Mapping the Life Science using Medical Subject Headings (MeSH)

Fei Shu[1], Junping Qiu[1] and Vincent Lairière[2]

[1] *fei.shu@hdu.edu.cn; casee.hdu@outlook.com; tjmxsyx@sina.com at*
Hangzhou Dianzi University, Chinese Academy of Science and Education Evaluation (CASEE), Xiasha, Hangzhou, Zhejiang (China P.R.)

[2] *vincent.lariviere@umontreal.ca at*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, Montréal, Quebec (Canada)

Abstract
Maps of scientific knowledge are generally based on citation analysis and therefore reveal how disciplines draw from each other to produce new knowledge. Although subject headings as well as their co-assignments represent the topics and their relationships within the journal article or book, they rarely have been used for mapping science. This study attempts to map the life science based on the Medical Subject Headings (MeSH) as well as their co-assignment at the paper level, which could advance the knowledge in mapping science.

Introduction
The purpose of mapping science is to visualize the scientific structure and the evolution of scientific inquiry (Börner, Theriault, & Boyack, 2015; Klavans & Boyack, 2015) by classifying science and relating the classes, which are generally derived from the analyses of scientific literature elements such as authors, journals, disciplines or other information (Klavans & Boyack, 2009). Although citation analysis is the dominant method for generating maps of science, expert judgements, subject categories, topic modelling, course descriptions, and subject headings could also be used to map the science.

Medical Subject Headings (MeSH) are controlled vocabularies for indexing journal articles and books in the life sciences, which represent all topics discussed within the journal article or book. Since a journal article or book could be assigned multiple MeSHs, the MeSH co-assignments could be used to measure the relationship between two medical topics by which the structure and evolution of life science could be mapped. The purpose of this study is to generate a map of life science using the MeSHs.

Related Works
Expert judgment was first used for mapping science. Bernal (1939) drew the first map of science representing the hierarchical structure of scientific topics by hand. Small and Griffith (1974) created the first citation-based map of science using co-citation analysis. Since then, citation analysis including direct citation (Boyack & Klavans, 2014b; Pan, Zhang, & Wang, 2013; Waltman & Eck, 2012), bibliography coupling (Boyack, 2008), co-citation (Boyack & Klavans, 2014a; Braam, 1991a, 1991b; Small, 1999) was widely used for mapping science.

Other methods in addition to the citation analysis were also used for mapping science. A map of science could be generated based on the co-occurrence of words in titles, abstracts or keywords using the co-word analysis (Ding, Chowdhury, & Foo, 2001; Leydesdroff, 1989; Peters & van Raan, 1993a, 1993b; Rip & Courtial, 1984). Balaban and Klein (2006) mapped science using undergraduate course pre-requisites at Texas A&M University. Suominen and Toivanen (2016) generated a map of science using topic modelling based on the latent patterns in texts retrieved from the Web of Science (WoS).

Subject headings was also applied to generating the map of science. Shu, Dinneen, Asadi, and Julien (2017) produced a map of science based on non-fiction books and their Library of Congress Subject Heading (LCSH) co-assignments. Leydesdorff, Comins, Sorensen, Bornmann, and Hellsten (2016) tried to compare the MeSH with cited sources using clustering and mapping. However, a map of life science based on MeSH has not been generated, which will be addressed by this study.

Method

In this study, in addition to MeSH co-assignment as discussed above, MeSH of citing/cited papers was also used to generate the map as the contrast. Although each MeSH term represents a topic discussed in the journal articles or papers, MeSH terms representing the major topics are marked in the PubMed. Each pair of the major MeSH terms between citing and cited papers also represents the relationship between two major medical topics. Leydesdorff et al. (2016) point out that the citation (citing/cited) map indicates a core structure of life science while the MeSH map shows the relevance of the life science research. Thus, two maps generated from two different approaches were compared in this study.

