This is a preprint version of a paper that has been accepted to appear in *Scientometrics*, 2/8/17.

Thesaurus-based methods for mapping contents of publication sets

Kevin W. Boyack[1]

[1] *kboyack@mapofscience.com*
SciTech Strategies, Inc., 8421 Manuel Cia Pl NE, Albuquerque, NM 87122 (USA)

Abstract

Visualization of literature-related information is common in scientometrics and related fields. Despite this, relatively little work has been done to visualize knowledge organization systems, such as controlled vocabularies or thesauri. In this paper we explore the creation and use of contextual visualizations based on thesauri. Two different methods are developed for creating maps of thesaurus terms that can then be used as templates or basemaps on which to display the contents of publication sets. The first example maps first-level terms from the Unified Astronomy Thesaurus (UAT) into a wheel-like (hub and spokes) configuration. This circular map can then be used to show relative positions of clusters of astronomy papers from different cluster solutions based on the thesaurus terms assigned to the papers in the clusters. The second example triangulates the entire Public Library of Science (PLOS) thesaurus onto a global map of science, and then uses the resulting map of thesaurus terms as the basis for an overlay map. This map can be used for several purposes, including mapping of subsets of PLOS content, and the identification of thesaurus terms whose rule bases may need to be changed.

Introduction

The use of visualization to display the results of literature-based analysis has become pervasive. Common visualizations include timelines (or alluvial diagrams), scatterplots (typically based on similarities), network graphs, heat maps and more. Any of these visualization types can be used for science mapping. Many visualizations show the objects of analysis directly. For instance, VOSviewer (van Eck & Waltman, 2010) can be used to create a scatterplot of journals (or documents, terms, etc.) where similar objects are near each other. Other scatterplots or network diagrams show the relative positions of clusters of objects (rather than the objects themselves) created from analysis of a dataset.

Some visualizations, however, show the results of a literature analysis within a larger context rather than directly. For these visualizations, a larger document set – often an entire database – is used to create a conceptual space that is then used as a template within which the results of an analysis can be projected. Overlay maps are one example of this type of visualization. Overlay maps exist for journals (Leydesdorff & Rafols, 2012; Leydesdorff, Rafols, & Chen, 2013), journal subject categories (Leydesdorff, Carley, & Rafols, 2013; Rafols, Porter, & Leydesdorff, 2010), clusters of papers (Boyack & Klavans, 2014b), and patent categories (Kay, Newman, Youtie, Porter, & Rafols, 2014; Leydesdorff, Kushnir, & Rafols, 2014), often with tools that make them available to researchers worldwide. Templates can also be based on triangulation. Objects such as journals, categories, or terms can be ordered around a circle, and the results of an analysis can then be placed within the circle using triangulation. For example, a "circle of science" which orders fields around a circle has been used to display competencies (clusters of papers) for universities and countries (Börner et al., 2012; Klavans & Boyack, 2010). Using this template, disciplinary clusters appear near the edge of the circle, while multi-disciplinary clusters are closer to the center.

Despite the common use of visualization, relatively little work has been done to visualize knowledge organization systems (KOS), such as controlled vocabularies or thesauri, whether expert-based or crowd-sourced (Akdag Salah, Gao, Suchecki, & Scharnhorst, 2011). In this paper we explore the creation and use of contextual visualizations based on thesauri. The first section of this paper will provide some context about thesauri, their structure, indexing, and mapping. This will be followed by presentation of two different methods for creating concept

spaces and templates based on two different thesauri, along with examples of visualizations using these templates and their utility.

Background

Thesauri and indexing

To most people, a thesaurus is a dictionary of synonyms and antonyms. It contains words or phrases grouped together by similarity of meaning. However, within the realm of information retrieval, a thesaurus is somewhat different. A thesaurus builds on the notion of a controlled vocabulary, which is a set of terms that have been defined to cover a particular information space. A taxonomy is a controlled vocabulary that is organized in a hierarchical or tree-like structure with parent/child relationships between terms. With a taxonomy, as one moves down the tree (deeper in the hierarchy) terms become more specific. Each term is typically related to either a broader term (at a higher level) or narrower term (at a lower level) within the tree structure, or to both broader and narrower terms. A thesaurus differs from a taxonomy in that it can include relationships that connect terms that are not already in a parent/child relationship (Hlava, 2015). Ideally, all of these relationships should be considered when creating a visualization based on thesaurus terms.

The method for creating and maintaining a thesaurus typically involves a combination of algorithmic and manual work (Hlava, 2015). In practice this is done using software (typically commercial) where each thesaurus term is represented by a set of textual triggers, or rules. The sum total of these rules is known as the rule base for the thesaurus. Rule bases can be modified as needed by those who maintain a particular thesaurus. One common operation is modification of the rules for a particular term to make it either less specific or more specific, thus leading to the term being triggered either more or less frequently, respectively. Other common operations include adding, deleting or splitting of terms. Single terms or even entire branches (hierarchical sets of terms) can occur in multiple places within a thesaurus.

