A Cancer Genomics Data Space within the Linked Open Data (LOD) Cloud

Durre Zehra*, Alokkumar Jha*, Yasar Khan*, Ali Hasnain*, Mathieu d'Aquin **
Ratnesh Sahay*

Data Science Institute, National University of Ireland, Galway
{*firstname.lastname, ** mathieu.daquin}@insight-centre.org

Abstract. The ongoing cancer research requires finding patterns and associations among genetic, cellular and molecular features residing in isolated and disparate repositories. The discovery of complex biological associations from these independent repositories will help advanced analysis and hypothesis generation over a network of coherent datasets. In this paper we provide a short overview of three types of cancer genomics datasets that are transformed from raw formats (csv, tsv, relational, etc.) into a set of linked datasets within the Linked Open Data Cloud. The three genomics datasets (Copy Number Variation (CNV), Methylation, & Gene Expression) are related to ovarian cancer studies and originally archived in three different repositories (The Cancer Genome Atlas (TCGA), Catalogue of Somatic Mutations in Cancer (COSMIC), and Copy Number Variation in Disease (CNVD)). Our key motivation is to create a network of coherent cancer genomic linked datasets within the widely accessible LOD cloud. We provide these three genomics datasets as a set – called Linked Open Data for Cancer Genomics (LOD4CG) – of five interlinked publicly accessible SPARQL endpoints that will help researchers and practitioners to exploring these datasets and links across them.
LOD4CG SPARQL Endpoints: https://github.com/drzehra14/LOD4CG.

1 Introduction

Cancer research is producing massive amount of data in heterogeneous data formats and in disparate repositories. It is already predicted that 2–40 exabytes of storage capacity will be needed by 2025 just for the human genomes which will continue to grow approximately 40 petabytes of additional genomic information each year [26]. Therefore, the heterogeneous nature of these data and their widespread distribution over numerous databases makes searching and pattern discovery a tedious and cumbersome task[11,9]. From a researcher perspective, a network of coherent and well-interlinked datasets, opens the possibilities of advanced search and analysis across such datasets sources in order to identify novel and meaningful correlations and mechanisms as explained by Hasnain et al[10].

In the recent years, there is a growing interest and adoption of open data infrastructures such as Linked Open Data (LOD)[3,1] by researchers particularly from the Health-care and Life Sciences (HCLS) domain. How to exploit open data infrastructures has become an important research agenda in the open science community. Our work is motivated by the needs of the BIOOPENER[1] project which aims to link cancer and

[1] http://bioopenerproject.insight-centre.org/

bio-medical data repositories by providing interlinking and querying mechanisms to understand cancer progression [16].

In this paper, we present a short overview of three types of cancer genomics datasets (Copy Number Variation (CNV), Methylation, & Gene Expression) and links among them, which are newly included in the Linked Open Data (LOD) Cloud. These genomics datasets are originally archived at three independent repositories (COSMIC[2], TCGA[3] & CNVD[4]) and we transformed them into a set of five interlinked publicly accessible SPARQL endpoints. The proposed LOD4CG aims to support Life Science's researchers in the exploration of cancer related data and links among different resources. We start the paper by presenting some related works of publishing bio-medical and health-care datasets with the LOD Cloud. We then present a motivating scenario on how well-interlinked datasets could help researcher in finding novel associations among biological entities (gene, protein, pathways, etc.). Finally, we then present the details of LOD4CG datasets and links among them.

2 Related Work

Jentzsch et al. [15] discuss the importance of linking open drug data for pharmaceutical research and development.Minarro-Gimenez et al. [2] introduced an extension of the OGO Knowledge Base with the OGOLOD system, having orthologs/diseases information using Linked Data. Saleem et al. [24] transformed TCGA data to RDF and linked it to elements of the LOD cloud, creating the Linked Cancer Genome Atlas dataset. Later, the authors also integrated publications from PubMed with the Linked Cancer Genome Atlas dataset [23]. Koide et al. [17] RDFize the Japanese WordNet and linking to the Japanse DBpedia as Linguistic LOD. McCrae et al. [20] defines the importance of Linguistic Linked Open Data cloud, and created LOD (sub-)cloud of linguistic resources, which covers various linguistic databases, lexicons, corpora, terminologies, and metadata repositories. The deployment of Linked Open Government Data is explained by Li Ding et al. [6]. In order to promote easy data access, reusability, extraction and analysis Bukhari et al. [5] transform Canadian health census data to LOD. Hasnain et al. presented biomedical resources ontologies, repositories, and other data resources relevant in the context of Drug Discovery and Cancer Chemoprevention[14].

