H-index for the determination of sufficient terms to describe a scientific field

ARTEMIS CHALEPLIOGLOU,
Department of Archives, Library Science and Museology,
Faculty of Information Science & Informatics
Ionian University
Ioannou Theotoki 72, 491 00 Corfu
GREECE
artemischal@ionio.gr
SOZON PAPAVLASOPOULOS,
Department of Archives, Library Science and Museology,
Faculty of Information Science & Informatics
Ionian University
Ioannou Theotoki 72, 491 00 Corfu
GREECE
sozon@ionio.gr
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: - The Semantic web offers the intelligent algorithms that could logical analyze scholarly data and retrieve accurate results in scientific research queries. It is based on the generation of ontologies that describe particular knowledge domains. The building of a new ontology is a challenging and demanding approach. Herein, we enable the Hirsch index (h-index) to define the critical terms needed for the description of the cardiology domain. To this end we generated a master vocabulary to describe cardiology derived from relative textbooks by allowing duplicates. More than 56,000 unique terms were collected. The frequency of appearances of each term was used as the sole criterion for the evaluation of its importance in the cardiology domain description. The power regression (log-log) model best fits to these data compared to different non-linear regression models. Therefore, we apply the h-index function to define the sufficient number of the multiple appeared cardiology terms that could describe this particular scientific field. We found that the h-index for the cardiology terms is 68, indicating the number of terms appearing equally or more than 68 times in the corpus of cardiology textbooks. The definite integral of the power function between the terms and their repeats for the 68 terms was found to represent 70% of the total area under curve. Thus, approximately 1.5‰ of the unique terms indexed in the Cardiology textbooks may be used as the core for the development of a cardiology ontology. We propose that this methodology may serve as a road map in similar librarian applications.

Key-Words: - Semantic web, ontology, bibliometrics, non-linear regression, cardiology

1 Introduction

The degree of success of an ontology dataset is strongly depending upon the understanding of the domain data by the publisher. This argument suggests that expert opinion is nodal for the development of an ontology. However, problems in organization, design, bias assessment, searching, management, source selection, data synthesis and document delivering strongly suggest the involvement of librarians and the utilization of bibliographic tools, metrics and reasoning [1].

The motivations for the deployment of a Cardiology ontology were: (a) the need for collaboration between this scientific field professionals, the cardiologists, with scientists from other disciplines, such as epidemiologists, biologists, pharmacologists and bioinformaticians [2]; (b) the continuous increasing incidence of cardiovascular diseases worldwide; (c) the complex interplay between patients' clinical phenotype and genetic and environmental factors, such as lifestyle and drugs; (d) the growing body of high-throughput genomics, transcriptomics, metabolomics, and proteomics data; (e) the restricted number of currently available tools such as the CardioVascular Research Grid (CVRG) [3, 4], the representation of heart development in the gene ontology [5], the circulatory system ontology based on ICD-11 and SNOMED CT [6], the implantable electronic devices recordings ontology [7], and the human disease network of electronic health record data [8]. A major challenge in Cardiology field is the requirement to combine different types of terms, representing clinical, physiology, pathology, and cell biology entities 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) molecular biology entities (genes and proteins that regulates heart physiology and pathophysiology, hereditary human diseases, and drug metabolism); and (c) therapeutic modalities (including both surgical and pharmacological interventions) (Fig.1). The frequency of each term appearances in the indices of cardiology textbooks was used as bibliometric tool. The criterion for the necessary and sufficient terms that may serve as a core for the description of cardiology in a semantic ontology was the Hirsch index (h-index).

2 Collection of Terms

To retrieve the terms describing the cardiology field we select a set of the available cardiological textbooks with the following criteria: (a) the recommendations and guidelines of the field professionals (AHA, ACC, ESC), (b) the popularity of textbooks based on sales analytics and bibliometrics (google trends, amazon best sellers and worldCat), and (c) the coverage of the field, by prioritising clinical cardiology, molecular cardiology, and cardiovascular pharmacology areas, was determined with bioportal annotator search of their indexed terms (Fig.2). The indices of the retrieved documents were extracted and saved in text (.txt) files.

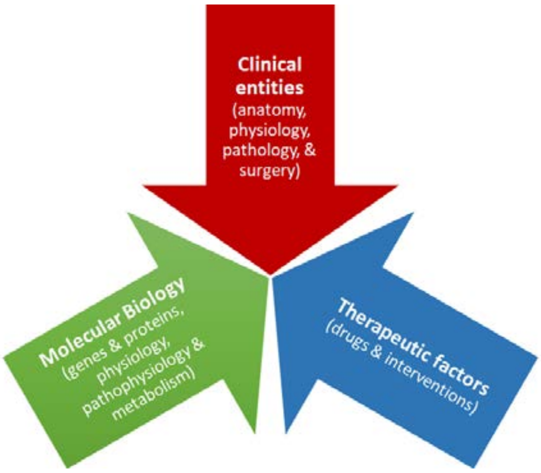
Clinical entities
(anatomy, physiology, pathology, & surgery)
Molecular Biology
(genes & proteins, physiology, pathophysiology & metabolism)
Therapeutic factors
(drugs & interventions)

Fig.1. Cardiology ontology design.