Data

In this study, 3,344 research papers published in four top medical journals (i.e., *The Journal of the American Medical Association, The Lancet, New England Journal of Medicine*, and *The British Medical Journal*) between 2015 and 2017 as well as their cited references were retrieved from Web of Science (WoS). A version of MEDLINE database integrated into the WoS was used as the linkage between WoS and PubMed in which a PubMed ID and MeSH terms were assigned to each journal article. As noted, not all papers are covered by both WoS and PubMed; in this study, only papers, either citing or cited, with a PubMed ID were included. Eventually, as Table 1 shows, 2,577 papers as well as their 80,782 cited references were collected under investigation; 5,119 and 16,582 MeSH terms were assigned to these citing papers and cited references respectively.

Table 1. Distribution of Papers and Cite References under Investigation in the Study

Journal	*Number of papers in WoS*	*Number of papers under investigation (citing paper)*	*Number of Cited Reference*	*Number of MeSH Terms (citing paper)*	*Number of MeSH Terms (cited reference)*
The Journal of the American Medical Association	658	516	13,889	1,749	9,589
The Lancet	963	658	25,459	2,062	8,696
New England Journal of Medicine	1,003	841	19,866	2,424	11,858
The British Medical Journal	720	562	19,704	2,926	11,812
Total	**3,344**	**2,577**	**80,782**	**5,119**	**16,582**

Note: Since one reference or MeSH term could be cited or assigned to different papers, the sum of the number of cited reference and the sum of number of MeSH term of four journals are higher than the totals in the last row.

Data Treatment

MeSH terms are organized as a 14-level tree structure, representing medical topics from broad to specific. This tree structure starts with 16 level-1 MeSH terms and 118 level-2 MeSH terms, on which the maps of life science were based. Assigned MeSH terms at level 3 or lower were

re-assigned to their parent level-2 or grandparent level-1 MeSH terms. This method of reassignment to broader or more general abstraction levels, has been used in library classification mapping where its robustness has been confirmed (Shu et al., 2017).

As shown in Table 2, four datasets were finalized to produce four maps of life science: MeSH co-assignment map at level 1, MeSH co-assignment map at level 2, MeSH citation map at level 1, MeSH citation map at level 2. For each dataset, MeSH terms as well as their co-assignments or citation pairs (major MeSH terms between citing and cited papers) were imported into graph-drawing software *Gephi* to generate the visual map of life science. Each MeSH term was a node while each MeSH co-assignment or citation pair was an edge in the map. The number of assignment of each MeSH term determined the size of node while the number of co-assignment or citation pair decided the weight of each edge. Although the number of citation pairs are much higher than the number of co-assignments in these datasets, they could be normalized when producing the map through Gephi settings.

Table 2. Four Datasets for Four Maps of Life Science

Map	*Number of MeSH term*	*Number of unique MeSH co-assignment or citation pair*	*Total number of MeSH co-assignment or citation pair*
Co-assignment at level 1	16	105	104,832
Co-assignment at level 2	107	3,305	96,776
Citing/cited at level 1	16	110	818,944
Citing/cited at level 2	113	4,767	1,015,203

Results

Figure 1 shows the two maps of life science at the MeSH term level 1 containing 16 nodes/105 edges (lower) and 16 nodes/110 edges (upper) respectively. Nodes are level 1 MeSH terms while edges represent their relationship (i.e., co-assignment and citation pair respectively). Edge width is proportional to the number of co-assignment or citation pair between the two MeSH terms, and the node and label sizes are proportional to the number of assignments or citations.