The summary about Linked Data-driven solution in different domains is given in Table 1.

Table 1: Overview of the Related work in different Domains

Publications	LOD	Domain
Minarro-Gimenez et al. [2]	OGOLOD	Orthologs/Diseases
Saleem et al. [24]	LinkedTCGA	Cancer
Saleem et al.[23]	LinkedTCGA	Publications
Koide et al. [17]	JLLOD	Japanese WordNet
P. McCrae et al. [20]	LLO	Linguistic
Li Ding et al. [6]	LOGD	Government
Bukhari et al. [5]	CHCLOD	Census
Jentzsch et al. [15]	LODD	Drug

[2] http://cancer.sanger.ac.uk/cosmic
[3] https://cancergenome.nih.gov/
[4] http://202.97.205.78/CNVD/

3 Motivating Scenario

Our work is motivated by the need of interlinking cancer genomics data resources with other bio-medical resources already available in the LOD cloud. The following section describes a scenario where having several cancer databases linked could facilitate the analysis of data by a researcher. For instance, if a bio-medical expert aims to mine information about the KRAS gene – across Web – which is one of the most frequently mutated genes in human cancers [27], suggesting that targeting one gene may not be sufficient to fully inhibit KRAS-driven oncogenesis[18].

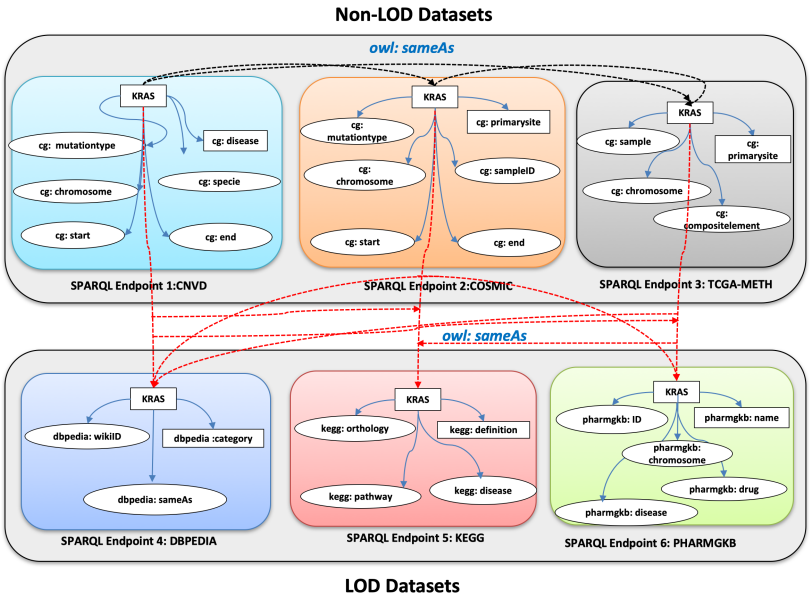

Fig. 1: KRAS Gene links across LOD4CG and LOD Datasets. Black dotted lines shows LOD4CG *owl:sameAS* links with in LOD4CG whereas Red dotted lines shows LOD4CG and LOD *owl:sameAS* links respectively

Consider a scenario in which datasets are hosted at six different SPARQL endpoints as shown in Figure 1. The source gene *KRAS* is located in the SPARQL endpoint 1 with disease, chromosome, mutation type, specie and cnv start and end locations information, whereas primary site, sample-ID, composite element, wiki-D, orthology, pathway, and drug, information is distributed across the SPARQL endpoints 2,3,4,5 and 6 respectively. Exploring widespread distribution of biomedical datasets over numerous databases makes searching and pattern discovery a tedious and manual task. However, linking across these six datasets will make it feasible to search across open data sources. By linking *KRAS* gene as *owl:sameAs*, we might find novel information about genes and their distributed properties i.e. diseases, drug, histology etc accross multiple datasets and reveal interesting opportunities for biomedical experts to pursue. We belive that healthcare research data level, specially opening up biomedical data, sharing and linking large healthcare datasets enables semantically to relate and enrich data. It enables

more efficient semantic access to the evidence base on symptoms, diseases, diagnosis, and treatments, grounds offering the potential for improvements in individuals and populations care. The Linking resuts are discussed in Section 6.

