Linked Open Data Life Cycle

Rihab Fahd Al-Mutawa and Arwa Yousuf Al-Aama
Computer Science Department,
Faculty of Computing and Information Technology,
King Abdul Aziz University,
Jeddah 21589, Saudi Arabia

ABSTRACT
Open data refers to the publishing of information in standard formats that are interoperable and open in order to ease its access and reuse for many purposes such as to encourage citizen participation in the development of their cities and to improve economic opportunities. The strategies of open data have gained the attention of government organizations to ensure transparency and reusability of the data. Many governments across the world have promoted openness as a strategy. Open data enables a government to be transparent and accountable to the citizens of the country. Smart cities is concerned with improving the urban environment by efficiently using existing resources. Those resources can be better understood using linked open data. In the contemporary context, organizations face difficulties in linked open data management, and there is a lack of awareness about the steps to be taken. Additionally, there is no standardized linked open data life cycle yet. This paper proposes a comprehensive open data life cycle model based on literature covering both data supply and demand processes and involving the roles of internal and external stakeholders. The paper presents thorough details of the suggested open data life cycle. The discussed models in the literature have highlighted technical aspects for a smooth and secure publication. Open data life cycle models have the potential of facilitating the procedure of opening up data. They offer phases and steps related to opening up data thus play a pivotal role in the successful application of open data.

General Terms
Applied Sciences, Data Science, Smart Cities.

Keywords
E-government, Open Data, Linked Data, Public Sector Information, Data Management, Life Cycle Model, Citizen Participation.

1. INTRODUCTION
Open data initiatives are growing rapidly around the world due to the high economic and social values of open data [1]. It enables organizations to ensure transparency, be held accountable to the citizens of a country, improve the economic activity through data reuse [2], facilitate the development of innovative software applications [3], improve government services [4], benefit service and policy designers, improve a given country's competitiveness [5], and is a foundation for citizen participation, which is one of the key factors of success for a smart city [5], [6]. Therefore, the organizations are interested in open data processes and services. Models that describe data objects such as dataset files through a set of stages ordered in a timely manner are called life cycle models [7]. Life cycle models are found beneficial for managing open data activities [8]. Those models are concerned about the data and metadata structures and reaching to an appropriate quality level [9]. Supporting the data through the linked data life cycle is a challenging task whether the support is through developing research approaches, standards, technology or tools [10]. There are only a few production-leveal quality tools available for modelling or publishing linked open data such as LOD2 Stack [9]. The LOD2 Stack is a publishing tool that supports the life cycle of linked data from extraction, enrichment, interlinking, fusing to maintenance [11]. Moreover, the life cycle of linked open data and metadata should address supply and demand sides, unlike the current life cycles, which mainly focus on the supply of data (identification and selection, modelling and cleansing, and publishing and linking). However, the demand side adds finding and retrieving, integration and reuse, and providing feedback [9]. All of the previously mentioned steps within those life cycles will therefore be embedded within the suggested life cycle.

Section 2, discusses related work. Section 3, represents the details of the suggested life cycle. Section 4, discusses implications of the study and suggests future work. Finally, the conclusion of this work is presented in Section 5.

2. RELATED WORK
There are numerous versions of varying complexity exist of an abstracted data management or data life cycle [7]. However, they can share some common steps. While reviewing linked open data life cycle models, it was found that all life cycles in [12], [13] and [14], go through five common processes as follows:

- Data selection.
- Data preparation.
- Data publishing.
- Data interlinking.
- Data discovery.
- Data reuse.

And they differ in the steps and roles that shown in table 1. In this paper, all the steps found in the literature are included in addition to the six identified roles that are adapted within each process of the life cycle: top management, information manager, legal advisor, community manager, data owner, and end users. Next section, describes the proposed linked open data life cycle model.