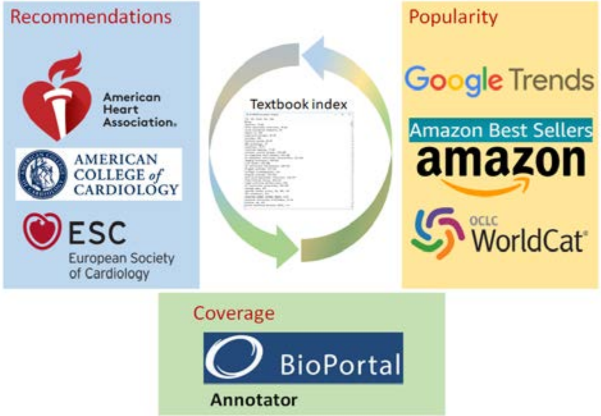

Fig.2. Criteria for cardiology textbooks selection.

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. Twenty-five cardiology textbooks were selected to explored utilizing library and information science tools, 7 of general cardiology [9-15], 8 of pathology [16-23], 4 of physiology [24-27], and 6 focusing in molecular biology of cardiovascular diseases [28-33]. The different degree of specialization of the selected textbooks ensured the widest coverage of the field.

3 Analysis of Terms by the h-index

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 (Fig.3). This master set, of 56134 unique terms, used to determine their impact in the description of the cardiology field 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.

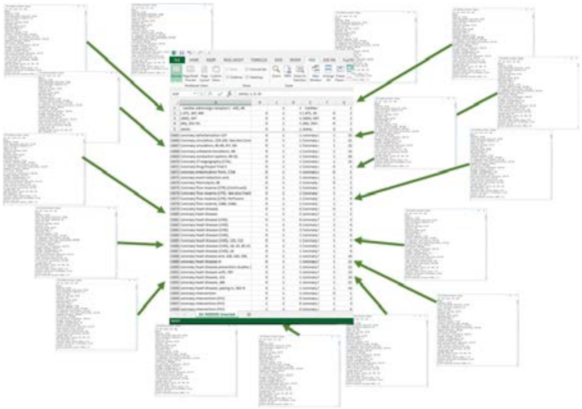

Fig.3. Combination of terms into a single file, alphabetically shortening and scoring.

We found that 38,005 terms mentioned only once in the master index, 11,786 terms twice, 2,590 trice, 1,130 four times, and 609 five times, whilst five terms appeared more than 350 times.

The Cardiology domain could be 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.4). 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, as a non-linear index to evaluate the terms impact [34]:

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

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:

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

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. 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.

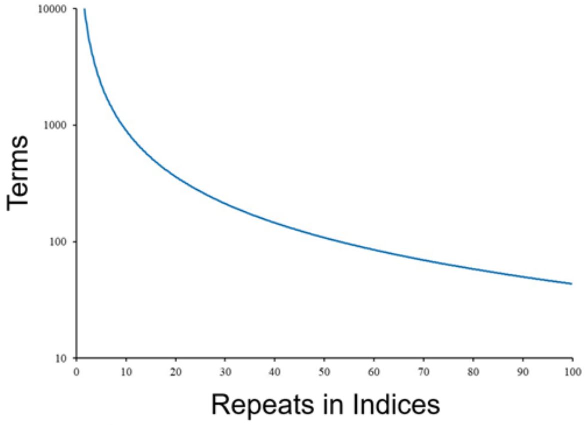

Fig.4. Diagrammatic representation of cardiology terms versus their repeats in textbook indices.

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.

4 Conclusion

Semantic web description of a knowledge domain is strongly depending on ontology building. In life sciences an ontology should serve a as a bridge connecting clinical, pharmacological and molecular biology metadata in order to allow complex reasoning and clinical decision making assistance. Therefore, the innumerable terms, single words, compound words and multi-word expressions descriptions refer to entities but also to interplays, hierarchical classifications and data structure, should be collected and build on logical basis. However, such an approach is hard or impossible to apply in real life situations because of the complexity of interactions, the number of terms, programming restrictions and hardware limitations. Therefore, to explore the possibilities of a semantic ontology system for cardiology, we tried to define a minimum set of terms that sufficiently describe the knowledge domain. Herein, we used bibliographic and bibliometric reasoning for the analysis of a composite master index of Cardiology. 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. We used the h-index as the criterion to identify the nodal number of terms appeared in the knowledge domain indices in such a frequency that can adequate describe it. The area under curve of the terms versus their repeats function represent in its completeness the corpus of the knowledge domain of interest. Our data suggest 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] Y.Martzoukos, S. Papavlasopoulos, M. Syrrou, M. Poulos. Bibliometrics & gene connections, *6th International Conference on Information, Intelligence, Systems and Applications*, 2015, pp. 1-5.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

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

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

[14] 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.

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

[16] 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.

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

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

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

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

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

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

[23] P.J. Wang, H.H. Hsia, A. Al-ahmad, Ventricular Arrhythmias and Sudden Cardiac Death : Mechanism, Ablation, and Defibrillation., John Wiley & Sons, Chichester, 2009.

[24] 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.

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

[26] 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.

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

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

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

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

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

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

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

[34] S. Papavlasopoulos, M. Poulos, N. Korfiatis, G. Bokos. A non-linear index to evaluate a journal’s scientific impact, Vol.180, 2010, pp. 2156-2175.