Life Sciences Linked Open Data Datasets Connections to SNOMED CT, RxNORM & GO

ARTEMIS CHALEPLIOGLOU, SOZON PAPAVLASOPOULOS, MARIOS POULOS
Department of Archives, Library Science and Museology
Faculty of Information Science & Informatics
Ionian University Corfu,
Ioannou Theotoki 72, Corfu 49100
GREECE
mpoulos@ionio.gr

Abstract: - Linked Open Data (LOD) in life sciences is about a movement dealing with an interconnected set of Big Data from clinical records, trials and pharmacologic interventions to next generation sequencing and proteomics. The combination of these evidence is a prerequisite for biomedical bibliometrics, conduction of systematic reviews and meta-analyses, evidence-based medicine and precision medicine. Therefore, we collected the BioPortal subdomain life sciences LOD datasets triples and links from the LOD Cloud diagram and old.datahub.io repository to explore their connections to the three top ontologies describing clinical, pharmacological and molecular biology information, SNOMED CT, RxNORM and GO respectively. We found that 96% of the datasets share links with SNOMED CT, 24% with RxNORM and 31.5% with GO. However, the datasets that share links with both RxNORM and GO are only 3.5%. Our data suggest the need of enrichment of life science LOD datasets with connections between pharmacology and molecular biology.

Key-Words: - Semantic web, Linked Data, Biomedicine, Life Sciences, Ontologies, Bibliometrics, Evidence-based medicine, Precision medicine.

1 Introduction

Modern high-throughput biomedical technologies generate enormous amounts of data and transform medicine, pharmacology and biology into Big Data fields [1]. In assisting clinical decision-making by physicians or artificial intelligence systems [2], several tools have been proposed include bibliometrics, quantitative analysis of biomedical literature [3], and the opposite but universally accepted and endorsed practices the evidence-based medicine and the personalized medicine [4], all based on life sciences Big Data analytics and synthesis.

The challenge for the library and information scientists is to deliver and describe this information, extracted from electronic bibliographic databases as well as gray literature resources, contained in the World Wide Web (web) in a meaningful manner easily exchangeable, shareable and retrievable, the semantic web, by both man and the machine [5]. The Linked Open Data is a semantic web community initiative aiming to massively publish open data to the Web identified by Uniform Resource Identifiers (URIs) with a knowledge representation language such as the Resource Description Framework (RDF)/schema or the Web Ontology Language (OWL) format, forming triplestore datasets that can be queried by Simple Protocol and RDF Query Language (SPARQL) [6, 7].

To meet the goals of biomedical bibliometrics, systematic reviews and meta-analyses, the different clinical, pharmacologic and molecular biology datasets should be connected with links that allow automatic reasoning to expose existing but currently cryptic relationships between Patients, Interventions, Comparisons and Outcomes, the PICO criteria [8]. Therefore, we explored the LOD cloud diagram, graphically depicting the linked data datasets published, its evolution and connectivity, focusing in specific on the Life Sciences Bioportal datasets subdomain, which represents the largest, active and well-structured corpus of biomedical information.

2 Experimental and Computational Details

Three sources of datasets information were used in this study the LOD Cloud (https://lod-cloud.net), old.datahub.io (https://old.datahub.io) and BioPortal

(https://bioportal.bioontology.org), last updated on August 2017. The LOD2 project methodology to produce, publish and maintain a semantic web linked dataset, the formation of structured data, the URIs as unique identifiers, and the inclusion of links to other datasets was employed in this study [9].

2.1 Image Analysis

The ImageJ software (version 1.50e) was used for the analysis of LOD Cloud diagrams. The domain datasets area, visualized as colour circles representing a subject (i.e. Geography, Government, Life Sciences), measurements were used as indicators of the domain volume (number of triples). The measurements of the integrated density of links (edges of LOD Cloud) were used as indicators of the connectivity within each domain.

2.2 Data Analytics

Descriptive statistics, linear regression, exponential regression, calculation of coefficients and data standardization of links was performed with Microsoft (MS) Excel. The links of each dataset with SNOMED CT, RxNORM and GO were used as variables for the connectivity with clinical, pharmacological and molecular biology knowledge domains. Data are presented as mean ± standard deviation.