4 Datasets

The LOD4CG includes five cancer genomics datasets that are loaded into 5 different SPARQL endpoints discussed in following section.

4.1 LOD4CG Datasets

In this work we target those cancer genomics datasets, which are not currently part of LOD. In this work, we have used Cosmic which is a comprehensive database for exploring somatic mutations in key cancer genes across different cancer samples, CNVD which is a database that aggregates data from publicly available literature related to CNV that has been published in recent years and TCGA which is a catalogue of the genomic alternations found in all cancers. From Cosmic and CNVD we have used Copy Number Variations (CNV) data type for all cancer types whereas from TCGA we have used three data types from the Ovarian Serous Cystadenocarcinoma (OV) disease i.e Copy Number Variation (CNV), Gene Expression (GE) and Methylation (METH) respectively. The data cumulatively is around 5.2196 GB. Briefly discussed these repositories below and Table 2 shows the number of size, triples, subjects, predicates, and objects in each dataset.

COSMIC [5] is a comprehensive database for exploring somatic mutations in key cancer genes across different cancer samples. COSMIC gives open access to the 1,343,214 tumour samples, with 1,180,789 copy number variations. It combines genome-wide sequencing results from 32,514 tumours, with complete manual curation of 25,501 individual cancer publications [8]. In this work, we have used COSMIC's Copy Number Variation (CNV) data for all cancer types.

TCGA [6] is a publicly funded project created by the National Cancer Institute and the National Human Genome Research Institute in 2005 . This project aims to catalogue the genomic alternations found in all cancers. The TCGA data portal stores 2 PB of open access cancer patient data, having 310,859 text archives for 33 different cancer types and 11,000 patients. Each disease data is categorized into tumour type (i.e. ovarian, breast) and data types (i.e. mutation, gene expression). In this work, we have used three different data types i.e. copy number variation. methylation and gene expression from the Ovarian Serous Cystadenocarcinoma (OV) disease and deployed on three different SPARQL endpoints.

CNVD [7] is a database that aggregates data from publicly available literature related to CNV that has been published in recent years. CNVD contains information on more than 500 diseases and includes different tumour types. A majority of the results documented

[5] https://cancer.sanger.ac.uk/cosmic
[6] https://cancergenome.nih.gov/
[7] http://202.97.205.78/CNVD/

in this database was derived from reliable CNV detection experiments. More than 28% of the disease data (from 22 species) in the CNVD data portal is related to neoplasms [22]. The most common tumour types described in CNVD are breast cancer, prostate cancer, lung cancer, gastric cancer and ovarian cancer. In this work, we have used CNVD's CNV data for all cancer types.

Table 2: Overview of LOD4CG Datasets

Dataset	Size (GB)	Triples	Subjects	Predicates	Objects
COSMIC	3.84	36537626	893293	14	2056753
CNVD	0.0346	1552025	194590	09	512307
TCGA-OV-CNV	0.212	11940187	1860004	06	3632669
TCGA-OV-METH	0.564	137540670	6920165	08	32212038
TCGA-OV-GE	0.573	54249644	20604	06	15489284
Total	5.2196	241817452	19016179	43	53903051

4.2 LOD Datasets

Bio2RDF[8] currently provides the largest network of Linked Data for the Life Sciences. We use three Bio2RDF datasets KEGG, PharmGKB and GOA. We have downloaded Bio2RDF datasets and deploy them locally to increase the reliability of the querying system. Whereas we also use live SPARQL endpoint of Dbpedia which is considered as a central hub of LOD.

