Minimisation of Terms to Describe a Knowledge Domain for Ontology Engineering and Linked Data Generation

ARTEMIS CHALEPLIOGLOU, SOZON PAPAVLASOPOULOS, MARIOS POULOS
Department of Archives, Library Science and Museology,
Faculty of Information Science & Informatics
Ionian University
Ioannou Theotoki 72, 491 00 Corfu
GREECE
mpoulos@ionio.gr

Abstract: - Access and retrieval of scholarly data through the web is often difficult because of the lack of intelligent algorithms that could logical analyze and compute the searching requests. The Semantic web represent the solution to this problem. The building of a new semantic ontology, to describe particular knowledge domain, is a challenging and demanding approach. Herein, we explore cardiology as a paradigm for the definition of a sufficient vocabulary. Firstly, we define a set of textbooks of the knowledge domain according to the following criteria: (a) degree of field coverage; (b) recommendations and guidelines of the field professionals; and (c) popularity of the textbooks based on sales analytics and bibliometrics. Secondly, we extract the terms indexed in these textbooks in worksheets allowing duplicates. Thirdly, we used the frequency of appearances of a term in the combined master index as an independent and objective quantitative variable of its impact in the knowledge domain description. Finally, we define the desirable power of the knowledge field description of our new ontology and use the subset of the most frequent terms as core for the building of the new ontology. This road map may serve in similar applications in ontology generation and maintenance.

Key-Words: - Semantic web, ontology, bibliometrics, intelligent algorithm, linked data, cardiovascular biology

1 Introduction

Domain and data knowledge is the first step to deploy a semantic dataset. The degree of success of an ontology such an effort is strongly depending on the understanding of the domain data by the publisher. Cardiology represents a major medical knowledge domain where the cardiologists in collaboration with scientists from other disciplines, epidemiologists, biologists, pharmacologists and bioinformaticians collectively fight against the cardiovascular diseases [1]. The incidence of cardiovascular diseases is dramatically increasing worldwide. Accumulating evidence suggest that the clinical phenotype of the cardiovascular diseases patients is the outcome of a complex array of interactions between genotype, lifestyle, and drugs alike. In the era of continuously growing and affordable omics technologies, high-throughput genomics, transcriptomics, metabolomics, and proteomics data are increasingly available. When this information is appropriate combined with clinical and research medical literature we may decipher important answers in cardiovascular disease pathology and treatment. However, analysis of the cardiology Big Data is depending on the generation of appropriate ontologies and the application of Linked Data technologies.

To this end significant efforts have been performed or are still in progress including the CardioVascular Research Grid (CVRG) [2, 3], the representation of heart development in the gene ontology [4], the circulatory system ontology based on ICD-11 and SNOMED CT [5], the implantable electronic devices recordings ontology [6], and the human disease network of electronic health record data [7]. However, a major challenge remains to achieve maximum interoperability, the requirement for submerge the different types of cardiological terms, clinical, physiology, pathology, and cell biology into a common dataset.

Herein, we utilized bibliographic reasoning to select the appropriate cardiological terms to describe: (a) clinical entities (anatomy, physiology, pathology, and surgery); (b) basic biological entities (genes and proteins that regulates heart physiology and pathophysiology, hereditary human diseases, and drug metabolism); and (c) therapeutic entities that could be applied in cardiovascular

[*] Corresponding author: mpoulos@ionio.gr

diseases (including both surgical and pharmacological interventions). The criterion for the selection of terms was the frequency of their appearances in the indices of cardiology textbooks.

2 Selection of Textbooks

To explore the cardiology data and to identify the specific needs of cardiology professionals, physicians and researchers, we focus in the scientific textbooks describing this field. Firstly, we defined the project priorities: (a) Clinical Cardiology, (b) Molecular Cardiology, and (c) Cardiovascular Pharmacology.