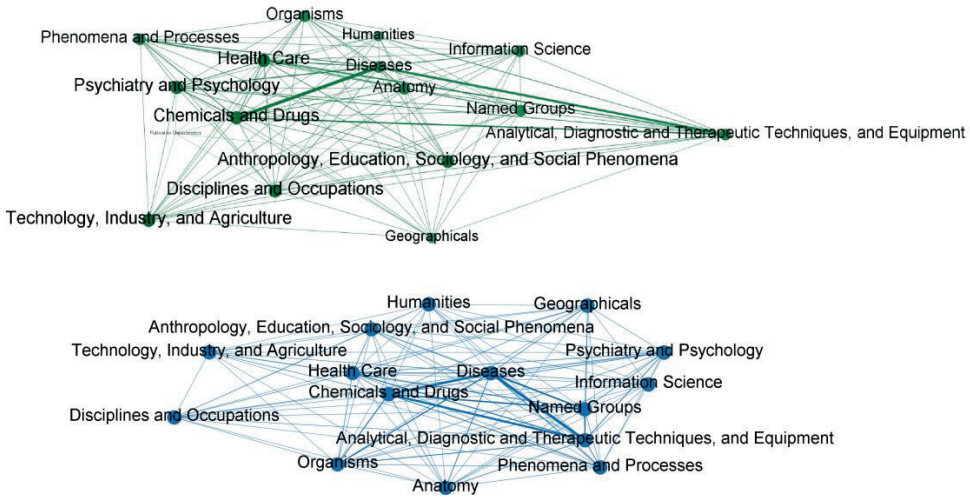

Figure 1. MeSH citing/cited map (upper) and co-assignment map (lower) at level 1.

The difference between the co-assignment map and the citation map is not significant. A strong triangle relationship among *Diseases*, *Chemicals and Drugs*, and *Analytical, Diagnostic and Therapeutic Techniques, and Equipment* was found in both maps. Indeed, as Table 3 indicates, the top 10 MeSH terms in the both maps are in the same order and similar shares.

Table 3. Top 10 MeSH terms (Level 1) in Co-assignment Map and Citing/cited Map

MeSH	*Co-assignment Map*	*Citation Map*
Diseases	23.36%	22.12%
Anatomy	16.82%	17.7%
Phenomena and Processes	14.95%	15.04%
Chemicals and Drugs	14.02%	14.16%
Analytical, Diagnostic and Therapeutic Techniques, and Equipment	6.54%	6.19%
Health Care	5.61%	5.31%
Organisms	3.74%	4.42%
Psychiatry and Psychology	3.74%	3.54%
Anthropology, Education, Sociology, and Social Phenomena	2.8%	2.65%
Technology, Industry, and Agriculture	2.8%	2.65%

The color-coded legend of level 1 MeSH terms (see right of Figure 2) were used in the level 2 maps as shown in Figure 2. Nodes are level 2 MeSH terms as the colours of nodes represent their parent MeSH terms at level 1. Some differences were found when comparing the co-assignment map (lower of Figure 2) and the citation map (upper of Figure 2). The distribution of MeSH of citing/cited papers is skewed as some large nodes and wide edges appear in the citation map, while the distribution of MeSH co-assignments is more balanced. However, comparing with Figure 1, Figure 2 is visually complex due to high connectivity between the nodes and overlapping edges.

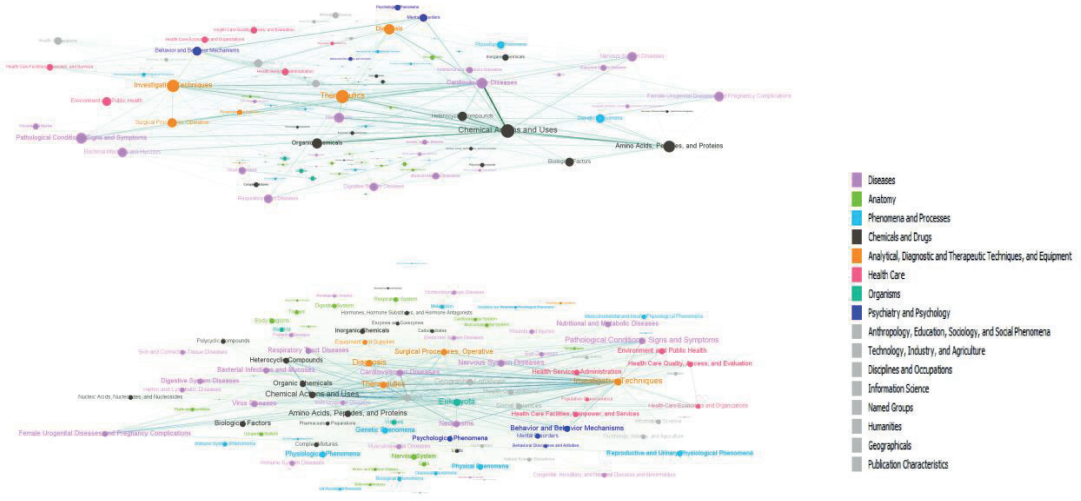

Figure 2. MeSH citing/cited map (upper) and co-assignment map (lower) at level 2.