Once a thesaurus is in place, it is commonly used to index the documents in a particular corpus. Perhaps the best known literature-based thesaurus is the Medical Subject Headings (MeSH) thesaurus that is maintained by the U.S. National Library of Medicine (NLM). Although MeSH is widely used, it is perhaps not well known that MeSH is actually a hierarchical thesaurus (with up to 12 levels) rather than simply a flat structure of keywords. MeSH has a sophisticated rule based that is maintained by NLM. When a new document is added to PubMed, the title and abstract of that document are run against the MeSH rule base, which suggests possible terms within the thesaurus that can be used to index the document. These index terms are then approved or discarded by human indexers at NLM; those that are approved become the MeSH terms for the document. Although human indexers are used by NLM for MeSH indexing, this is not necessarily common. For many thesauri, once the rule base is complete, indexing is done solely by machine. Thesauri are more common in the literature world than one might think. For example, PubMed, JSTOR, INSPEC, PLOS, and many other literature aggregators and publishers actively maintain thesauri by which they index their database contents to make them more usefully retrievable by their clients. Many journals (such as JASIST) include a taxonomy or thesaurus in their submission systems, asking authors to tag their submissions using the controlled vocabulary terms.

Thesaurus visualization

As mentioned above, relatively little work has been done to visualize information systems, and particularly to use such visualizations as visual reference systems for sets of publications.

Visualization of a thesaurus is typically straight-forward. Given its hierarchical structure, a thesaurus is often visualized using a tree graph or radial tree graph. For example, Figure 1 shows the thesaurus from the Public Library of Science (PLOS) as a tree graph and a radial tree graph with 11 first-level terms and a maximum depth of 8 levels. Each of the branch structures associated with a first-level term uses a different color (with the exception of 'Science Policy' and 'Social Sciences', which have the same color), making it easy to see the relative sizes of the different branches. The black coloring in these graphs shows those terms and branches in the PLOS thesaurus that occur in multiple locations. This poly-hierarchy is intentional. It indicates that a term has multiple meanings – and thus associations with multiple parent terms – that are equally valid. Note that the graphs of Figure 1 are based only on the parent/child relationships, and do not show the term links between terms in different branches.

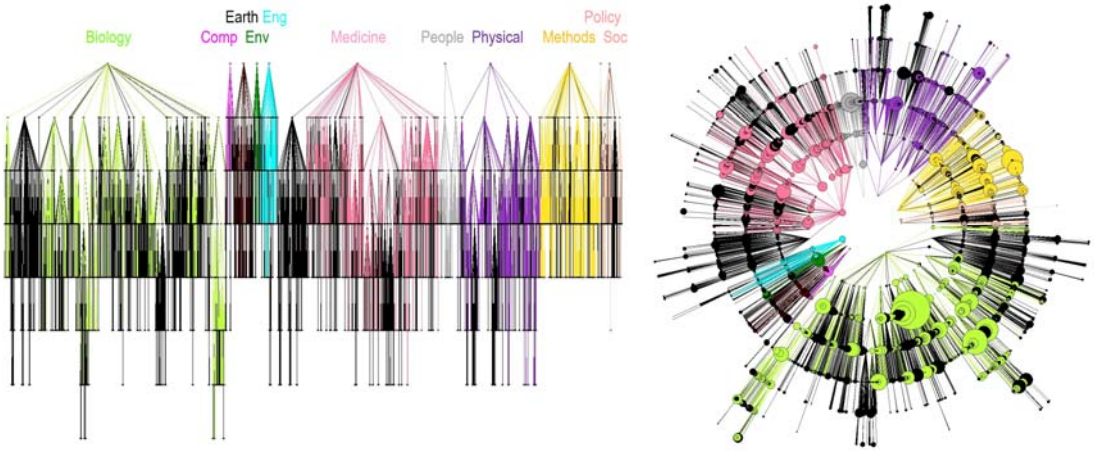

Figure 1. The PLOS thesaurus shown as a tree graph (left) or radial tree graph (right). Nodes in the radial tree graph are sized to reflect the number of PLOS documents indexed by each term in the thesaurus.

Tree graphs are very useful for some applications. However, the ordering of the upper levels of a thesaurus (e.g., first and second levels) is often not based on similarity. For example, the PLOS tree graph of Figure 1 orders terms alphabetically within levels and branches rather than by similarity. If they were ordered by similarity, the biology and medicine branches would likely be next to each other. However, this would not assure that similar terms lower in the hierarchy would be proximate to each other, either within or across branches. This is particularly true for similar terms from different levels of the thesaurus. Thus, these graphs cannot be used to display the kind of information that relies on a proximity-based visualization.