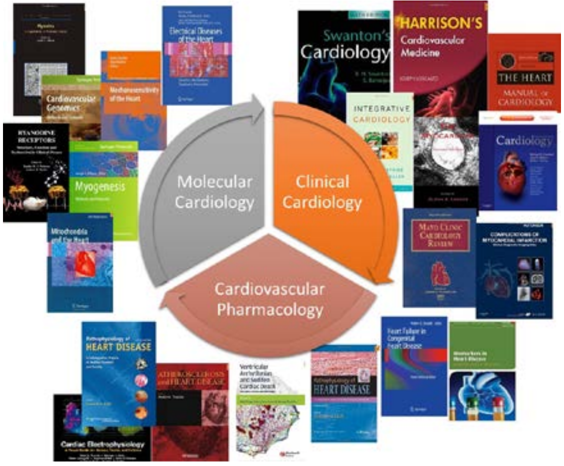
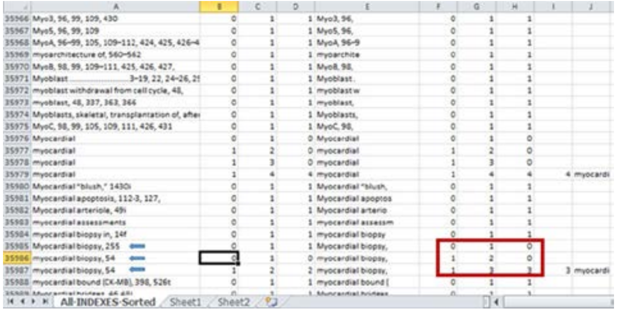


Fig.1. Representative examples of the textbooks analyzed.

We anticipated that the collection of all terms describing the concepts of these specific scientific subfields would formulate a compiled vocabulary that facilitates cardiologic clinical decision-making. Secondly, we defined the criteria for the textbook selection: (a) the degree of field coverage, (b) the recommendations and guidelines by the field professionals, and (c) the popularity of the textbooks. Twenty-five cardiology textbooks were selected to explored utilizing library and information science tools, 7 of general cardiology [8-14], 8 of pathology [15-22], 4 of physiology [23-26], and 6 focusing in molecular biology of cardiovascular diseases [27-32]. The different degree of specialization of the selected textbooks ensured the widest coverage of the field (Fig.1).

3 Analysis of Indexed Terms

The index of each textbook selected was extracted as a set of terms Ai = (w1, w2,…,wn), whereas i is the serial number of the textbook and w the terms. The indexed terms included single words, compound words and multi-word expressions. We merged the indices from the twenty-five textbooks into a single file without removing duplications. This master set, of 56134 terms, used to determine the impact of each term description to the knowledge domain. The master list of terms sorted alphabetically, including compound words and multi-word expressions. The frequency of appearances of a term in the combined master index was calculated through the combination of logical algorithms and counting. In specific, when wi = wi+1 was true, a counter formula add plus one for each appearance of the term, but when wi ≠ wi+1 then the counter restarted from one. The alphabetically sorted list of terms was scored for the determination of the frequency of appearances of each term in the compile master index. In the example in Fig.2 the term "myocardial biopsy" found repeated 3 times.



	A	B	C	D	E	F	G	H	I	J
35966	Myo3, 96, 99, 109, 430	0	1	1	Myo3, 96,	0	1	1		
35967	Myo5, 96, 99, 109	0	1	1	Myo5, 96,	0	1	1		
35968	MyoA, 96–99, 105, 109–112, 424, 425, 426–4	0	1	1	MyoA, 96–9	0	1	1		
35969	myoarchitecture of, 560–562	0	1	1	myoarchite	0	1	1		
35970	MyoB, 98, 99, 109–111, 425, 426, 427,	0	1	1	MyoB, 98,	0	1	1		
35971	Myoblast ....3–19, 22, 24–26, 2!	0	1	1	Myoblast .	0	1	1		
35972	myoblast withdrawal from cell cycle, 48,	0	1	1	myoblast w	0	1	1		
35973	myoblast, 48, 337, 363, 366	0	1	1	myoblast,	0	1	1		
35974	Myoblasts, skeletal, transplantation of, after	0	1	1	Myoblasts,	0	1	1		
35975	MyoC, 98, 99, 105, 109, 111, 426, 431	0	1	1	MyoC, 98,	0	1	1		
35976	Myocardial	0	1	0	Myocardial	0	1	0		
35977	myocardial	1	2	0	myocardial	1	2	0		
35978	myocardial	1	3	0	myocardial	1	3	0		
35979	myocardial	1	4	4	myocardial	1	4	4	4	myocardi
35980	Myocardial "blush," 1430i	0	1	1	Myocardial "blush,	0	1	1		
35981	Myocardial apoptosis, 112-3, 127,	0	1	1	Myocardial apoptos	0	1	1		
35982	Myocardial arteriole, 49i	0	1	1	Myocardial arterio	0	1	1		
35983	myocardial assessments	0	1	1	myocardial assessm	0	1	1		
35984	myocardial biopsy in, 14f	0	1	1	myocardial biopsy	0	1	1		
35985	Myocardial biopsy, 255 ⟸	0	1	1	Myocardial biopsy,	0	1	0		
35986	myocardial biopsy, 54 ⟸	0	1	0	myocardial biopsy,	1	2	0		
35987	myocardial biopsy, 54 ⟸	1	2	2	myocardial biopsy,	1	3	3	3	myocardi
35988	myocardial bound (CK-MB), 398, 526t	0	1	1	myocardial bound (	0	1	1		
35989	Myocardial bridges, 46 48i	0	1	1	Myocardial bridges	0	1	1		