5 Methodology

The transformation of COSMIC, CNVD and TCGA data has three main steps:

1. Retrieving data in the free-text format;
2. Annotating and transforming data of different cancer diseases/data types to RDF using standard vocabularies and deploying it to SPARQL endpoints and;
3. The discovery of quality links between LOD4CG datasets as well as across LOD datasets,

5.1 Data Transformation

Data from TCGA, COSMIC and CNVD first get annotated for the ease of transformation, which was based on the Semantic Science Integrated Ontology (SIO) [7].However, SIO is primarily and upper-level ontology, i.e. describes high level concepts of the domain. Therefore, in this work, we use and extend this ontology to fulfil our annotation needs. To maximize the reuse of existing terms, we use the MIREOT guidelines [21] to import single classes from the the National Cancer Institute Thesaurus (NCIT) [25] and the Experimental Factor (EFO) [19]. Both the imported and the newly created classes/properties were integrated into the SIO structure. The extended ontology was called Cancer Genomics. Afterwards this raw data is RDFized and deployed to various SPARQL Endpoints for further experiments.

[8] http://bio2rdf.org/

5.2 Link Discovery

One of the best practices for creating LD includes linking it to different sources [4]. The creation of links between the CNVD, COSMIC and TCGA datasets was essential to guarantee that the information contained in these datasets is publicly available, allows federated SPARQL queries [13,12,29], facilitates data integration and data analytics, and is linked to the LOD cloud. We used the SILK framework [28] to discover links between the CNVD, COSMIC and TCGA knowledge bases. The SILK framework is a flexible link discovery tool that provides time efficient link discovery between entities within different Web data sources. The framework uses a declarative language for specifying which types of RDF links should be discovered between data sources, as well as which conditions entities must fulfil in order to be interlinked. As genes, chromosomes and disease have unique identifiers used across several bio-medical knowledge bases, we used SILK's *owl:sameAs* measure for linking the identifiers.

6 Results and Discussion

The experimental setup comprises of two parts; (i) finding links between the LOD4CG datasets; and (ii) finding links across the LOD4CG and LOD datasets. Table 3 shows

Table 3: Overall link Statistics.

Source	Target	Entity	Links	Total
COSMIC	CNVD	Gene	18,068	48,826
	TCGA-OV-GE		16,219	
	TCGA-OV-METH		13,935	
	Dbpedia		521	
	PharmgKB		83	
	Dbpedia	Chromosome	1188	1,236
	TCGA-OV-CNV		24	
	TCGA-OV-METH		24	
CNVD	TCGA-OV-GE	Gene	17,712	32,933
	TCGA-OV-METH		14,517	
	Kegg		38	
	Dbpedia		544	
	PharmgKB		82	
	GOA		40	
	Dbpedia	Disease	100	180
	PharmgKB		68	
	Kegg		12	
	Dbpedia	Chromosome	1,188	1,238
	TCGA-OV-CNV		25	
	TCGA-OV-METH		25	
TCGA-OV-METH	Dbpedia	Gene	427	482
	PharmgKB		40	
	Kegg		15	
TCGA-OV-GE	TCGA-OV-METH	Gene	13,500	14,114
	Dbpedia		500	
	PharmgKB		79	
	Kegg		35	
				143,109

that cumulatively discovered links are 143,109 across 8 chosen datasets. The first part of our experimant shows that cumulative discovered links between LOD4CG datasets are 94,049 which includes: (i) links of *gene* between COSMIC, CNVD, TCGA-OV-METH and TCGA-OV-GE, (ii) links of *chromosome* between TCGA-OV-METH, TCGA-OV-CNV, COSMIC and CNVD datasets, (iii) COSMIC and CNVD have highest number of links 18,068 linked with *gene*, and finally (iv) TCGA-OV-METH, TCGA-OV-CNV and COSMIC have least number of 24 links via instances of *chromosome*. For second

experimental setup, the total discovered links are 4960 across LOD4CG and LOD datasets, which includes: (i) Dbpedia, PharmGKB, Kegg, GOA are linked with COSMIC, CNVD, TCGA-OV-METH and TCGA-OV-GE via *gene* instances, (ii) Dbpedia is linked with COSMIC and CNVD with *gene* and *chromosome* instances, (iii) Kegg, PharmGKB and Dbpedia are linked with CNVD via *disease* instances (iv) CNVD, COSMIC and Dbpedia have 1188 links via *chromosome*, (v) TCGA-OV-METH and Kegg have least number of 15 links using *gene* instances.

