





Constructing Knowledge Graphs from Data Catalogues

Adegboyega Ojo¹  and Oladipupo Sennaiké² 

¹ Insight Centre for Data Analytics, Data Science Institute, NUI Galway, Galway, Republic of Ireland

adegboyega.ojo@nuigalway.ie

² Department of Computer Sciences, Faculty of Science, University of Lagos, Lagos, Nigeria

Abstract. We have witnessed about a decade's effort in opening up government institutions around the world by making data about their services, performance and programmes publicly available on open data portals. While these efforts have yielded some economic and social value particularly in the context of city data ecosystems, there is a general acknowledgment that the promises of open data are far from being realised. A major barrier to better exploitation of open data is the difficulty in finding datasets of interests and those of high value on data portals. This article describes how the implicit relatedness and value of datasets can be revealed by generating a knowledge graph over data catalogues. Specifically, we generate a knowledge graph based on a self-organizing map (SOM) constructed from an open data catalogue. Following this, we show how the generated knowledge graph enables value characterisation based on sociometric profiles of the datasets as well as dataset recommendation.

Keywords: Open data · Knowledge graphs · Self-organising maps · Dataset recommendation · Dataset value

1 Introduction

Many government institutions around the world have publicly published data about their services, performance and programmes on open data portals. These data portals are built on a myriad of open data platforms including CKAN, DKAN, Socrata, PublishMyData, Information Workbench, Enigma, Junar and OpenDataSoft. Despite the increasing number of datasets in these data portals, there has been limited use of the data by the public. While these islands of data resources have been exploited to create some economic and social value particularly in the context of city data ecosystems [1], there is a general acknowledgement the promises of open data are far from being realised [2]. In fact, usage of and engagement around open data has been and remains poor even with mediated use through apps. This problem could be associated with a number of factors. The first includes the failures of government to advertise available datasets and benefit obtained from their use [3]. The second factor is related to how the data are published on the data portals and the limited features on the underlying open data platforms to simplify access and consumption of data by ordinary users.

Current generation of open data platforms essentially provide basic dataset search capabilities and features for filtering search results. These platforms do not provide capabilities for discovering related or important datasets. Without prior knowledge of what to search for, a typical user finds it very difficult to get any meaningful information out of these data portals. Users have no way of discovering how datasets are related or what other datasets could be of interest or potentially valuable to them. Prototypes of next-generation open data platforms are beginning to emerge with features to support the recommendation of datasets [4], social engagement around data [5, 6] and automatic extraction of facts from datasets in form data stories that are more meaningful for users [7].

Some of the recent ideas in unlocking the knowledge embedded within the vast amount of data on open data portals include the use of knowledge graphs [8]. Knowledge graphs which were popularised by Google in 2012 and now increasingly available in different forms [9] have enabled richer information search experience on the web. They allow entities in different domains to be described along with their interrelations in the form of a graph [10].

In this paper, we show how knowledge graphs could be constructed from open data catalogs to reveal latent relationships (including relatedness) among datasets and also the inherent values of these datasets based on their sociometric profiles. Our approach comprises two basic steps. The first step involves computing dataset relatedness on a Self-organising map (SOM) constructed in [4]. The second step entails the transformation of the SOM to a knowledge graph using the topological distances between datasets on the map. The resulting SOM-based knowledge graph enables the discovery of clusters and themes in datasets, enables the discovery of interesting datasets and enhances the recommendation of related datasets.

2 Knowledge Graphs

There are several definitions of Knowledge Graphs (KG). It may be defined as an object for describing entities and their interrelations, by means of a graph which are usually large in which arbitrary entities may be interrelated, thus covering various topical domains [8, 10]. One of the most detailed characterisation of KG is provided by the participants of the Dagstuhl Seminar 18371, Sept. 2018 [10]. The collective understanding of a KG is: “(1) a graph-structured knowledge-base, (2) any dataset that can be seen through the lens of a graph perspective, (3) something that combines data and semantics in a graph structure, (4) structured data organisation with formal semantics and labels that is computationally feasible and cognitively plausible, (5) defined by examples such as Bablenet, OpenCyc, DBpedia, Yago, Wikidata, NELL and their shared features”. A very similar definition is provided in [9] and also indicates that KG defines possible classes and relations of entities in a schema.