All-INDEXES-Sorted / Sheet1 / Sheet2

Fig.2. The alphabetically sorted list of terms scoring approach.

We found that 38,005 terms mentioned only once in the master index, 11,786 terms twice, 2,590 trice, whilst five terms appeared more than 350 times (Table 1).

Table 1. Repeats of terms in textbooks.

Times	Terms
1	38005
5	2623
10	1021
20	401
30	240
40	164
50	117
100	40
150	23
200	14
250	7
300	6
350	5

The Cardiology domain could sufficiently described with a minimum set of the terms indexed in textbooks. The number of terms versus the number of their repetition were diagrammatically semi-logarithmical designed (Fig.3A). The cardiology indices data tested versus different non-linear regression models. We found that the power regression model (log-log regression model) exhibited the best fit to our data following the equation:

$$y = a\,x^{\beta} \tag{1}$$

with an R^2=0.998 (α = 18764 and β = 1.318). We used the *Hirsch index* function,

$$h - index = \max_{i} \min(f(i), i) \tag{2}$$

to define the *h*-index for the cardiology terms multiple appearances. We found that it’s value is 68, indicating that at least 68 terms appears 68 or more times in the compiled cardiology textbooks indices. The indefinite integral of the power function between the terms and their repeats is:

$$\int \alpha\, x^{\beta}\, dx = \frac{ax^{\beta+1}}{\beta+1} + C \tag{3}$$

The definite integral of this function for the 68 terms repeated at least 68 times and above found to represent 70% of the total area under curve of the function of terms versus their repeats (Fig.3B). We found that 1.5‰ of the unique terms indexed in the Cardiology textbooks indices could be used as a core for the formation of an ontology describing this knowledge field.

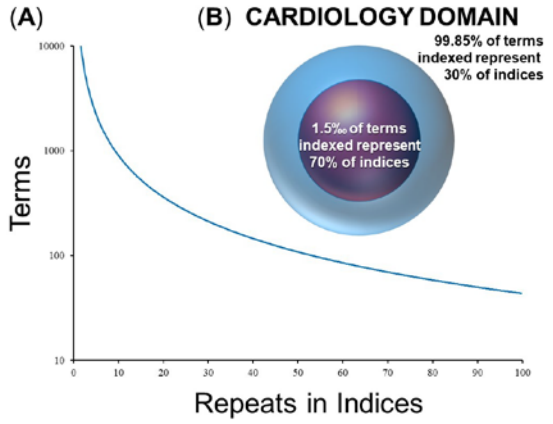


Fig.3. (A) Diagrammatic representation of cardiology terms versus their repeats in textbook indices. (B) Graphical representation of the volume of the whole cardiology knowledge domain and the highly repetitive terms.

Among the terms identified were: heart, aorta, echocardiography, tomography, cardiomyopathy, arrhythmia, atrial fibrillation, ventricular tachycardia, pacemaker, implantable cardioverter defibrillator, syncope, sudden cardiac death, cardiac hypertrophy, atherosclerosis, anticoagulants, angiotensin-converting enzyme, beta blockers, antiarrhythmics, digitalis, diuretics, and genes related to cardiovascular diseases and hereditary syndromes, such as APOB, MYH7, TNNT2, MYBPC3, NOTCH1, and PLN.