3 Experimental Results

The LOD Cloud has been well-established over the last decade as a point of reference to the Linked Data community. It covers a wide range of human knowledge domains including Government, Geography, Linguistics, Media, Publications, Social Networking, Life Sciences and Cross-Domain, the latest enriched with Wikipedia related metadata. The LOD Cloud is consisting of linked datasets provided by the community to the LOD Cloud group, until August 2017 located to the datahub.io registry (https://datahub.io/group/lodcloud) but then the metadata were transferred to the old.datahub.io. The presenting analysis was performed on the basis of these data.

To the best of our knowledge, generally accepted criteria for the description and comparison of structured datasets are not currently available. Therefore, we proposed the use of volume, general connectivity, and special connectivity to particular nodal points datasets describing fundamental scientific fields that have to be combined for logical reasoning and the extraction of broader conclusions from specific observations. We predominantly focus in the Life Sciences domain LOD Cloud datasets to explore the efficacy of combination of clinical, pharmacologic and molecular biology in clinical decision-making, either on the basis of bibliometrics, evidence-based medicine or precision medicine.

Since 2009 until 2017 the LOD Cloud diagram total number of datasets was increased from 89 to 1146 with an approximately pace of 130 new datasets per year, while for the Life Sciences domain in particular from 28 to 336 with an approximately pace of 35 on the basis of the information from the continuous updates of the diagram by the LOD Cloud team, Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak (Figure 1).

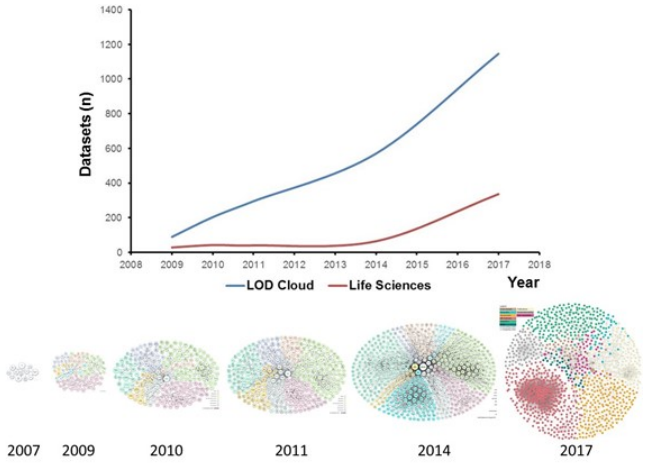

Figure 1. The growth of the LOD Cloud diagram over time. Diagrammatic representation of the number of the LOD Cloud total datasets (blue line) and LOD Cloud Life Sciences domain datasets (red line) from 2009 to 2017 (top). Evolution of the graphical view of the LOD Cloud from 2007 to 2017 (bottom).

Considering that the predominant task for Linked Data is the generation of interlinking metadata we explored the density of LOD Cloud diagram links (connectivity) until 2017, relative to the 2009 diagram (Figure 2). In 2017, the connectivity within the whole LOD Cloud datasets was increased by 2.3-fold and within the LOD Cloud Life Sciences domain datasets by 3.4-fold when compared to 2009. The LOD Cloud diagram relative connectivity was increased yearly by 27%, while the Life Sciences domain by 40%.

The rapid increase of LOD Cloud connectivity suggest an overwhelming adoption of Linked Data technologies by the publishers which continuously support the updates of existing datasets and their enrichment in new interlinking metadata.

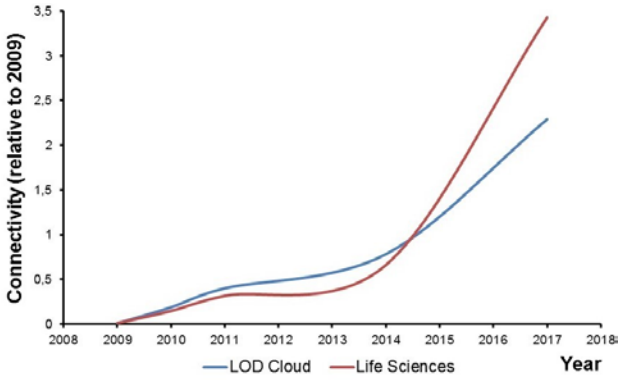

Figure 2. Evolution of the density of LOD Cloud links (connectivity) relative to the 2009 diagram. The connectivity of the whole LOD Cloud datasets (blue line) and LOD Cloud Life Sciences domain datasets (red line) are presented until 2017.