Table 1. Differences between linked open data life cycle models

Difference	[13]	[12]	[14]
Number of steps	9	5 main and 10 internal	5
Managing strategy		✓	
Data creation	✓		✓
Advertise		✓	

data & build community			
Managing feedback & proposition		✓	
Data and metadata enrichment	✓		✓
Data curation	✓	Partially with managing the data	
Data reuse	✓	Reuse is included within the main stage only	✓
Roles in each step	n/a	Information manager, top management, community manager, legal advisor, and data owner	Data owners, citizens, businesses, and civil society

3. OPEN DATA LIFE CYCLE

This section, explains the proposed open data life cycle model based on [12], [13] and [14]. All the processes in the life cycle need to be covered, thus stakeholders can follow the standard process. The six stakeholders involved in this life cycle are top management, information manager, legal advisor, community manager, data owner, and potential users such as businesses and enterprises, citizens, civil society, developers, and researchers [12], [15], [14]. The life cycle is shown in figure 1. It is made up of three sections, namely, a pre-processing section that is represented by the rectangular shapes (i.e., preparing the data to be published), an exploitation section that is represented by the oval shapes (i.e., using the published data), and a maintenance section that is represented by the hexagon shape (i.e., maintaining the published data in order to be sustainable) [13]. This open data life cycle model consists of eight phases: identification, preparation, publication, add-value, promote, reuse, evaluation and curation, each consisting of a range from one to three steps. The detail of each step of the life cycle is provided below.

3.1 Identification

The first phase of opening up data consists of creating data, strategy setting and identification of datasets to be opened up.

3.1.1 Setting the strategy

This step defines an open data strategy based on use of open data inside and outside the organization for all research executed with public funds. It evaluates the pilot project and gets advantage of assessing data proposition at the evaluation phase thus open data and the needed modification can be embedded in the organizational strategy and work processes including the issue of community building and support to create more value from the datasets that are opened up. The involved roles in this step are top management, information manager, community manager and data owners. The first two roles are necessary to define the scope of the strategy internally, while community manager is connected with potential users as well as the data owners as some of them should actively be involved in communities around the data. Top management is responsible for embedding open data in the organizational strategy and processes in addition to investigating how the process of opening up data is taking place in practice for the whole organization. Information manager can balance the renewals from top-down and bottom-up [12].

3.1.2 Data creation

Creation of data is the technical beginning of open data life cycle. In government agencies, the identification and preparation of relevant raw data is usually a daily procedure, and data can be collected only for the purpose of publishing it or because it was requested [13], [14]. The stakeholders involved in this step are the data owners such as civil servants [14].

3.1.3 Data selection

This process involves selecting the suitable data for publishing which requires removing any private data or personal identifying information, in addition to specifying the open data policies [13]. This step is very complex and widely acknowledged as challenging in the literature, as there are many issues concerning the fear of misinterpretation of data and privacy sensitivity of data [16]. It is mandatory for datasets to comply with regulations regarding three criteria: privacy, security and ownership including intellectual property in order to be opened up. Some data owners may not yet be aware of what open data is. They should be supported to open up their data with the help of information managers. Supporting data owners is highly required as they are responsible for preparing the data to be opened up [12].

3.2 Preparation

The second phase of opening up datasets consists of setting the requirements for the data to be opened up and preparing the data and metadata technically.

3.2.1 Setting requirements

In this step, both information manager and legal advisor formulates the requirements for the data in order to be published. These requirements include technical requirements that include certain standards such as publishing standards, data quality level, and metadata; legal requirements such as the license for reuse; and economic requirements such as the business model and value proposition. In addition, the legal advisor checks whether the selected data for publishing is suitable to be public [12].

3.2.2 Data harmonization

This involves data preparation for publishing with conforming to technical requirements in the previous step including the publishing standards, such as the Eight Open Government Data Principles or the Ten Open Government Data Principles [13], [17], [18]. However, data that can be tracked to individuals cannot be prepared for publishing or published thus if there is part of data that can be linked to individuals then it must be anonymized before publishing. This step, requires specifying the ownership of the data clearly so that data can be published freely in the next step. Moreover, data needs to be modeled as data is often captured in an unstructured way that fits its original purpose. And data should be labelled uniquely according to a Unique Resource Identifier strategy. Finally, to enable data reuse later during this life cycle, metadata is required, data must be converted into an open structure and machine readable format, and data gets stored following a specified formats. This step is carried out by both the information manager and the data owner [12].

3.3 Publication

The actual act of opening up the government data is by publishing it on a government platform by the data publisher, i.e. making the data available on the Web [13]. This phase involves ensuring findability of published datasets and metadata.

