct positioning of the datasets without categories. We use Formal Concept Analysis for building hierarchy lattices that adequately illustrate the use and importance of the tags within one category.

B. Formal Concept Analysis

Formal Concept Analysis (FCA) was introduced in early 1980s by Rudolf Wille as a mathematical theory for formalization of concepts [19]. The foundation of formal description is constituted with the following definitions:

Definition 1 (Wille, 1982): A formal context is a triple $K := (G, M, I)$ which consists of a set G of objects,

a set M of attributes, and a binary relation $I \subseteq G \times M$. (g, m) ∈ I is read as "object g has attribute m".

Definition 2 (Wille, 1982): For $A \subseteq G$, let $A^I := \{m \in M \mid \forall g \in A: (g, m) \in I\}$, and dually, for $B \subseteq M$, let $B^I := \{g \in G \mid \forall m \in B: (g, m) \in I\}$.

If the following conditions are met: $A \subseteq G$, $B \subseteq M$, $A^I = B$, $B^I = A$, then a pair (A, B) is a formal context. Set A is named concept extent while set B is named concept intent.

Definition 3 (Wille, 1982): The set S(C) of all concepts of a formal context C together with a partial order $(A_1, B_1) \leq (A_2, B_2) :\leftrightarrow A_1 \subseteq A_2$ (which is equivalent to $B_1 \supseteq B_2$) is a complete lattice of C.

Formal concept analysis is mainly used for data analysis, more precisely for deriving implicit relationships between set of objects defined by attributes and those attributes. The most significant output of FCA is the concept lattice. Concept lattice is created by generating a collection of formal concepts logically organized into a hierarchy of concepts interconnected using subconcept-superconcept relations [20].

Over the last 20 years, FCA has been applied in the various filed of research. One of them is classification, where FCA is successfully applied for classification and creation of classification rules. Overview of simple classification methods based on FCA divided into categories is presented by Prokasheva et al. in [21]. Further, the extensive overview of FCA application in research fields like knowledge discovery, software engineering, information retrieval, and data mining are presented in research papers [22], [23], [24] and [25].

Visualization of concept lattice is very important for the analysis when using FCA. As reported in [25], visualization of hierarchical order of concept lattice structure is an important concern for practical applications of FCA. The main problem for visualization can be the size of the concept lattice due to the large number of the formal concepts which can lead to the very complex and impractical concept lattice.

In this paper, we will apply FCA for the creation of data structure that will reveal shared conceptualization based on tags' usage for describing datasets in the same category. Since open data portals provide a large number of datasets, visualization of concepts can be a challenge. For that reason, in this paper we will propose interactive visualization mechanisms to overcome the complexity of the concept lattice and provide a clear preview of concepts and relations relevant for the user to reach the desired conclusions.

III. INTERACTIVE VISUAL ANALYSIS

Our focus in this research is a development of visualization tool for interactive analysis of the usage as well as the hierarchy of tags that describe different datasets in the category. The visualization relies on the concept lattices that contain hierarchical order of tags for every category available on the open data portal being analyzed.

Firstly, all datasets that contain information about the category on the open data portal being analyzed are separated into groups based on the categories they belong to. After that, for each category, the creation of hierarchical order among tags is done by performing Formal Concept Analyses on the combination of tags appearing together in the datasets within the same category. The input for the algorithm is an incident matrix where the rows present all datasets belonging to the same category and columns present all tags appearing in the datasets. The output of the algorithm is one concept lattice per category.

Due to the nature of the tags and the fact that tags used for describing a dataset can be assigned an arbitrary value, the number of distinct tag values used across ODPs can be very large. Consequently, generated concept lattices may become extremely complex and hard to analyze. To address this issue, we have developed a visualization tool that overcomes the problem of complexity. To achieve this, the tool supports several options for the analysis of concept lattice:

- The analysis of the whole lattice – gives an overview of all nodes in the concept lattice and all links between the nodes. This preview is valuable for gaining a general impression of the whole structure of the hierarchy.
- Preview of only a selected number of levels in the lattice – gives an overview of the part of the concept lattice where only nodes that belong to the selected number of levels are presented. This option is especially useful when the concept lattice is complex because the user has an option to analyze only the upper part of the lattice which contains tags with higher importance.
- Browsing positions of particular tags in the lattice – this is a search option available to the user. The user is given an opportunity to browse specific tags in the lattice. The search results will be shown in the concept lattice by highlighting nodes using different colors based on the subset of the search tags those nodes cover. The tool does not highlight all the nodes that cover some of the searched tags but only the most significant ones. Node's significance is determined based on the number of searched tags covered by the particular node and number of additional tags. Based on the significance, nodes are colored with different colors: red, purple, green, and blue. Red color is used to mark nodes that contain all searched tags and no additional tags. Purple is used to mark nodes that contain part of the searched tags and no additional tags. Green is used to mark nodes that contain all of the searched tags and minimum possible number of additional tags. Blue color is used to mark nodes that contain part of the searched tags and minimum possible number of additional tags. This option is especially useful for fast discovery of the particular tags.
- Browsing connections to all top and all bottom nodes from the selected node – this option allows the user to select a node, and by doing so, all parent and children nodes as well as the links between them will be highlighted. This allows users to, easily and clearly, analyze connections between particular subset of nodes.

Additionally, presented visualization options can be combined which further facilitates analysis of the position, usage and importance of tags in one category, especially if the concept lattice contains huge number of nodes, tags and datasets.

Along with this visualization options, the user is given the statistical information regarding the lattice. This functionality visualizes information such as total number of nodes in the lattice, total number of datasets in the lattice, total number of attributes in the lattice, and total number of levels. Further, the tool supports options for the selection of a particular node which gives a preview of all the information regarding the selected node: node name, level at which the node is positioned, list of tags covered by the selected node, list of attributed covered by the selected node and the number of datasets that are represented by the covered attributes.

An example of the usage of the presented tool is given using a concept lattices generated from the available datasets on the Canada's open data portal. This portal contains over 80000 datasets organized into 19 categories where some of the categories contain a significant number of datasets and tags used for describing those datasets.

For the purpose of demonstration of the visualization, we have used the category *Law* from the Canada's open data portal. This category is one of the smaller categories on this portal in terms of both the number of datasets and the number of different tags used for describing the datasets belonging to this category. Although it is one of the smaller categories, the concept lattice generated for this category is rather complex – it contains 325 nodes organized in 8 levels. The preview of the summary information about the whole concept lattice for category *Law* and an example of the summary information about the single selected node in the concept lattice is presented in Fig. 1.

The visualization option for analyzing the complete lattice, especially complex as this one, can be used for some initial analysis of the overall positions of some nodes and their position in the lattice. In Fig. 2 we have

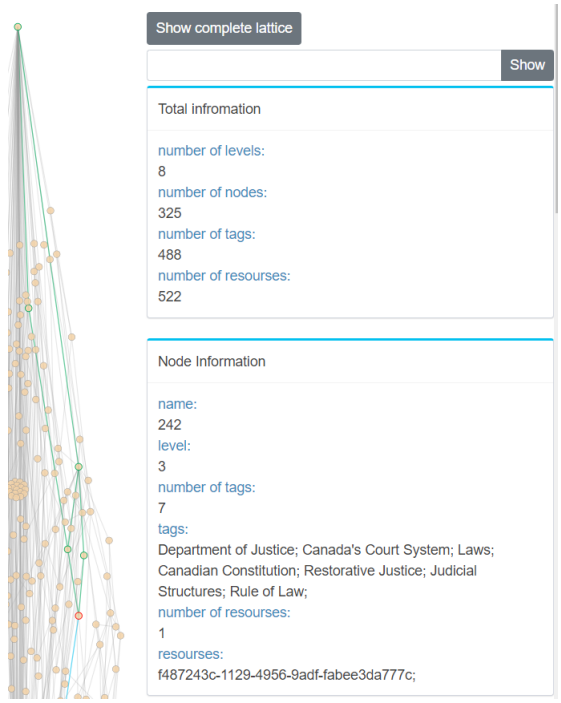

Figure 1. Preview of the summary information about the whole concept lattice and example of the summary information about a node