The Life Sciences domain datasets exhibit a stronger dynamics towards Linked Data generation possibly because: (a) of the introduction of next-generation omics technologies that demands automatic high-throughput analysis of metadata [9-11], (b) of the widespread adoption systems biology methods [12], (c) of the need for excessive systematic reviews and randomized controlled trials for clinical purposes [13, 14], (d) of the development of novel or reuse of existing drugs and natural products in therapeutics or disease prevention [15, 16], and (d) of the introduction of Big Data analytics into General Biology, especially into Evolution [17], Ecology [18] and Taxonomy [19]. This trend is emphatically demonstrated into the 2017 LOD Cloud diagram where it is clearly demonstrated that the Life Sciences domain overtook the diagram by producing the highest number of datasets, with the biggest volume in triples, and the strongest connectivity (Figure 3).

3.1 The BioPortal Subdomain of Life Sciences

The LOD Cloud Life Sciences domain datasets exhibit two distinct subdomains the BioPortal and Bio2RDF (Figure 3). The two subdomains follow different metadata building strategies and exhibited different number of datasets, different average volume and relative connectivity.

The Bio2RDF (http://bio2rdf.org/) datasets were developed directly from the web sources of Life Sciences relational databases, text files, XML documents, and HTML pages, by data crawling and conversion to RDF, in a mashup approach led by the Bio2RDF administrators [20]. The Bio2RDF ultimate goal was the generation of a triplestore with interlinked metadata accessible form a unique URL. However, Bio2RDF datasets were last updated by July 2014 with the publication of their third release [21]. The LOD Cloud diagram (version August 2017) data suggest that the Bio2RDF subdomain contains less datasets, with an average smaller volume and less connectivity than the BioPortal subdomain.

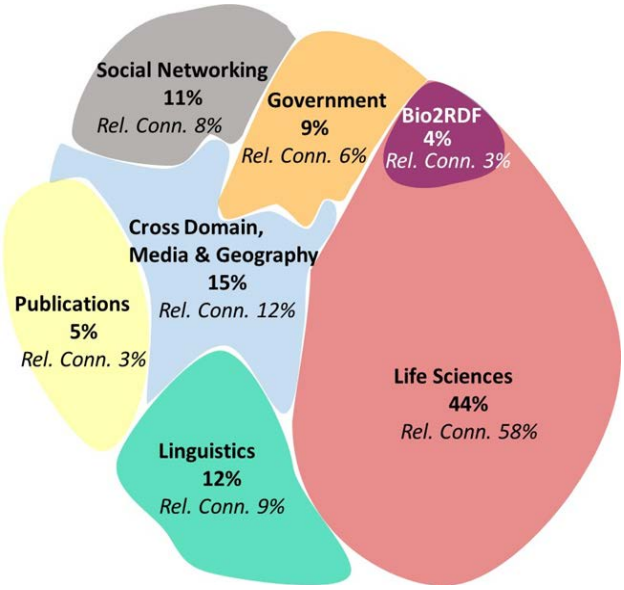

Figure 3. Diagrammatic representation of the per cent of LOD Cloud datasets per domain (in bold) with the relative connectivity per domeain (in italics). The Life Sciences domain dominates in number of datasets, volume of datasets and connectivity (version August 2017).

The BioPortal (https://bioportal.bioontology.org) repository ultimate is the efficient delivery of the existing biomedical semantic metadata to investigators, ontology and software developers to drive data integration [22]. Under the supervision of the National Center for Biomedical Ontology (NCBO) [23], the BioPortal supports a variety of knowledge representation formats, OWL, RDF or Biomedical Ontologies (OBO) and collects submitted biomedical ontologies from government, academic or private developer groups, under different types of license. These metadata content is openly available for evaluation, reused, reviewing, commenting and editing by the community of developers, as well as for embodiment in software applications [24]. Besides the enabling of ontology developers community participation, through the BioPortal tools, in particular the Ontology Recommender and the Annotator, BioPortal dynamically endorse metadata reuse and interlinking, thus performs as the driving force of the Life Sciences Linked Data movement.

Through the utilization of these tools and the proper management, the Life Sciences BioPortal

subdomain LOD Cloud exhibits an impressive number of extensively intelinked datasets, from the highly specific to the broader biomedical subject and from the smaller to the enormous in number of triples ones (Fig.3).