Information systems can also be visualized using standard network representations such as those produced by force-directed placement algorithms (e.g., the Fruchterman-Reingold implementations in Pajek or Gephi). Petersen, Rotolo & Leydesdorff (2016) provide an example of this with their recent visualization of MeSH terms. Other examples, both of tree and network visualizations, can be found in Katy Börner's *Atlas of Knowledge* (Börner, 2015, pp. 124, 154, 162).

In contrast, the kind of term mapping often associated with co-term analysis does produce proximity-based visualizations. Co-term analysis was introduced over 30 years ago by Callon (Callon, Courtial, Turner, & Bauin, 1983), and has since been widely used to characterize and visualize document sets and fields. The relationship between terms is typically based on the number of documents in which those terms co-occur, often normalized using cosine or a related measure, thus producing a proximity-based visualization. Today, many term mapping studies

are done using sophisticated tools such as VOSviewer (van Eck & Waltman, 2010) and Sci2 (Börner et al., 2010).

Although term mapping based on co-term analysis is a commonly used technique, it seems to have been used but rarely to create a map template using thesaurus terms. In fact, we are only aware of a single instance – there may, of course, be more of which we are unaware – where a map has been created from thesaurus terms as basemap for overlays. Leydesdorff, Rotolo & Rafols (2012) created a map of 822 first and second level categories from three branches (Diseases, Drugs & Chemicals, and Techniques & Equipment) of the U.S. National Library of Medicine's Medical Subject Headings thesaurus – commonly known as MeSH – using PubMed documents from 2010. Even though the MeSH thesaurus is 10 or 11 levels deep in these branches, only the first two levels were used to restrict the visualization to high level terms, and to keep it from becoming too complex. Leydesdorff et al. (2012) counted co-occurrences between directly indexed first and second level terms, produced cosine-normalized matrices from the counts, and then used both Pajek and VOSviewer to create maps of terms. This is a fairly standard approach to co-term mapping. Terms from lower levels (3-11) were ignored and not accounted for. Ignoring the lower levels of a tree can have a large effect on the similarity values and resulting maps. For example, if a particular branch of the tree contains a large number of terms in its lower levels that are heavily indexed, this branch is likely to have much smaller similarity values when based on only one or two levels than if it were based on all levels, and can thus lead to less accurate visualizations than if all terms are considered. Alternately, as we show below, one can calculate weights and similarities for entire branches of the tree at a particular level. All terms within a branch can be counted as belonging to their higher level terms for the purposes of computing weights and similarities. This has the advantage of accounting for all terms, while restricting the visualization (for purposes of simplicity and readability) to those terms at a particular level. The patent category maps of Boyack & Klavans (2008) and Kay et al. (2014) based on the hierarchical IPC structure are two examples of maps where counts rolled up from lower level categories to higher level categories were used to calculate similarity between categories.

The MeSH map of Leydesdorff et al. (2012) showed relatively clean separation between the three branches that it mapped. Terms from the Techniques & Equipment branch, while mostly congregated in one portion of the map, were more dispersed than terms from either of the other two branches. This map was used to show the dispersion of publications by a single author (using an overlay), and also to show the dispersion of research related to RNA interference. In the following sections, we show two examples of mapping of thesaurus terms that are less conventional, both in their methodology of construction and in their display. In both cases, the lack of convention arises because of the specific aims of the visualizations.

Methods for thesaurus visualization

Creation of a simple non-overlay thesaurus visualization

This issue of *Scientometrics* contains a number of papers reporting studies of 111,616 papers in an Astronomy dataset (hereafter called Astro-set). This data set has been clustered in a number of different ways by several teams of researchers. As a group, we sought means to compare the different cluster solutions. However, as non-experts in astronomy we were not capable of the type of detailed analysis of cluster contents that could be done by an expert in the field who recognizes papers, authors, terms, and knows of their inherent relationships. Thus, we sought methods of comparison that could be used by non-experts. We became aware that there was a public domain thesaurus specific to astronomy, the Unified Astronomy Thesaurus

(UAT, http://astrothesaurus.org), and decided to investigate the possibility of using this thesaurus as a means of comparing cluster solutions. Further, since visual comparisons can be very effective in showing similarities and differences between cluster solutions, we decided to see if this thesaurus could be used as the basis for a visualization. In essence, a visualization based on thesaurus terms is intended to add an interpretive layer to cluster solutions produced using bibliometric techniques. This section describes the UAT thesaurus and our efforts to create a visualization template based on this thesaurus as specifically applied to the Astro-set.