4 Conclusion

A knowledge domain terminology may include innumerable terms, single words, compound words and multi-word expressions. These descriptions not only express knowledge domain entities but also refer to interplays, hierarchical classifications and data structure, all necessary components for the logical analysis and decision making of this domain information. Ideally, an ontology describing the knowledge domain should contain all terms and interactions between them to accurate depict the domain. However, such an approach is hard or impossible to apply in real life situations because of the complexity of interactions, programming restrictions and hardware limitations. In ontology engineering, is common to start from a core set of terms, setting the interactions between them, setting the interactions with other ontologies in the Linked Data context, and gradually update the data by incorporating more terms and functions. How can one define a minimum set of terms that sufficiently describe a knowledge domain?

In this work, we present an example from cardiology scientific field as a pathway for the definition of terms adequate describing the field that may serve as a core for ontology building. Our approach based on the fundamental textbooks describing the knowledge domain, collected through pre-specified criteria, from which the indices extracted and compiled into a set by allowing the repetition of terms. The number of repeats of each term in this set used as the quantitative variable define its significance in the description of this knowledge domain. The mathematical description of terms as a function of their repeats and the application of non-linear logistic regression allow the determination of the best-fitted equation as well as data integration. The area under curve of the terms versus their repeats function represent in its completeness the corpus of the knowledge domain of interest. We used the *Hirsch h*-index to identify the nodal number of terms appeared in the knowledge domain indices is such a frequency that can adequate describe it. We anticipate that this approach is applicable in many different knowledge domains for the generation of ontological and linked data contextualization of variable data resources.

References:

[1] K.W. Johnson, K. Shameer, B.S. Glicksberg, B. Readhead, P.P. Sengupta, J.L.M. Björkegren, J.C. Kovacic, J.T. Dudley, Enabling Precision Cardiology Through Multiscale Biology and Systems Medicine, *JACC: Basic to Translational Science*, Vol.2, 2017, pp. 311-327.

[2] S. Steinert-Threlkeld, S. Ardekani, J.L. Mejino, L.T. Detwiler, J.F. Brinkley, M. Halle, R. Kikinis, R.L. Winslow, M.I. Miller, J.T. Ratnanather, Ontological labels for automated location of anatomical shape differences, *Journal of biomedical informatics*, Vol.45, 2012, pp. 522-527.

[3] R.L. Winslow, J. Saltz, I. Foster, J.J. Carr, Y. Ge, M.I. Miller, L. Younes, D. Geman, S. Graniote, T. Kurc, R. Madduri, T. Ratnanather, J. Larkin, S. Ardekani, T. Brown, A. Klasny, K. Reynolds, M. Shipway, M. Toerper, The CardioVascular Research Grid (CVRG) Project, in: Proceedings of the AMIA *Summit on Translational Bioinformatics* 2011, pp. 77-81.

[4] V.K. Khodiyar, D.P. Hill, D. Howe, T.Z. Berardini, S. Tweedie, P.J. Talmud, R. Breckenridge, S. Bhattarcharya, P. Riley, P. Scambler, R.C. Lovering, The representation of heart development in the gene ontology, *Developmental biology*, Vol.354, 2011, pp. 9-17.

[5] J.M. Rodrigues, S. Schulz, A. Rector, K. Spackman, J. Millar, J. Campbell, B. Ustun, C.G. Chute, H. Solbrig, V. Della Mea, K.B. Persson, ICD-11 and SNOMED CT Common Ontology: circulatory system, *Studies in health technology and informatics*, Vol.205, 2014, pp. 1043-1047.

[6] A. Rosier, P. Mabo, M. Chauvin, A. Burgun, An ontology-based annotation of cardiac implantable electronic devices to detect therapy changes in a national registry, *IEEE journal of biomedical and health informatics*, Vol.19, 2015, pp. 971-978.

[7] B.S. Glicksberg, L. Li, M.A. Badgeley, K. Shameer, R. Kosoy, N.D. Beckmann, N. Pho, J. Hakenberg, M. Ma, K.L. Ayers, G.E. Hoffman, S. Dan Li, E.E. Schadt, C.J. Patel, R. Chen, J.T. Dudley, Comparative analyses of population-scale phenomic data in electronic medical records reveal race-specific disease networks, *Bioinformatics*, Vol.32, 2016, pp. i101-i110.