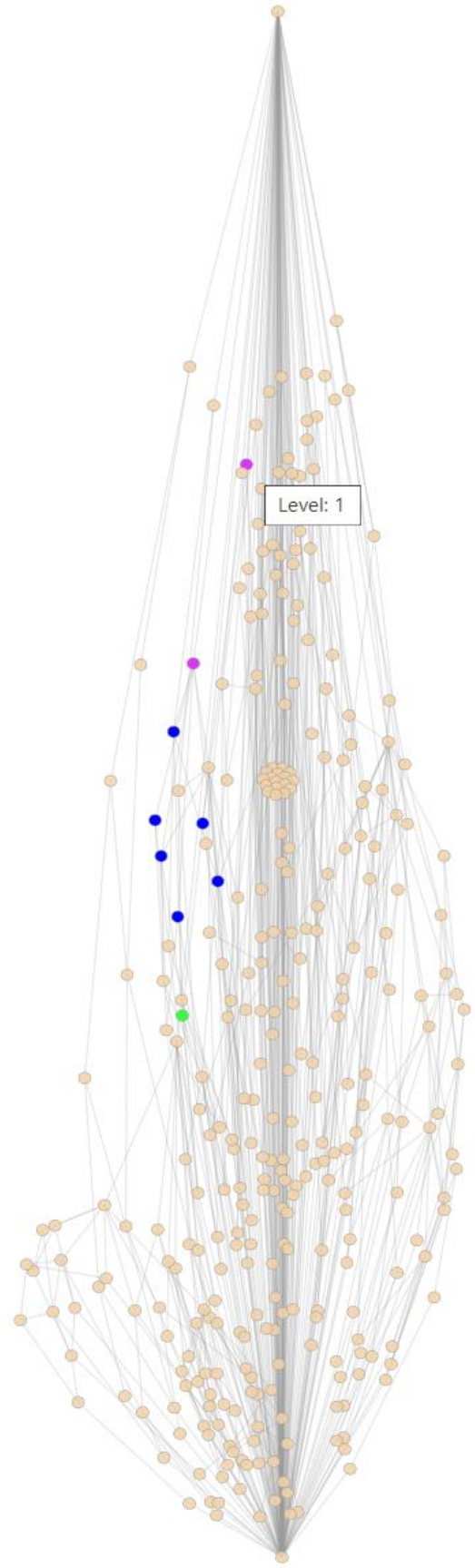

Figure 2. Complete concept lattice for category Law from Canada's open data portal

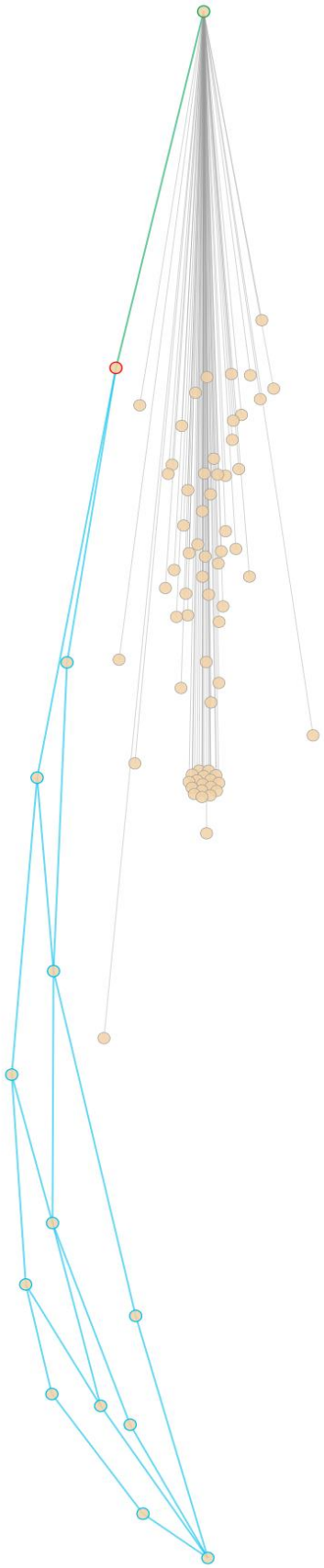

Figure 3. Combination of visualization options for category *Law* from Canada's open data portal