3.3.1 Ensuring findability

During the process of publishing data, the datasets must be registered in such a way that potential users can find what they are looking for. From a technical perspective, this is done by registering both data and metadata in data catalogues of the national open data platform [12], [14]. The metadata includes aspects such as data description, its purpose, licensing, and maintainer [14]. The responsibility of this step relies on the information manager and the data owners [12].

3.4 Value-increasing

This phase involves interlinking and integrating published datasets to add more values for the end-users.

3.4.1 Data interlinking

This allows the published data to increase its value, as linking or combining data reveals relationships between data and gives context to its interpretation according to the five star maturity model for linked open data [19], [8], [13]. The responsibility of this step is carried out by both the information manager and the data owner. There are several tools that can be used for linking structured data on the web such as SILK, LIMES, KnoFuss, RDF-AI, SERIMI, OKKAM [20].

3.5 Promotion

This phase involves advertising the published datasets to highly encourage data reuse by potential end-users [13].

3.5.1 Data advertising

This step is concerned with encouraging data consuming and reuse through proper advertisement of published data and actively raising awareness on its existence such as through organizing hackathons [13] and engaging with a community of potential re-users. Building and engaging in a community around the published data with external stakeholders is necessary to advertise the data and make sure that data would be used. Active community building can also motivate the external stakeholders to give feedback in the next phase of the life cycle, which helps to improve the quality of the data. Moreover, the reuse licenses which includes the conditions of reuse must be communicated to make sure that the end-users understand them. The community manager is responsible for active advertisement and collaborative communication with potential users that may want to use the data [12].

3.6 Reusing

This phase is for exploring and reusing published datasets and metadata, and collecting feedbacks from end-users.

3.6.1 Data exploration

This step is for trivial consuming of data [13]. The end-user can discover the published datasets and metadata by using a faceted browser features such as filtering, searching and RSS feed notifications in addition to examining open data by scrutinizing or visualizing it [14], [13] or may find datasets or metadata by executing a SPARQL query on a SPARQL endpoint [9].

3.6.2 Data reuse

This is an advanced way for consuming data. It enables the end-user to pro-actively use, reuse or distribute open data by creating mashups, leading out analysis, or innovating based on open data [13]. Using open data apps via Web and mobile devices by the end-users may generate new data, which will later go through the phases of the open data life cycle [14].

3.6.3 Feedback

Community is in a better position to think about reuse than data owners. Feedback is important as the end-users can give their opinions and suggestions about the open data platform or the published datasets; thus the public organization can make available the requested datasets which has a likelihood of being reused by programmers. The interest of the public in certain datasets is an important reason for delivering this data as open. Additionally, asking users for feedback helps increasing data quality. The community manager guides the information manager and data owners through this step of the process to achieve data reuse, while keeping track of visitors and users of published datasets [12].

3.7 Evaluation

The last phase of the life cycle is the evaluation for the provided feedback on open data and its strategy by end-users.

3.7.1 Feedback assessment

This step is for assessing the provided feedback on open data and its strategy by external stakeholders. All stakeholders except end-users are involved in this step, to determine the value of opening up data based on the received feedback. Both financial and societal gains are considered [12].

3.8 Curation

This phase of the life cycle is for managing data as needed throughout the life cycle according to any new requirements accommodated with unplanned changes.

3.8.1 Data curation

Data curation can happen at any stage and it is vital in ensuring sustainability of the data. It involves a number of processes, such as updating data and stale data, data cleansing, data and metadata enrichment [13]. Data enrichment is an optional task that increase the potential for dataset utilization in applications. Datasets and metadata are transformed into semantically richer formats (e.g., RDF) and express the context and the relation between data using annotation thus enabling machine-readability, data interpretation and seamless processing [14], [21], [22]. The responsibility of this step is relying on both information manager and data owners. Data owners make a plan for how to manage data and make sure that the quality of the data remains as required. The information manager is prepared for feedback and requests for support [12].