4 Results and Discussion

Because of this unique characteristics, herein, we focus in the BioPortal datasets to explore the pattern of their connectivity in relation to clinical, pharmacological and molecular biology concepts as expressed in the SNOMED CT, RxNORM and GO datasets, respectively. We postulated that if a dataset has links to these particular datasets, it would be possible to retrieve information and provide answers to complex biomedical queries that would assist clinical decision making.

The acquired LOD Cloud diagram data of BioPortal subdomain datasets links were analysed and we found that 96% of the BioPortal subdomain datasets share links with SNOMED CT, 24% with RxNORM, with an average of 1,821 ± 14,660 links, and 31.5% with GO, with an average of 258 ± 2,447 links. However, only 8 out of the 230 datasets, 3.5%, share links with both RxNORM and GO. These datasets are: (1) SNOMED CT, (2) MeSH, (3) CTV3, (4) Logical Observation Identifiers Names and Codes (LOINC), a database and universal standard for identifying medical laboratory observations, which developed in 1994 and since then maintained by a US non-profit medical research organization the Regenstrief Institute [25], (5) NCIt, (6) MeSH-OWL, (7) NDF-RT, and (8) Computer–Retrieval of Information on Scientific Projects (CRISP or CSP) an NIH developed biomedical terminology last updated in 2006 [26].

Noteworthy, although GO shares links with nearly one-third of the BioPortal domain datasets the number of links are one order of magnitude smaller than RxNORM. This finding suggests that, while molecular biology contribution in disease development and drug response or adverse reactions is well appreciate, the established connections between specific molecular biology concepts and clinical or pharmacological end-points are significant fewer than the connections between clinical and pharmacology entities. In a recent report the RxNORM representations were acknowledged as a starter tool in clinical decision making that needs expansion [27].

5 Conclusions and Future Work

The implementation of both evidence-based and precision medicine in clinical arena for efficient clinical decision making demands the development of semantic bioinformatics tools. The Life Sciences BioPortal subdomain LOD Cloud diagram include many biomedical datasets the combination of which may successfully deliver this task. The dynamic expansion of the LOD Cloud is accompanied by the explosive development of the BioPortal subdomain and the increase in the connectivity between its datasets. To meet the modern medicine targets the PICO criteria, combining patients, interventions, comparisons and outcomes data, should be applied. Herein, we focus in the retrieval of clinical, pharmacological and molecular biology information from different resources that could be combine through automatic reasoning and deliver accurate answers in complex medical questions. We used as base the triptych of SNOMED CT, RxNORM and GO datasets as the best representations of this information, respectively. The LOD Cloud data suggest that SNOMED CT is the core of the BioPortal weaving, while its closely associated datasets organize the subdomain thematically. However, pharmacological and molecular biology data sharing is possible for only a small fraction of BioPortal datasets. Thus, there is a need to emphasize in the development of bridging datasets that combine this information [28-30].

As future work, we plan to analyze the relationships between the BioPortal datasets in clusters. This analysis will shed light to the affinities of individual datasets with clinical, pharmacological and molecular biology data. The produced metadata may served as a pathway in the building of novel bridging datasets between these knowledge fields that can automatically resolve and deliver answers to complex clinicopathological queries

References:

[1] Issa NT, Byers SW, Dakshanamurthy S, Big data: the next frontier for innovation in therapeutics and healthcare, *Expert review of clinical pharmacology*, 7, 2014, 293-298.

[2] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y, Artificial intelligence in healthcare: past, present and

future, *Stroke and vascular neurology*, 2, 2017, 230-243.

[3] Adunlin G, Diaby V, Xiao H, Application of multicriteria decision analysis in health care: a systematic review and bibliometric analysis, *Health expectations : an international journal of public participation in health care and health policy*, 18, 2015, 1894-1905.

[4] Beckmann JS, Lew D, Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities, *Genome medicine*, 8, 2016, 134.

[5] Berners-Lee T, Hendler J, Publishing on the semantic web, *Nature*, 410, 2001, 1023-1024.

[6] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z, Dbpedia: A nucleus for a web of open data, in: *The semantic web*: Springer, 2007, pp. 722-735.

[7] Hartig O, Bizer C, Freytag J-C, Executing SPARQL queries over the web of linked data, in: International Semantic Web Conference: Springer, 2009, pp. 293-309.

[8] Singh S, How to Conduct and Interpret Systematic Reviews and Meta-Analyses, *Clinical and translational gastroenterology*, 8, 2017, e93.