Since the popularization of knowledge graphs by Google in 2012, major companies including AirBnB, Facebook, Microsoft, Yahoo!, Elsevier and Ebay have adopted this idea and developed their own variant [10]. These industry variants all employ graph abstraction as the underlying data structure. Other examples of knowledge graphs in academic literature include a knowledge graph of connected things [11], product knowledge graph to support sales assistants [12], open research KG [13].

However, knowledge graphs are distinguished from conventional web-based publishing schemes such as linked data [14]. Specifically, some contributors to [10] argue that KGs are products of collaborative efforts and brings together techniques from scientific disciplines such as Knowledge Representation, Machine Learning, Semantic Web, Databases, Natural Language Processing, Multimedia Processing and Information Extraction, amongst others [10].

Similar to our approach in this work to KG development, many works report on automatically building knowledge graphs out of textual medical knowledge and medical records [14]. Lastly, Knowledge graphs need to be able to evolve and capture the changes made to the knowledge it contains [10].

3 Self-organizing Map (SOM)-Based Dataset Relatedness

3.1 Self-organising Maps

The self-organising map is an unsupervised artificial neural network proposed by Kohonen [15] that projects high dimensional data to two or three-dimensional map while preserving the topological order in the data. The map consists of an array of units or nodes arranged in a regular rectangular or hexagonal grid. Each node has an associated n -dimensional model vector $\mathbf{m}_k = [m_{k1}, \dots, m_{kn}] \in \mathbb{R}^n$ that approximates the set of input data, where n is the dimension of the input space. The SOM is trained by iteratively presenting the input data to the nodes in parallel with a winning node emerging based on some distance metric, usually the Euclidean distance metric. The model vectors of the best matching node and its neighbors are adjusted to better match the input data.

$$\mathbf{m}_k(t+1) = \mathbf{m}_k(t) + h_{c(x),k}(t)[\mathbf{x}(t) - \mathbf{m}_k(t)], \quad (1)$$

where t is a time step and $h_{c(x),k}(t)$ is the neighborhood function [16], and

$$c(x) = \arg \min_k \{\|\mathbf{x} - \mathbf{m}_k\|\}, \quad (2)$$

is the best matching unit.

The neighborhood function is usually a Gaussian function

$$h_{c(x),k}(t) = \alpha(t) \exp\left(-\frac{\|r_k - r_{c(x)}\|^2}{2\sigma^2(t)}\right), \quad (3)$$

where $0 < \alpha(t) < 1$ is the learning-rate, $r_k \in \mathbb{R}^2$ and $r_{c(x)} \in \mathbb{R}^2$ are vectorial locations on the display grid, and $\sigma(t)$ corresponds to the width of the neighborhood function. Both $\alpha(t)$ and $\sigma(t)$ decreases monotonically with the time steps.

Some applications of SOM include Image Compression, Image Segmentation, Density Modeling, Gene Expression Analysis and Text Mining [17].

3.2 Dataset Relatedness

Relatedness defines an established or discoverable connection or association between two concepts. The study of relatedness spans a number of domains including genetics [18], management [19], computational linguistics [20], etc. Of particular interest to us is semantic relatedness, which considers how much two concepts are associated via any type of relationship between them [20]. Semantic relatedness is used in word sense disambiguation [21], information extraction [22], biomedical informatics [23], etc. Semantic relatedness goes beyond semantic similarities because it explores other kinds of relationships (beyond hyponymy/hyperonymy) between concepts.

A number of approaches have been used to measure semantic relatedness between concepts. In [20], Budanitsky and Hirst gave an overview of lexical resource-based approaches to measuring semantic relatedness. Other approaches include Latent Semantic Analysis (LSA) [24], Extended Semantic Analysis (ESA) [25], Title vector Extended Semantic Analysis (ESA) [26].