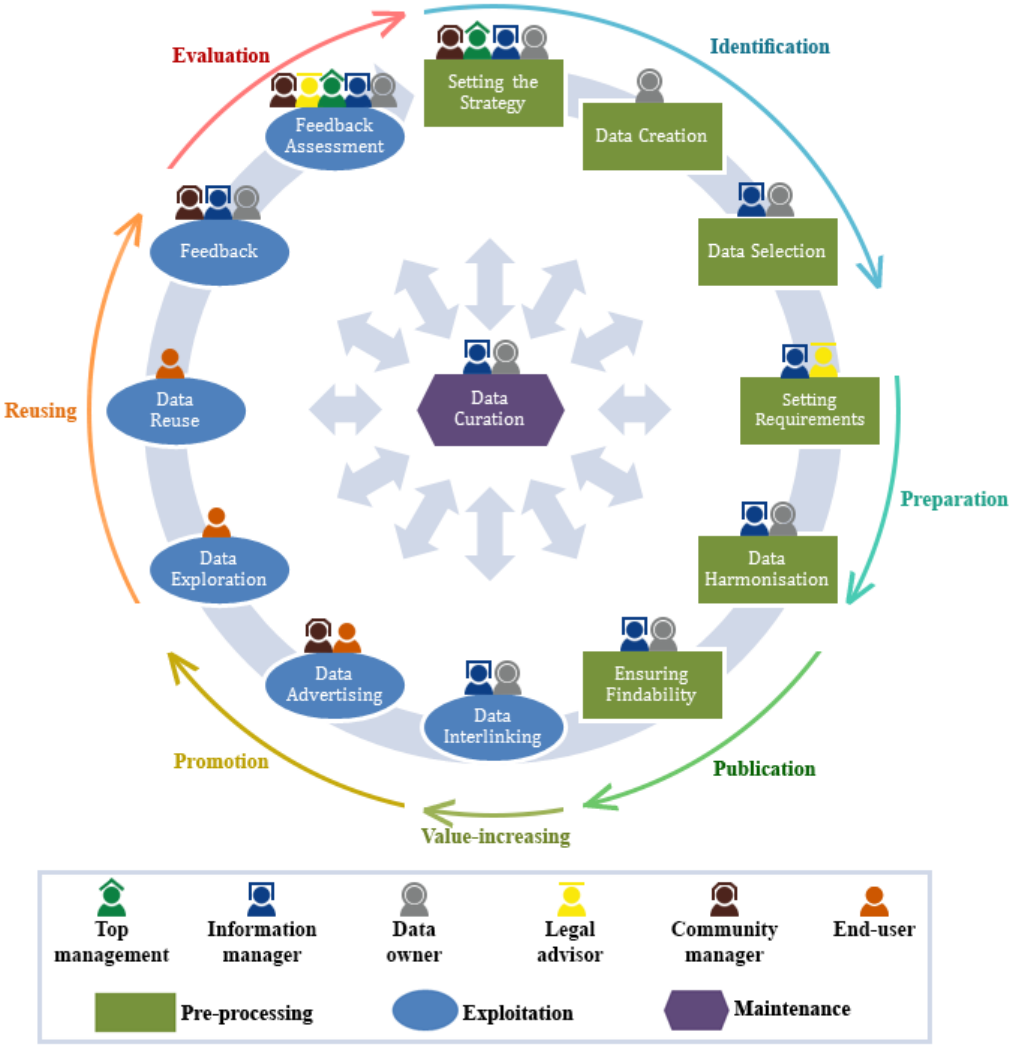

Fig 1: Linked Open Data Life Cycle Model

4. DISCUSSION AND FUTURE WORK

Any organization that offers open data is practicing and following a data management life cycle for their open data. The open data life cycle should manage the open data effectively thus enhance the level of transparency and accountability in addition to support the economic growth as data is the new oil for innovation and digital economy. According to [23], the economic value of open data estimated with many studies for the European Union at several tens of billions of Euros annually. All the studies of open data life cycle suggest their procedures and adds more value to the current life cycles that manage open data in order to improve the potential of data which has a positive impact on many aspects of life. Therefore, organizations need to keep a check on the new emerging life cycles. The proposed linked open data life cycle model of this research is aiming to contribute to literature related to building the foundation of providing successful and comprehensive guide of handling the supply and demand sides of open data in any organization through its high level overview of the generic stages and low level description of the executive activities associated with assigned roles of responsibility that are involved in its internal processes to ease and support the flow and progress of tasks during the multiple stages of the life cycle. This study collected and organized eight phases and twelve steps based on investigation of relevant studies with the results shown in table 1 then represented the proposed model. Every stage is important starting from data identification until data evaluation in addition to data curation that provides active management of data which can be needed any time during the life cycle.