[9] Nekrutenko A, Taylor J, Next-generation sequencing data interpretation: enhancing reproducibility and accessibility, *Nature Reviews Genetics*, 13, 2012, 667.

[10] Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S, Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement, *Theoretical and Applied Genetics*, 126, 2013, 867-887.

[11] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE, Big data: astronomical or genomical?, *PLoS biology*, 13, 2015, e1002195.

[12] Wilkinson J, Goff M, Rusoja E, Hanson C, Swanson RC, The application of systems thinking concepts, methods, and tools to global health practices: An analysis of case studies, *Journal of evaluation in clinical practice*, 24, 2018, 607-618.

[13] Greenhalgh T, Thorne S, Malterud K, Time to challenge the spurious hierarchy of systematic over narrative reviews?, *European journal of clinical investigation*, 48, 2018, e12931.

[14] Senthilkumar S, Rai BK, Meshram AA, Gunasekaran A, Chandrakumarmangalam S, Big Data in Healthcare Management: A Review of Literature, *American Journal of Theoretical and Applied Business*, 4, 2018, 57-69.

[15] Jovanovik M, Trajanov D, Consolidating drug data on a global scale using Linked Data, *Journal of biomedical semantics*, 8, 2017, 3.

[16] Wild DJ, Ding Y, Sheth AP, Harland L, Gifford EM, Lajiness MS, Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research, *Drug discovery today*, 17, 2012, 469-474.

[17] Fidler F, Chee YE, Wintle BC, Burgman MA, McCarthy MA, Gordon A, Metaresearch for evaluating reproducibility in ecology and evolution, *BioScience*, 67, 2017, 282-289.

[18] Andrew C, Heegaard E, Kirk PM, Bässler C, Heilmann-Clausen J, Krisai-Greilhuber I, Kuyper TW, Senn-Irlet B, Büntgen U, Diez J, Big data integration: Pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology, *Fungal Biology Reviews*, 31, 2017, 88-98.

[19] Altenhoff AM, Glover NM, Train C-M, Kaleb K, Warwick Vesztrocy A, Dylus D, de Farias TM, Zile K, Stevenson C, Long J, The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces, *Nucleic acids research*, 46, 2017, D477-D485.

[20] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J, Bio2RDF: towards a mashup to build bioinformatics knowledge systems, *Journal of biomedical informatics*, 41, 2008, 706-716.

[21] Dumontier M, Callahan A, Cruz-Toledo J, Ansell P, Emonet V, Belleau F, Droit A, Bio2RDF release 3: a larger connected network of linked data for the life sciences, in: Proceedings of the 2014 International Conference on Posters & Demonstrations Track, 2014, pp. 401-404.

[22] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA, BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic acids research*, 37, 2009, W170-173.

[23] Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story M-A, Smith B, team N, The national center for biomedical ontology, *Journal of the American Medical Informatics Association*, 19, 2011, 190-195.

[24] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, *Nucleic acids research*, 39, 2011, W541-W545.

[25] Forrey AW, Mcdonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, Fiers T, Charles L, Griffin B, Stalling F, Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results, *Clinical chemistry*, 42, 1996, 81-90.

[26] Bair AH, Brown LP, Pugh LC, Borucki LC, Spatz DL, Taking a bite out of CRISP. Strategies on using and conducting searches in the Computer Retrieval of Information on Scientific Projects database, *Computers in nursing*, 14, 1996, 218-224; quiz 225-216.

[27] Freimuth RR, Wix K, Zhu Q, Siska M, Chute CG, Evaluation of RxNorm for Medication Clinical Decision Support, *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*, 2014, 2014, 554-563.

[28] Martzoukos, Y., Papavlasopoulos, S., Syrrou, M., & Poulos, M. (2015, July). Biobibliometrics & gene connections. In *Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on* (pp. 1-5). IEEE.

[29] Poulos, M., & Papavlasopoulos, S. (2013, July). Automatic stationary detection of time series using auto-correlation coefficients and LVQ—Neural network. In Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on (pp. 1-4). IEEE.

[30] Poulimenou, S., Stamou, S., Papavlasopoulos, S., & Poulos, M. (2014). Keywords Extraction from Articles' Title for Ontological Purposes. In *Proceedings of the 2014 International Conference on Pure Mathematics, Applied Mathematics, Computational Methods (PMAMCM 2014)* (pp. 120-125).