demonstrated how this can be done. We have used the possibility of previewing the whole concept lattice together with browsing for nodes that contain some tags. In this example we searched for tags: *Research*, *Annual reports.* As shown in Fig. 2, the concept lattice does not contain a node that covers both searched tags and no additional tags. Additionally, there are nodes in the lattice that cover only searched tags separately (purple nodes), and by placing the cursor on them it can be noted that both nodes are on the first level in the lattice. Also, there are several nodes that contain part of the searched tags and the same minimum number of additional tags (in this case one additional tag). Finally, in the results, there is one green node that contains both searched tags and some additional tags (in this case three additional tags). All detailed information about the nodes can be acquired by selecting the node of interest.

However, for deeper analysis such as analyzing connections between the nodes, huge number of nodes and links may cause confusion and problems. For that reason, to enable clearer preview, our tool offers a possibility for hiding some of the nodes trough an option of selecting only few levels to be shown.

One example of analysis with some of the nodes being hidden is shown in Fig. 3. In this figure, a preview of the selected number of levels in the lattice is combined with the preview for browsing connections to all top and all bottom nodes from the selected node. The selected number of levels is set to one, which significantly reduced the number of nodes. After that, browsing the connections can be done with a simple click on the node of interest. After the selection, the chosen node is marked with a red border. Links from the selected node to nodes higher in the hierarchy, as well as the upper nodes' borders are colored green. Additionally, links from the selected node to nodes lower in the hierarchy and lower nodes' border are colored blue.

As it can be seen, such preview is much clearer comparing to the one that can be seen in Fig. 2. This type of preview can be used for a detailed analysis of the relations between subsets of interrelated tags within one category. Our tool provides several modes for concept lattice preview and analysis. These modes can be combined to enable easy overview analysis depending on the user's needs.

IV. CONCLUSION

As the amount of data on open data portals continues to grow, it is becoming extremely challenging to maintain the data discoverability. This difficulty becomes even greater when dataset's metadata values are missing, particularly information like dataset's category. Therefore, categorization of uncategorized datasets based on the existing metadata, like available tags, is very important. However, prior to categorizing datasets, it is crucial to analyze the usage and hierarchy of tags within different categories on the open data portal. Due to the nature and number of available tags, an adequate visualization mechanism is needed.

In this paper, our aim was to introduce the interactive visualization tool we developed, that will enable clear preview, adequate for analyzing popularity, importance and usage of different tags and combination of tags within one category. Through multiple preview options and the

search mechanism, we have provided the ability for the user to examine the position of a particular tag and combination of tags in the concept lattice. Further, we provided the ability for the user to inspect the connections between tags with both the parent and children nodes as well as main statistical information that can further give the information about the importance of a particular node in the concept lattice. Due to the possibility of showing only significant nodes and connections, our tool overcomes the problem of complexity of the concept lattice, which is one of the main problems when working with a such large number of datasets, and such diversity of tags and combination of tags which open data portals have. Therefore, the presented visualization can be used as means for making decisions regarding the appropriate category for uncategorized dataset.

REFERENCES

[1] J. Attard, F. Orlandi, S. Scerri, S. Auer, "A systematic review of open government data initiatives. Government Information Quarterly", 32(4), 399-418, 2015

[2] S. Kubler, J. Robert, S. Neumaier, J. Umbrich, Y. Le Traon, "Comparison of metadata quality in open data portals using the Analytic Hierarchy Process," *Government Information Quarterly*, vol. 35, no.1, pp.13-29, 2018.

[3] S. Neumaier, J. Umbrich, A. Polleres, "Automated quality assessment of metadata across open data portals," *Journal of Data and Information quality*, vol. 8, no.1, pp. 2:1-2:29, 2016.

[4] S. van der Waal, K. Węcel, L. Ermilov, V. Janev, U. Milošević, M. Wainwright, "Lifting open data portals to the data web," *In Linked Open Data--Creating Knowledge Out of Interlinked Data*, Springer, Cham, pp. 175-195, 2014 .