Two datasets are related if they share some concepts in common. Dataset relatedness is a measure of the proportion of shared concepts between two datasets in a catalog. Two datasets are related if they are associated by a shared concept. An attempt can be made to explicitly relate two datasets by assigning them the same theme or tagging them with the same keywords. However, explicit methods can sometimes be subjective and incomplete. In a number of cases, these tags or themes are absent.

3.3 Computing Dataset Relatedness Using SOM

The SOM is an ordered map, thus nodes close on the map are more similar than nodes further away and by extension, datasets that report to them. We used this ordering property of the SOM as the basis of computing dataset relatedness. A node A is in the neighborhood of another node B if nodes A and B are adjacent to the SOM grid. A dataset X is related to another dataset Y if the node that Y belongs to on the SOM is in the neighborhood of the node that X belongs to. The degree of relatedness is defined as the neighborhood size, thus, a degree of zero means that the datasets are in the same node while a degree of 1 means that the datasets are in the node under consideration and nodes that are immediate neighbors.

The SOM was used to cluster the Dublin City Council (DubLinked)¹ instance of the CKAN platform. The data portal contains 205 datasets for the Dublin region (www.dublinked.ie). Metadata for these datasets, along with the field names of underlying data were extracted and saved in a csv file. The data was transformed into a term frequency-inverse document frequency (tf-idf) matrix after removing stop words from the documents and stemming. The resulting matrix was a 205 by 1026 matrix, with 205 datasets and 1026 terms which serves as input to a 20 by 20 SOM. The resulting SOM is shown in Fig. 1.

From the generated map, related datasets can easily be determined based on a specified neighborhood radius. Increasing the radius increases the number of related datasets. Determining the optimal radius requires experimentation like a typical model selection problem.

¹ <http://dublinked.ie/>.

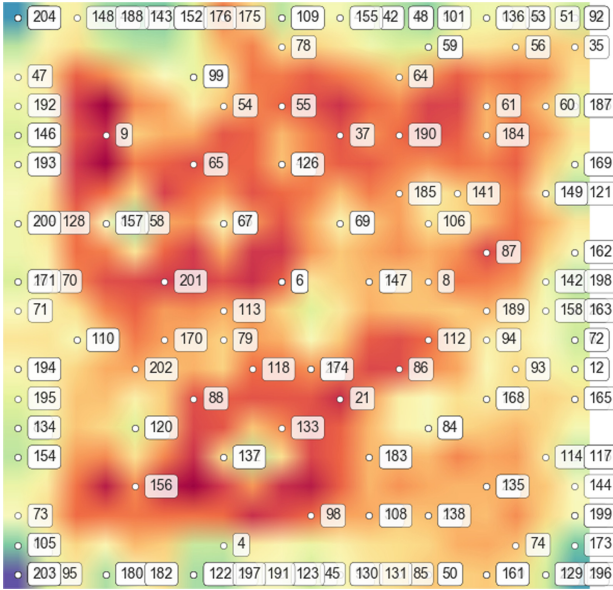


Fig. 1. The SOM

This SOM-based relatedness procedure has already been implemented as extensions to instances of the CKAN data platform with very good results [6]. The number of datasets returned is based on the degree of relatedness specified. When a high degree of relatedness is specified, datasets that are members of the same node with the dataset of interest are returned. However, when the relatedness is relaxed, datasets associated with neighbouring nodes (within a given radius) on our SOM map are also included. The model has been extended to the Dutch Language with equally good results. It is also been used as a basis for identifying datasets that can be merged [6].

4 Generating the SOM-Based Knowledge Graph

4.1 Graph Schema

The graph schema is designed to capture salient relationships among datasets in an open data catalogue - a collection of datasets descriptions or metadata. We highlight the set of important properties and relations of datasets that we consider in our graph schema specification as follows. One or more resources (in a variety of formats) are attached to each dataset in the catalogue. A dataset is associated with one or more themes usually specified by the its publisher. These themes are either associated explicitly or derived from the dataset. A dataset may also be associated with a set of *entities* (names, places, organisations, etc.). A dataset can be used in queries or visual artefacts, which are saved. Facts can be produced from a dataset. A dataset can be derived from one or more dataset, for example, a dataset can be split to form two or more datasets, or two datasets merged to form a new dataset. A dataset can also be related to another dataset. Datasets are used in sessions. Our proposed graph schema for our knowledge graph is presented in Fig. 2.