Both public and private sector organizations can practice the proposed life cycle as they collect a wide range of different data types for performing their tasks and there are enormous amount of valuable untapped data by public and private sector organizations that could be get published and reused. However, the significance of government open data in general is increasing because of two reasons. First, the government data is mostly public data by law hence be made available for others to use. Second, the centrality and quantity of the data that government collects [23]. But with respect to smart cities context, private organizations can provide equal support for economic growth. This finding is different from the prior research in [13], in which researchers associated the use of linked open data life cycle to the government organizations. A private company may get buy-in from the city manager to collects and processes the resulting data for the city based on

installed sensors by this company in a transit system [24]. Consequently, it contributes to the smart city initiative if it offers any open data that may support smart mobility or other enablers of smart city. According to [25], the smart city ecosystem consists of eight smart components: economy, environment, governance, infrastructure, living, mobility, people, and services. There are many examples that could be provided by any sector and shows the importance of open data for the smart city. For example, for more efficient travel [26], some traffic data can be used to inform public transit riders of delays and suggest alternative routes or it can be used to help operations managers to understand how their services may be affected, such as, an incoming shipment of goods will arrive four hours later than expected. Other examples for open data related to cities are food safety inspections, crime reports, business licenses, etc. As can be seen, open data represents a powerful opportunity for cities to connect with the end-users in meaningful and life-changing ways which force the need for engaging the community through getting their feedback during the life cycle then assessing the feedback to deliver the open data as needed which would improve the chances of their acceptance and reuse. According to the case study of [12], building an engaged community of stakeholders in an early stage also outside the organization, is pivotal to the success of open data. The developers are among the end-users that participate in creating added value with their innovative applications in many aspects of smart city such as reduce energy consumption. The linked data and semantic web technologies that are imposed with the data interlinking and data curation processes respectively, are providing developers during the reuse with options for using or creating a seamless machine-readable mashups which increase the value and amount of the available data for reuse thus the chances of developing creative applications. This contributes to enabling the smart city as smart people is one of the smart city components. Therefore, serves the purpose of smart city; increasing the quality of urban life in the city.

As a smart sustainable city needs open data technology and its life cycle. Moreover, open data can be small or big data and cities need both. Therefore, for future work, this life cycle can be adapted and further extended and specialized to handle the case of linked open big data in smart cities as researches around smart cities are currently an active field of study and managing all the big data coming from the cities is one of the challenges of the smart cities [27] and [28].

5. CONCLUSION

Many public and private organizations are aiming to adopt an open data strategy to accomplish availability of data to the public. Data life cycle models provide effective and efficient management ways for attaining this objective. As demonstrated with the previously mentioned studies, each study imposed some significant phases and steps in their life cycle and being collected and analyzed with consideration to the different roles of stakeholders is contributing to raising the sufficiency level of identifying the processes within the linked open data life cycle. The suggested life cycle model is consisted of eight stages beginning from data identification and ending at data evaluation covering both the supply and the demand side regarding data. Life cycle models should not only concentrate on the technical aspects but also handle the involvement of the stakeholders. The internal and external interaction with stakeholders enhances the reusability of the data. It also makes open data strategy as an integral part of the corporate strategy. For future work, an idea of specializing the proposed life cycle is suggested to handle the case of big data as a combination of both open data and big data is useful for improving the urban quality of life.

6. REFERENCES

[1] Jetzek, T., Avital, M., & Bjorn-Andersen, N. (2014). Data-driven innovation through open government data. Journal of theoretical and applied electronic commerce research, 9(2), 100-120.

[2] Zuiderwijk, A., Janssen, M., & Dwivedi, Y. K. (2015). Acceptance and use predictors of open data technologies: Drawing upon the unified theory of acceptance and use of technology. Government information quarterly, 32(4), 429-440.

[3] Vosough Jr, P. (2013). Implementing an open data system and showing its benefits.

[4] Cowan, D., Alencar, P., & McGarry, F. (2014, June). Perspectives on Open Data: Issues and Opportunties. In Software Science, Technology and Engineering (SWSTE), 2014 IEEE International Conference on (pp. 24-33). IEEE.