[5] P. Milic, N. Veljkovic, L. Stoimenov, "Comparative analysis of metadata models on e-government open data platforms," *IEEE Transactions on Emerging Topics in Computing*, 2018.

[6] M. Frtunić Gligorijević, M. Bogdanović, N. Veljković and L. Stoimenov, "Open data categorization based on formal concept analysis," in IEEE Transactions on Emerging Topics in Computing, 2019.

[7] F. Desimoni, L. Po, "Empirical evaluation of Linked Data visualization tools," *Future Generation Computer Systems*, vol.112, pp. 258–282, 2020.

[8] K. J. Reiche, E. Höfig, "Implementation of Metadata Quality Metrics and Application on Public Government Data," *2013 IEEE 37th Annual Computer Software and Applications Conference Workshops*, Japan, pp. 236-241, 2013.

[9] The Open Definition, retrieved November, 2020 from https://opendefinition.org/

[10] Open Knowledge Foundation: Open data handbook, retrieved November, 2020 from https://opendatahandbook.org/

[11] The Annotated 8 Principles of Open Government Data, retrieved November, 2020 from https://opengovdata.org/

[12] A. Assaf, R. Troncy, A. Senart, "HDL-Towards a Harmonized Dataset Model for Open Data Portals," *In USEWOD-PROFILES@ ESWC*, pp. 62-74, 2015.

[13] Data Catalog Vocabulary, https://www.w3.org/TR/vocab-dcat/

[14] E. Duval, "Metadata Standards, What, Who & Why," *Journal of Universal omputer Science*, Springer, 7(7), 591-601, 2001.

[15] F. Maali, R. Cyganiak, V. Peristeras, "Enabling Interoperability of Government Data Catalogues," *In Proceedings of EGOV 2010*, pp. 339-350, 2010.

[16] J. Umbrich, S. Neumaier and A. Polleres, "Quality assessment & evolution of open data portals," in *Proceedings IEEE International Conference on Open and Big Data*, IEEE, Rome, 2015, pp. 1-8.

[17] J. Kučera, D. Chlapek, M. Nečaský, "Open Government Data Catalogs: Current Approaches and Quality Perspective," *In: Technology-Enabled Innovation for Democracy, Government and Governance. EGOVIS/EDEM 2013. Lecture Notes in Computer Science*, vol. 8061. Springer, Berlin, Heidelberg, 2013.

[18] N. Veljković, S. Bogdanović-Dinić, L. Stoimenov, "Benchmarking open government: An open data perspective," *Government Information Quarterly*, 31(2), 278-290, 2014.

[19] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts," Ordered Sets, Springer, Dordrecht, pp. 445–470, 1982.

[20] R. Belohlavek, "Introduction to formal concept analysis," 2008, retrieved November, 2020 from https://phoenix.inf.upol.cz/esf/ucebni/formal.pdf

[21] O. Prokasheva, A. Onishchenko, S. Gurov, "Classification methods based on formal concept analysis," *FCAIR 2012 – Formal Concept Analysis Meets Information Retrieval*, p. 95, 2012.

[22] J. Poelmans, S.O. Kuznetsov, D.I. Ignatov, G. Dedene, "Formal Concept Analysis in Knowledge Processing: a Survey on Models and Techniques," *Expert Systems with Applications*, Vol. 40, Issue 16, pp. 6601-6623, 2013.

[23] J. Poelmans, D. I. Ignatov, S. O. Kuznetsov, G. Dedene, "Formal concept analysis in knowledge processing: A survey on applications", *Expert Systems with Applications*, Vol. 40, Issue 16, pp. 6538-6560, 2013.

[24] S. Doerfel, R. Jaschke, G. Stumme, "Publication analysis of the formal concept analysis community," *ICFCA 2012, LNCS*, Springer, 7278, pp. 77–95, 2012.

[25] P.K. Singh, C. Aswani Kumar, G. Abdullah, "A comprehensive survey on formal concept analysis, its research trends and applications," Int J Appl Math Comput Sci, vol. 26, no. 2, pp. 495–516, 2016.