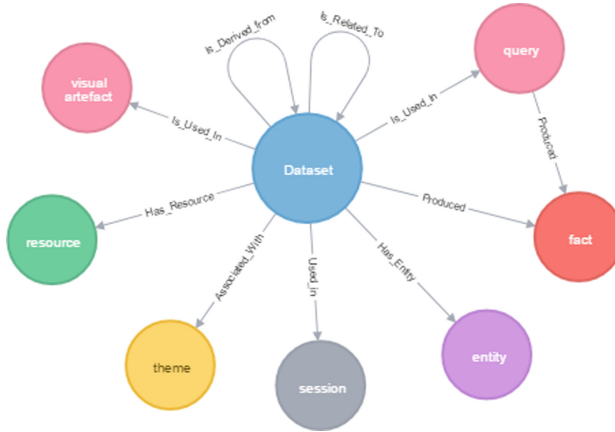


Fig. 2. The open data knowledge graph schema

4.2 Generating the Graph

The knowledge graph is generated from the SOM by eliciting the relatedness among the datasets and modeling it on a graph. The graph creation is broken down into three phases. In phase one, an initial graph is created strictly based on data relatedness information elicited from the SOM. In phase two, the graph is augmented to include relations between pairs of datasets that are very similar based on some distance metric (the Euclidean distance metric). In phase three, the graph is pruned to remove relationships between pairs of nodes with a distance exceeding a threshold.

For our experiment, we focused only on the dataset and the *is_related_to* relationship in Fig. 2. We chose a degree of 1 for our dataset relatedness, thus two datasets are related if their best matching nodes are in a neighborhood of radius 1 on the SOM. Each dataset in the catalog is represented by a node while the edges represent the relatedness between a pair of nodes. We had 205 nodes and 956 edges. Each node is labelled with the serial number of the dataset as used in the SOM from 0 to 204. Properties of the node include the title and the extracted features. Each edge connecting a pair of related dataset and is labelled “RELATED_TO” and has the following properties: the distance between the datasets, and the common terms between the datasets. Figure 3 shows an example graph with three nodes and six relationships. Table 1 summarises the parameters used in the experiment. The complete graph is presented in Fig. 4.

4.3 Structural Properties of the Graph

To analyse the generated KG, we compute some sociometric measures for each node of the knowledge graph. Specifically, we compute the degree of centrality, betweenness centrality and closeness centrality. The results of our analyses are below.

Degree of Centrality - assigns an importance score based purely on the number of links held by each node. Table 2 shows the top ten datasets in terms of degree centrality while Fig. 5 shows the subgraph containing the dataset with the highest degree centrality labelled 155.



Fig. 3. Sample graph

Table 1. Summary of parameters

SOM size	20 by 20
Inclusion threshold	1.15
Exclusion threshold	1.4
No of graph nodes	205
Number of relationships	956

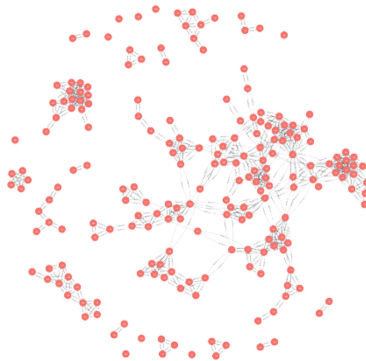


Fig. 4. Graph generated from SOM

Betweenness Centrality - measures the number of times a node lies on the shortest path between other nodes. Table 3 shows the top ten datasets in terms of betweenness centrality while Fig. 6 shows the subgraph containing the dataset with the highest betweenness centrality labelled 75.

Closeness Centrality - This measure scores each node based on their ‘closeness’ to all other nodes within the network. This measure calculates the shortest paths between all nodes, then assigns each node a score based on its sum of shortest paths.

Table 4 shows the top ten datasets in terms of closeness centrality while Fig. 7 shows the subgraph containing the dataset with the highest closeness centrality labelled 56.