Discussion and Conclusion

In this study, four maps of life science were generated using the MeSH terms assigned to 2,577 papers published in four top medical journals between 2015 and 2017 as well as their cited reference. Few difference was found when comparing the co-assignment map with the citation map at the MeSH level 1. It indicates that the MeSH co-assignment representing the relationship among different medical topics could also be used to map the life science comparing to traditional citation-based maps generated by the citing/cited relations. The results of this study could form a foundation for future studies mapping the life science using MeSH terms.

In addition, this study found the difference in terms of the MeSH term distribution between the co-assignment map and the citation map at the MeSH level 2, which could partly be due to the different functions between subject headings and citations. Subject headings emphasize the correlation of all related topics discussed in the journal articles or books while citations measure the similarity of citing and cited documents, which has been addressed by Leydesdorff et al. (2016).

As a research-in-process paper, this study only sampled 2,577 research papers from four top medical journals, a full dataset containing all medical articles should be investigated in the future studies. In addition, different visualization methodologies, in turn stemming from choice of visualization software, may also influence the visualization of the map, which should also be addressed in the future studies.

References

Balaban, A. T., & Klein, D. J. (2006). Is chemistry "The Central Science"? How are different sciences related? Co-citations, reductionism, emergence, and posets. *Scientometrics, 69*(3), 615-637.

Bernal, J. D. (1939). The social function of science. *The Social Function of Science.*

Börner, K., Theriault, T. N., & Boyack, K. W. (2015). Mapping science introduction: Past, present and future. *Bulletin of the American Society for Information Science and Technology, 41*(2), 12-16.

Boyack, K. W. (2008). Using detailed maps of science to identify potential collaborations. *Scientometrics, 79*(1), 27-44.

Boyack, K. W., & Klavans, R. (2014a). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology, 65*(4), 670-685.

Boyack, K. W., & Klavans, R. (2014b). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics, 8*(3), 569-580.

Braam, R. R. (1991a). Mapping of Science by Combined Co-Citation and Word Analysis. I. Structural Aspects. *Journal of the American Society for Information Science, 42*(4), 233-251.

Braam, R. R. (1991b). Mapping of Science by Combined Co-Citation and Word Analysis. II: Dynamical Aspects. *Journal of the American Society for Information Science, 42*(4), 252-266.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management, 37*(6), 817-842.

Klavans, R., & Boyack, K. W. (2009). Toward a Consensus Map of Science. *Journal of the American Society for Information Science and Technology, 60*(3), 455-476.

Klavans, R., & Boyack, K. W. (2015). Exploring the relationships between a map of altruism and a map of science. *Bulletin of the American Society for Information Science and Technology, 41*(2), 30-33.

Leydesdorff, L., Comins, J. A., Sorensen, A. A., Bornmann, L., & Hellsten, I. (2016). Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering

and mappings at the paper level. *Scientometrics, 109*(3), 2077-2091. doi:10.1007/s11192-016-2119-7

Leydesdroff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*(4), 209-223.

Pan, J., Zhang, X., & Wang, X. (2013). Mapping the Structure and Evolution of Science. *Journal of information processing and management, 52*(8), 255-266.

Peters, H. P. F., & van Raan, A. F. J. (1993a). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy, 22*(1), 23-45.

Peters, H. P. F., & van Raan, A. F. J. (1993b). Co-word-based science maps of chemical engineering. Part II: Representations by combined clustering and multidimensional scaling. *Research Policy, 22*(1), 47-71.

Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics, 6*(6), 381-400.

Shu, F., Dinneen, J. D., Asadi, B., & Julien, C.-A. (2017). Mapping science using Library of Congress Subject Headings. *Journal of Informetrics, 11*(4), 1080-1094.

Small, H. (1999). Visualizing Science by Citation Mapping. *Journal of the American Society for Information Science, 50*(9), 799-813.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science studies*, 17-40.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human‐assigned subject classification. *Journal of the Association for Information Science and Technology*.

Waltman, L., & Eck, N. J. (2012). A new methodology for constructing a publication‐level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392.