KRAS Gene: Finally, we discover that KRAS gene which is highly mutated in lung cancer patients is linked via *owl:sameAs* with COSMIC, CNVD, TCGA-OV-METH, TCGA-OV-GE, Kegg and PharmGKB respectively. The Figure 1 shows our results in detail where the source gene KRAS is located in the SPARQL endpoint 1 with disease, chromosome, mutation type, specie and cnv start and end locations information, whereas the primary site, sample-ID, composite element, wiki-D, orthology, pathway, and drug, information is distributed across the SPARQL endpoints 2,3,4,5and 6 respectively.

1. COSMIC, CNVD, TCGA-METH, Kegg, Dbpedia and PharmGKB hosted at six different SPARQL endpoints.
2. The source gene KRAS are located in SPARQL endpoint 1 CNVD.
3. The LOD4CG has (*owl:sameAs*) links for KRAS starting from SPARQL endpoint 1: CNVD where the target is available at SPARQL endpoint 2:COSMIC; and SPARQL endpoint 3:TCGA-METH.
4. In LOD datasets, the target is available in the SPARQL endpoint 4:DBpedia, SPARQL endpoint 5:Kegg and SPARQL endpoint 6: PharmGKB via *owl:sameAs* link.

7 Conclusion and Future Work

In this work, we introduce LOD4CG, a new cancer genomic data space within the LOD cloud. The proposed LOD4CG aims to support Life Science's researchers in the exploration of cancer related data and links among different resources. We have introduce 5.2196 GB of data from Cosmic, CNVD and TCGA having 143,109 discovered links across LOD4CG and LOD. In future, we plan to extend LOD4CG with breast cancer data (Copy Number Variation (CNV), Methylation, & Gene Expression) from these three responsories (COSMIC, CNVD, & TCGA). We realise that links between datasets can become invalid or broken due to the changes in datasets and URIs, therefore, maintenance of links is a necessary task for LOD4CG. We plan to employ an approach that will ensure maintenance of links with evolving datasets in the LOD4CG.

8 Acknowledgement

The work presented in this research paper has been funded by Science Foundation Ireland under Grant No. SFI/12/RC/2289.

9 Availability of data and materials

The BIOOPENER online demonstration website `http://bioopenerproject.insight-centre.org/` is available for the scientific uses and the relevant datasets(in RDF) shown in the Table 2 are available at `https://github.com/drzehra14/LOD4CG`.