Table 2. Datasets with highest Degree Centrality

Sn	Label	Degree	Title
1	155	14	Planning Register
2	52	13	DLR Goatstown Local Area Plan
3	150	13	Parking Meters location tariffs and zones in Dublin City
4	56	13	DLR Martello Towers - Location & Gun Range
5	40	12	Dublin City Council Development Planning
6	39	12	Development Planning
7	14	11	DLR - Blackrock LAP
8	48	11	DLR Cherrywood SDZ
9	38	11	Dun Laoghaire-Rathdown Development Planning
10	26	11	Citizens Information Centres

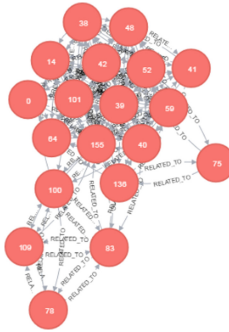


Fig. 5. Subgraph for dataset with highest degree centrality

Table 3. Datasets with highest Betweenness Centrality

Sn	Label	Betweenness	Title
1	75	2179.299417	Dublin City Council Administrative Area Maps
2	23	1802.105915	Cemeteries
3	56	1461.044824	DLR Martello Towers - Location & Gun Range
4	73	1383.155299	Dublin City Centre Litter Bin Survey
5	169	1325.938235	National Transport Authority Public Transport Information
6	90	1301.547113	Roads and Streets in Dublin City
7	51	1220.452887	DLR Cycle Counter Data
8	0	974.8005196	2010–2016 Amenities Areas
9	136	914.2101601	DLR Local Electoral Areas
10	52	878.0981247	DLR Goatstown Local Area Plan

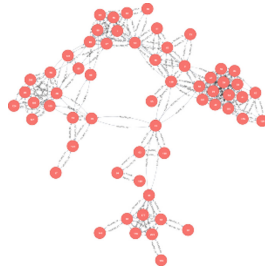


Fig. 6. Subgraph for dataset with highest betweenness centrality

Table 4. Datasets with highest Closeness Centrality

Sn	Label	Closeness	Title
1	56	0.013577496	DLR Martello Towers - Location & GunRange
2	75	0.013570882	Dublin City Council Administrative Area Maps
3	53	0.013570882	DLR Ice cream vending permits
4	23	0.013569937	Cemeteries
5	136	0.013568049	DLR Local Electoral Areas
6	0	0.013565217	2010–2016 Amenities Areas
7	73	0.013554845	Dublin City Centre Litter Bin Survey
8	52	0.013550136	DLR Goatstown Local Area Plan
9	46	0.013539786	DLR Casual Trading Locations
10	10	0.01353133	Arts Centres

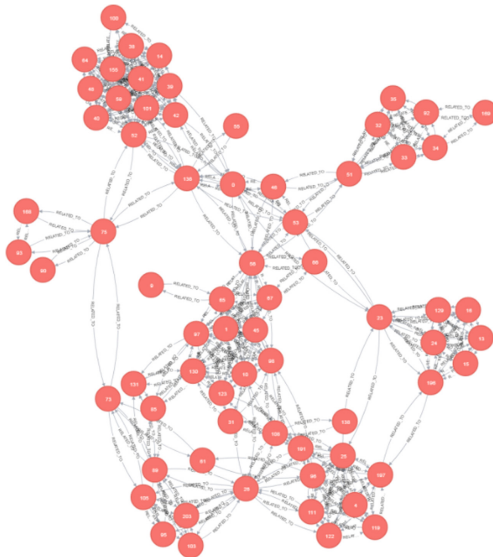


Fig. 7. Subgraph for dataset with highest closeness centrality

Clusters - A total of 32 clusters were obtained from the graph. An example cluster shown in Fig. 8 contains the labels 45, 56, 123, 10, 31, 130, 97, 67, 66, 98, 131, 1, 65, 85 and 9. The corresponding titles of these datasets are ‘DLR Arts Venues’, ‘DLR Martello Towers - Location & Gun Range’, ‘Heritage Venues’, ‘Arts Centres’, ‘South Dublin Council Offices’, ‘Libraries’, ‘DLR Libraries’, ‘DLR WW1 Hospitals’, ‘DLR War Memorials’, ‘DLR Offices and Depots’, ‘Locations of Libraries and Mobile Libraries in Fingal’, ‘ACA Boundaries’, ‘DLR Sculpture Trail Map’, ‘Dublin City Councils Libraries November Adult Fiction Issues & Renewals List’, ‘Art in the Parks - A Guide to Sculpture in Dublin City Council Parks’. This cluster coherently contains datasets that are strongly related to “Culture” including arts, heritage and leisure.