[5] e-Government Program. (2014, November). Open Data. Retrieved from http://www.yesser.gov.sa/ar/mediacenter/events/Documents/opendata2014/Open_Data_Introduction_last_Nov19.pptx.

[6] Clarke, R. Y. (2013). Smart cities and the internet of everything: The foundation for delivering next-generation citizen services. Alexandria, VA, Tech. Rep.

[7] Ball, A. (2012). Review of data management lifecycle models.

[8] Zuiderwijk, A., Janssen, M., & Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. Information Polity, 19(1, 2), 17-33.

[9] European Commission. (2014). The Linked Open Government Data & Metadata Lifecycle [Powerpoint slides]. Retrieved from https://joinup.ec.europa.eu/sites/default/files/d2.1.2_training_module_2.1_the_linked_open_government_data_lifecycle_v1.00_en.pdf.

[10] Auer, S. (2014). Introduction to LOD2. In Linked Open Data--Creating Knowledge Out of Interlinked Data (pp. 1-17). Springer International Publishing.

[11] Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., ... & Stadler, C. (2012, November). Managing the life-cycle of linked data with the LOD2 stack. In International semantic Web conference (pp. 1-16). Springer, Berlin, Heidelberg.

[12] van Veenstra, A. F., & van den Broek, T. (2015). A community-driven open data lifecycle model based on literature and practice. In Case Studies in e-Government 2.0 (pp. 183-198). Springer International Publishing.

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

[14] Lapi, E., Tcholtchev, N., Bassbouss, L., Marienfeld, F., & Schieferdecker, I. (2012, July). Identification and utilization of components for a linked open data platform. In Computer Software and Applications Conference Workshops (COMPSACW), 2012 IEEE 36th Annual (pp. 112-115). IEEE.

[15] Schaffers, H., Komninos, N., Pallot, M., Trousse, B., Nilsson, M., & Oliveira, A. (2011). Smart cities and the future internet: Towards cooperation frameworks for open innovation. The future internet, 431-446.

[16] Zuiderwijk, A., Janssen, M., Choenni, S., & Meijer, R. (2014). Design principles for improving the process of publishing open data. Transforming Government: People, Process and Policy, 8(2), 185-204.

[17] Open Government Working Group. (2007). 8 principles of open government data. Retrieved from https://public.resource.org/8_principles.html.

[18] Sunlight foundation. (2010). Retrieved from https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

[19] Berners-Lee T. (2010). Linked data. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html.

[20] Nafis, F., & Chiadmi, D. (2016). Methods and Systems for the Linked Data. In Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015 (pp. 587-592). Springer, Cham.

[21] Nagao, K., Shirai, Y., & Squire, K. (2001). Semantic annotation and transcoding: Making web content more accessible. IEEE MultiMedia, 8(2), 69-81.

[22] Rusu, D., Fortuna, B., & Mladenic, D. (2011). Automatically Annotating Text with Linked Open Data. LDOW, 813.

[23] Open Knowledge Foundation. Open Data Handbook Documentation. (2012). Retrieved from http://www.edinburgh.gov.uk/download/downloads/id/3392/open_data_handbook.pdf.

[24] Rosenbaum, D. (2017, May 17). Smart cities: Who owns the data?. Retrieved from https://insights.hpe.com/articles/smart-cities-who-owns-the-data-1705.html.

[25] Anthopoulos, L. G. (2017). The Rise of the Smart City. In Understanding Smart Cities: A Tool for Smart Government or an Industrial Trick? (pp. 5-45). Springer International Publishing.

[26] Citron, R. (2016, May 17). While Big Data Grabs Headlines, Small Data Is What Cities Need. Retrieved from https://www.navigantresearch.com/blog/while-big-data-grabs-headlines-small-data-is-what-cities-need.

[27] González, A., Villazón-Terrazas, B., & Gómez, J. M. (2014, October). A linked data lifecycle for smart cities in Spain. In Proceedings of the Fifth International Conference on Semantics for Smarter Cities-Volume 1280 (pp. 9-14). CEUR-WS. org.

[28] Song, H., Basanta-Val, P., Steed, A., & Jo, M. (2017). Next-generation big data analytics: State of the art, challenges, and future research topics. IEEE Transactions on Industrial Informatics.