References

1. Bio2RDF Release 3: A larger, more connected network of Linked Data for the Life Sciences, CEUR Workshop Proceedings, vol. 1272. CEUR-WS.org, Riva del Garda, Italy (2014)
2. A Minarro-Gimenez, J., Egana-Aranguren, M., Villazón-Terrazas, B., T Fernandez-Breis, J.: Publishing orthology and diseases information in the linked open data cloud. Current bioinformatics 7(3), 255–266 (2012)
3. Bauer, F., Kaltenböck, M.: Linked open data: The essentials. Edition mono/monochrom, Vienna (2011)
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Semantic services, interoperability and web applications: emerging concepts pp. 205–227 (2009)
5. Bukhari, A.C., Baker, C.J.: The canadian health census as linked open data: towards policy making in public health. In: Data integration in the life sciences (2013)
6. Ding, L., Lebo, T., Erickson, J.S., DiFranzo, D., Williams, G.T., Li, X., Michaelis, J., Graves, A., Zheng, J.G., Shangguan, Z., et al.: TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web 9(3), 325–333 (2011)
7. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., Klassen, D., McCusker, J.P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M.D., Hoehndorf, R.: The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery. Journal of Biomedical Semantics 5(1), 14 (Mar 2014)
8. Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al.: COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic acids research 39(suppl_1), D945–D950 (2010)
9. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences LOD Cloud. In: 1st International Workshop on Ontology Engineering in a Data-driven World collocated with EKAW12 (2012)
10. Hasnain, A., Kamdar, M.R., Hasapis, P., Zeginis, D., Warren Jr, C.N., Deus, H.F., Ntalaperas, D., Tarabanis, K., Mehdi, M., Decker, S.: Linked biomedical dataspace: lessons learned integrating data for drug discovery. In: The Semantic Web–ISWC 2014, pp. 114–130. Springer (2014)
11. Hasnain, A., Mehmood, Q., e Zainab, S.S., Hogan, A.: Sportal: profiling the content of public sparql endpoints. International Journal on Semantic Web and Information Systems (IJSWIS) 12(3), 134–163 (2016)
12. Hasnain, A., Mehmood, Q., e Zainab, S.S., Hogan, A.: Cataloguing the context of public sparql endpoints. In: Innovations, Developments, and Applications of Semantic Web and Information Systems, pp. 295–328. IGI Global (2018)
13. Hasnain, A., Mehmood, Q., e Zainab, S.S., Saleem, M., Warren, C., Zehra, D., Decker, S., Rebholz-Schuhmann, D.: Biofed: federated query processing over life sciences linked open data. Journal of biomedical semantics 8(1), 13 (2017)
14. Hasnain, A., Rebholz-Schuhmann, D.: Biomedical semantic resources for drug discovery platforms. In: European Semantic Web Conference. pp. 199–218. Springer (2017)
15. Jentzsch, A., Zhao, J., Hassanzadeh, O., Cheung, K.H., Samwald, M., Andersson, B.: Linking open drug data. In: I-SEMANTICS (2009)
16. Jha, A., Khan, Y., Mehdi, M., Karim, M.R., Mehmood, Q., Zappa, A., Rebholz-Schuhmann, D., Sahay, R.: Towards precision medicine: discovering novel gynecological cancer biomarkers and pathways using linked data. J. Biomedical Semantics 8(1), 40:1–40:16 (2017)
17. Koide, S., Takeda, H.: RDFization of japanese electronic dictionaries and LOD. In: Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data. pp. 64–69 (2013)
18. Kostyrko, K., Han, K., Kelly, M., Jeng, E., Morgens, D., Bassik, M., Jackson, P., Sweet-Cordero, A.: Spot-007 identifying novel combinatorial synthetic lethal vulnerabilities in kras-driven lung cancer (2018)

19. Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., Parkinson, H.: Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 26(8), 1112–1118 (2010)
20. McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., et al.: The open linguistics working group: Developing the linguistic linked open data cloud. In: LREC (2016)
21. Mélanie, C., Frank, G., L, L.A., James, M., Daniel, S., R, B.R., Alan, R.: MIREOT: The minimum information to reference an external ontology term. Applied Ontology (1), 23–33 (2011)
22. Qiu, F., Xu, Y., Li, K., Li, Z., Liu, Y., DuanMu, H., Zhang, S., Li, Z., Chang, Z., Zhou, Y., et al.: CNVD: Text mining-based copy number variation in disease database. Human mutation 33(11) (2012)
23. Saleem, M., Kamdar, M.R., Iqbal, A., Sampath, S., Deus, H.F., Ngomo, A.C.N.: Big linked cancer data: Integrating linked TCGA and PubMed. Web Semantics: Science, Services and Agents on the World Wide Web 27, 34–41 (2014)
24. Saleem, M., Padmanabhuni, S.S., Ngomo, A.C.N., Almeida, J.S., Decker, S., Deus, H.F.: Linked cancer genome atlas database. In: Proceedings of the 9th International Conference on Semantic Systems. pp. 129–134. ACM (2013)
25. Sioutos, N., Coronado, S.d., Haber, M.W., Hartel, F.W., Shaiu, W.L., Wright, L.W.: NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics 40(1), 30–43 (2007)
26. Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E.: Big data: astronomical or genomical? PLoS biology 13(7), e1002195 (2015)
27. Torjesen, I.: Breakthrough in targeting kras gene mutation implicated in many cancers. Lung cancer 15, 05 (2018)
28. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk–a link discovery framework for the web of data (2009)
29. e Zainab, S.S., Saleem, M., Mehmood, Q., Zehra, D., Decker, S., Hasnain, A.: Fedviz: A visual interface for SPARQL queries formulation and execution. In: Proceedings of the International Workshop on Visualizations and User Interfaces for Ontologies and Linked Data co-located with 14th International Semantic Web Conference (ISWC 2015), Bethlehem, Pennsylvania, USA, October 11, 2015. p. 49 (2015), http://ceur-ws.org/Vol-1456/paper5.pdf