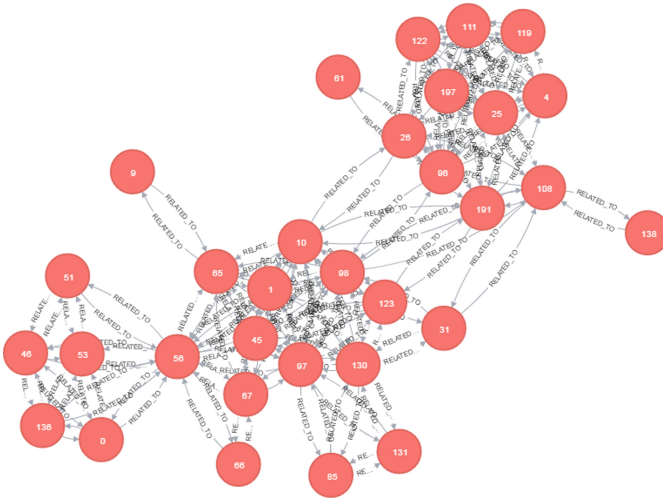


Fig. 8. Example cluster

5 Using the Generated Knowledge Graph

5.1 Dataset Centrality as an Indication of Its Value

We discuss here how the sociometric metrics above can inform the inherent value of datasets. Our notion of value here is related to the extent to which the dataset is related to or can be combined with other datasets. The dataset with the highest degree centrality has the highest number of dataset it is related to. Most of the datasets with the highest degree centrality in Table 2 are in the same cluster, so this is not very useful when considering the entire dataset. However, when the different clusters are considered, the dataset with highest degree centrality for each cluster serves as entry point to the different clusters through which majority of other datasets can be reached. These datasets can be recommended to users as entry points for their search and their exploration.

The datasets with the highest betweenness centrality are datasets that provide a bridge for two apparently different concepts. From our experiment, the dataset labelled 75, Dublin City Council Administrative Area Maps has the highest betweenness degree. It bridges dataset on:

- *Roads* - Road Collisions, Roads Maintenance Annual Works Programme, Winter Salting Routes, DLR Road Sweeping Schedule,
- *Bin locations* - Dublin City Centre Litter Bin Survey, DLR Refuse Bins Locations, Enterprise Centres,
- *Plans* - DLR Goatstown Local Area Plan, DLR Documentation for Local Area Plans, DLR Cherrywood SDZ, DLR Proposed Plan Areas,
- *Locations of amenities* - DLR Local Electoral Areas, 2010–2016 Amenities Area, DLR Ice cream vending permits, DLR Casual Trading Locations.

The datasets with the highest closeness centrality is the DLR Martello Towers - Location & Gun Range dataset labelled 56. This dataset is closely related to datasets on arts and heritage on one hand, and locations of amenities on the other.

5.2 Graph Segment Membership as a Basis for Recommendation

The second area of application for the developed knowledge graph is in the recommendation of related datasets to end-users. Based on our knowledge graph design, three types of recommendations could be potentially supported - content-based recommendation, collaborative recommendation and hybrid approaches [27, 28]. A content-based recommendation entails recommending datasets that are similar to a dataset being explored. In cases where user profiles are available, recommendations could consider datasets explored in the past. However given that most data platforms are explored or used anonymously (this is also true for our case), content-based recommendations will only consider current activities. For collaborative recommendation, users are recommended datasets that have been explored in the past by the same category of users. Hybrid collaboration combines both content and collaborative filtering methods. Against this background, we proceed to describe how the content-based recommendation works in our case.