[8] M.H. Crawford, Cardiology, 3rd ed., Mosby/Elsevier, Philadelphia, 2010.

[9] S.R. Devries, J.E. Dalen, Integrative cardiology, Oxford University Press, Oxford ; New York, 2011.

[10] J.W. Hurst, R.A. Walsh, V. Fuster, J.C. Fang, Hurst's the heart manual of cardiology, 13th ed., McGraw-Hill, New York, 2013.

[11] G.A. Langer, The myocardium, 2nd ed., Academic Press, San Diego, 1997.

[12] J. Loscalzo, T.R. Harrison, Harrison's cardiovascular medicine, 2nd ed., McGraw-Hill Education/Medical, New York, 2013.

[13] J.G. Murphy, M.A. Lloyd, Mayo Clinic., Mayo Clinic cardiology : concise textbook, 4th ed., Mayo Clinic Scientific Press/Oxford University Press, Oxford ; New York, 2013.

[14] R.H. Swanton, S. Banerjee, Swanton's Cardiology : a Concise Guide to Clinical Practice., 6 ed., John Wiley & Sons, Chichester, 2009.

[15] F.R. Breijo-Marquez, M.P. Ríos, The Variations in Electrical Cardiac Systole and Its Impact on Sudden Cardiac Death, INTECH Open Access Publisher, 2012.

[16] J.A. De Lemos, American Heart Association., Biomarkers in heart disease, Blackwell Pub., Malden, Mass., 2008.

[17] S.J. Hutchison, Complications of myocardial infarction : clinical diagnostic imaging atlas, Saunders/Elsevier, Philadelphia, PA, 2009.

[18] T.B. Levine, A.B. Levine, Metabolic syndrome and cardiovascular disease, Saunders/Elsevier, Philadelphia, PA, 2006.

[19] J. Marin-Garcia, M.J. Goldenthal, G.W. Moe, Aging and the heart: a post-genomic view, Springer, New York, 2008.

[20] R.E. Shaddy, Heart failure in congenital heart disease : from fetus to adult, Springer, London [u.a.], 2011.

[21] A. Tonkin, Atherosclerosis and Heart Disease, Martin Dunitz, New York, 2003.

[22] P.J. Wang, H.H. Hsia, A. Al-ahmad, Ventricular Arrhythmias and Sudden Cardiac Death : Mechanism, Ablation, and

Defibrillation., John Wiley & Sons, Chichester, 2009.

[23] I. Gussak, C. Antzelevitch, A.A.M. Wilde, P.A. Friedman, M. Ackerman, W.K. Shen, Electrical Diseases of the Heart : Genetics, Mechanisms, Treatment, Prevention, Springer-Verlag, London, 2008.

[24] A.G. Kamkin, I. Kiseleva, Mechanosensitivity of the heart, Springer Verlag, Dordrecht ; New York, 2010.

[25] L.S. Lilly, Harvard Medical School., Pathophysiology of heart disease : a collaborative project of medical students and faculty, 5th ed., Wolters Kluwer/Lippincott Williams & Wilkins, Baltimore, MD, 2011.

[26] D.P. Zipes, J. Jalife, Cardiac electrophysiology : from cell to bedside, 4th ed., Saunders, Philadelphia, 2004.

[27] L.M. Coluccio, Myosins : a superfamily of molecular motors, Springer, Dordrecht, 2008.

[28] J.X. DiMario, Myogenesis : methods and protocols, Humana Press ; Springer, New York, 2012.

[29] K. DiPetrillo, Cardiovascular genomics : methods and protocols, Humana Press, Totowa, N. J., 2009.

[30] V.J. Dzau, C.-C. Liew, Cardiovascular genetics and genomics for the cardiologist, Blackwell Futura, Malden, Mass., 2007.

[31] J. Marín-García, M.J. Goldenthal, Mitochondria and the heart, Springer, New York, 2005.

[32] X.H.T. Wehrens, A.R. Marks, Ryanodine receptors : structure, function, and dysfunction in clinical disease, 2005, New York, 2005.