In our content-based recommendation, we note two possible strategies. The first entails cluster membership-based recommendation and the second is based on artefacts shared by datasets including visual artefacts, queries and entities. For the cluster-based recommendations, members of cluster of the reference dataset are offered as recommendations to users. For instance, suppose the reference dataset d^* is titled “*Multi-story car parking space availability*” (see Table 5 below) belonging to Cluster 12 of the 32 clusters identified in Sect. 4. The set of recommendations for this dataset are the other datasets in *cluster 12* e.g. *Accessible Parking Places*, *DCC Derelict Site Register*, *Disabled Parking spaces*, etc. Cluster membership relations are implemented through the *is_related_to* relation on the graph. Therefore, the recommendations for reference dataset d^* , is simply the result of the query to produce the other members of the set/cluster in which d^* belongs.

Table 5. Datasets in cluster 12 of 32

- Accessible Parking Places
- DCC Derelict Site Register
- Disabled Parking Spaces
- DLR Commercial Parking Locations Numbers and Charges
- DLR County Council Parking Meters
- DLR Landscape Maintenance & Additional Sites
- DLR Parking Tag Information
- Dublin City Council Clamping Appeals
- **Multi Story Car Parking Space Availability**
- On Street Disabled Parking Bay in Dublin City Council area
- Parking Meters location tariffs and zones in Dublin City
- Parking Tag Weekly Reports
- Parks
- Play Areas
- Residential Parking permits for Dublin City Council Area
- Residential Permit Parking Area in Dublin City Council
- Suspension of Parking Bays in Dublin City Council Area

In the case of shared artefacts, recommendations for the reference dataset d^* is simply obtained by combining the results from querying the graph for the sets $is_used_in(d^*)$ and $has_entity(d^*)$.

6 Conclusions

We have shown how a Knowledge graph could be constructed from a self-organising map guided by the KG schema. Albeit, our KB could potentially include all the relationships shown in Fig. 2, we have only chosen for the purpose of illustration, the relatedness relation. In some other work, we have laid the foundation for writing facts generated from datasets directly into the KB in terms of data stories [7]. What we have shown in particular is how our KG can be generated based on a single relationship between the datasets. Since there is yet no specific prescribed procedure in literature for developing a KG [10], our approach is simply one of the possible ways to construct one. In terms of the quality of the results from our experimentations, we note that the use of knowledge graphs constitutes a very promising approach for discovering some of the latent values and similar datasets. Richer KG can be generated automatically by considering all the relationships shown in Fig. 2.

Ultimately, our vision is that the islands of open data portals on the web be meaningfully connected into a web-scale knowledge graph to truly open up open data. However, there are several research challenges that must be tackled to fully harness knowledge

graphs. These challenge among others include: how to efficiently manage the evolution of the knowledge graphs, providing explanations for information retrieved from knowledge graph, and how to effectively evaluate KGs.

Acknowledgment. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the “European Regional Development Fund”.

References

1. Ojo, A., Curry, E., Zeleti, F.A.: A tale of open data innovations in five smart cities. In: 2015 48th Hawaii International Conference on System Sciences, pp. 2326–2335 (2015)
2. Ojo, A., et al.: Realizing the innovation potentials from open data: stakeholders’ perspectives on the desired affordances of open data environment. In: Afsarmanesh, Hamideh, Camarinha-Matos, Luis M., Lucas Soares, António (eds.) PRO-VE 2016. IAICT, vol. 480, pp. 48–59. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45390-3_5
3. Hogan, M., et al.: Governance, transparency and the collaborative design of open data collaboration platforms: understanding barriers, options, and needs BT. In: Ojo, A., Millard, J. (eds.) Government 3.0 – Next Generation Government Technology Infrastructure and Services: Roadmaps, Enabling Technologies & Challenges, pp. 299–332. Springer, Cham (2017). <https://doi.org/10.1007/978-3-319-63743-3>
4. Sennaiké, O.A., Waqar, M., Osagie, E., Hassan, I., Stasiewicz, A., Ojo, A.: Towards intelligent open data platforms, pp. 414–421, September 2017
5. Scarano, V., et al.: Fostering citizens’ participation and transparency with social tools and personalization. In: Ojo, A., Millard, J. (eds.) Government 3.0 – Next Generation Government Technology Infrastructure and Services: Roadmaps, Enabling Technologies & Challenges, pp. 197–218. Springer, Cham (2017). <https://doi.org/10.1007/978-3-319-63743-3>
6. Ojo, A., et al.: A comprehensive architecture to support Open Data access, co-creation, and Dissemination. In: ACM International Conference Proceeding Series, pp. 0–1 (2018)
7. Janowski, M., Ojo, A., Curry, E., Porwol, L.: Mediating open data consumption - identifying story patterns for linked open statistical data. In: ACM International Conference Proceeding Series, vol. Part F148155, pp. 156–163 (2019)
8. Musa Aliyu, F., Ojo, A.: Towards building a knowledge graph with open data – a roadmap. In: AFRICOMM 2017: International Conference on e-Infrastructure and e-Services for Developing Countries, pp. 157–162 (2018)
9. Noy, N., Paulheim, H.: Knowledge graph refinement: a survey of approaches and evaluation methods. *Semant. Web* **8**, 489–508 (2016)
10. Bonatti, P.A., Decker, S., Polleres, A., Presutti, V.: Knowledge graphs: new directions for knowledge representation on the semantic web. Report from Dagstuhl Seminar, vol. 8, no. 9, pp. 29–111 (2019)
11. Le-Phuoc, D., Nguyen Mau Quoc, H., Ngo Quoc, H., Tran Nhat, T., Hauswirth, M.: The Graph of Things: a step towards the Live Knowledge Graph of connected things. *J. Web Semant.* **37–38**, 25–35 (2016)
12. Halaschek-wiener, C., Kolovski, V.: Towards a sales assistant using a product knowledge graph. *Web Semant. Sci. Serv. Agents World Wide Web* **6**, 171–190 (2017)
13. Jaradeh, M.Y., et al.: Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge (2019)

14. Kejriwal, M.: What is a knowledge graph? Springer Briefs in Computer Science, pp. 1–7 (2019)
15. Kohonen, T.: The self-organizing map. Proc. IEEE **78**(9), 1464–1480 (1990)
16. Kohonen, T., et al.: Self organization of a massive text document collection. Kohonen Maps **11**(3), 171–182 (1999)
17. Yin, H.: The self-organizing maps: background, theories, extensions and applications. Stud. Comput. Intell. **115**, 715–762 (2008)
18. Weir, B.S., Anderson, A.D., Hepler, A.B.: Genetic relatedness analysis: Modern data and new challenges. Nat. Rev. Genet. **7**(10), 771–780 (2006)
19. Nocker, E., Bowen, H.P., Stadler, C., Matzler, K.: Capturing relatedness: comprehensive measures based on secondary data. Br. J. Manag. **27**, 197–213 (2016)
20. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006)
21. Patwardhan, S., Banerjee, S., Pedersen, T.: Using measures of semantic relatedness for word sense disambiguation. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 241–257. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-36456-0_24
22. Stevenson, M., Greenwood, M.A.: A semantic approach to IE pattern induction. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL 2005 (2005)
23. Zhu, Y., Yan, E., Wang, F.: Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. BMC Med. Inform. Decis. Mak. **17**, 95 (2017)
24. Lee, M.D., Pincombe, B., Welsh, M.: An empirical evaluation of models of text document similarity. In: Proceedings of the Annual Meeting of the Cognitive Science Society, pp. 1254–1259 (2005)
25. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. IJCAI Int. Jt. Conf. Artif. Intell. **7**, 1606–1611 (2007)
26. Rybinski, M., Aldana-Montes, J.F.: tESA: a distributional measure for calculating semantic relatedness. J. Biomed. Semantics **7**, 67 (2016)
27. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
28. Shambour, Q., Lu, J.: Government-to-business personalized e-services using semantic-enhanced recommender system. In: Andersen, K.N., Francesconi, E., Grönlund, Å., van Engers, T.M. (eds.) EGOVIS 2011. LNCS, vol. 6866, pp. 197–211. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22961-9_16