The UAT is an open-source thesaurus created by the astronomy community. It builds upon several existing thesauri, including those from the International Astronomical Union, Institute of Physics Publishing, and the American Institute of Physics. The UAT thesaurus (and rule base) are the product of decades of experience in the Astronomy community. It contains 15 first-level terms, and 1,915 unique terms occurring in 3,778 locations at a maximum depth of 12 levels. Many of the branches of terms occur in more than one location in the tree structure.

The Astro-set corpus was indexed to generate thesaurus terms for each document. This was done (pro bono) by Access Innovations, a company that specializes in thesaurus creation and maintenance, using their MAI (Machine Aided Indexer) software package and the UAT rule base that they maintain. Titles and abstracts for each document were used as the input for indexing. Up to 10 terms were kept per document. After indexing, roughly 10% of the documents had no index terms, meaning that for those papers the text (title and abstract) did not trigger any of the rules in the thesaurus rule base. It is not uncommon for a document to not trigger any rules, and thus have no index terms. Often, this information is later used by the thesaurus designers to help them add or modify rules to increase the accuracy and coverage of their thesaurus.

Table 1 shows the number of term occurrences and their summed fractional weights for the first-level UAT terms in the Astro-set. It also shows the total number of term occurrences and summed weights of all terms in each level 1 branch of the thesaurus. Comparison of the term and branch occurrence numbers shows the stark differences that can occur if terms from the lower branches are ignored in an analysis or mapping of first-level terms. 'Cosmology' dominates the direct first-level term occurrence, and will thus dominate any similarities based only on first-level term co-occurrence. In addition, three terms of these terms are not indexed directly and two others are indexed only twice. In contrast, branch occurrences and weights are dominated by 'Astronomical objects', as are the resulting similarities (as we will see later). It is clear from these numbers that a map that ignores information from the lower levels may be substantially different than one that includes information 'rolled-up' to the branch level.

Table 1. Comparison of prevalence of level 1 UAT terms and their associated branches in the Astro-set.

Level 1 term	# papers	term weight	# branch	branch weight
Astronomical objects	100	28.35	440795	35454.89
Astrophysical magnetism	0	0.00	9582	2850.04
Celestial mechanics	52	16.57	47339	10902.73
Cosmology	2598	735.99	70474	11292.97
Equipment and apparatus	32	9.94	25068	4863.43
Galactic physics	2	0.27	4956	743.91
Interdisciplinary astronomy	0	0.00	16136	3098.31
Lunar physics	0	0.00	95	22.90

Methods and techniques	5	1.42	66506	9453.86
Nuclear astrophysics	20	5.96	9093	527.47
Observational astronomy	25	5.31	90394	9371.29
Planetary science	57	15.65	14744	3169.53
Positional astronomy	2	0.42	7856	1350.43
Space exploration	75	30.68	3781	474.66
Stellar physics	66	11.34	55898	8094.55

With our ultimate goal of creating a visualization that could be used to compare different cluster solutions of a single dataset, we embarked on a two-step process – first, to create the visualization template using the index terms from all Astro-set papers, and second, to use the visualization to display the different cluster solutions. Simple visualizations are better than complex visualizations at enabling intuitive understanding. Thus, we set out to create a very simple visualization that would be similarity-based.

As mentioned above, circular visualizations are often quite intuitive, and we have experience with their use. If a set of objects that are the basis for visualization are placed sequentially around a circle in an order that is based on similarity between those objects, then various representations composed of those objects can be displayed and compared visually (Börner et al., 2012; Klavans & Boyack, 2010). In our case, we chose to create a circle composed of first-level thesaurus terms from all Astro-set papers with these terms ordered by similarity. Using the positions of each term, one can then calculate the average position of each cluster in each solution based on the terms assigned to the papers in each cluster. If clusters are located towards the middle of the circle their content 'smears' across the thesaurus structure while if they are nearer the edge of the circle, there is a greater correspondence with the organization of content as defined by the thesaurus.

The first step in this process was to see if the first-level terms would align themselves roughly around a circle based on similarities between them. Similarities between first-level terms based on the Astro-set corpus, and a preliminary visualization were created using the following method:

1) Terms were fractionally assigned to documents. For instance, if a document had 9 index terms, each term was assigned a weight of 1/9 for that document.
2) Each indexed term was rolled up to its first-level term. For example, if the level 3 term 'Ring nebulae' was assigned to a document, the corresponding first-level term in the hierarchical tree directly above this term was used for the analysis (in this case 'Astronomical objects').
3) First-level terms and their weights were summed for each document.
4) Similarities between first-level terms were calculated on a per-document basis by multiplying the weights associated with each term, and then summing over all documents in the corpus.
5) Visualizations of the first-level terms were created using their similarities as edge weights using both DrL and the Fruchterman-Reingold algorithm (in Pajek), and are shown in Figure 2.

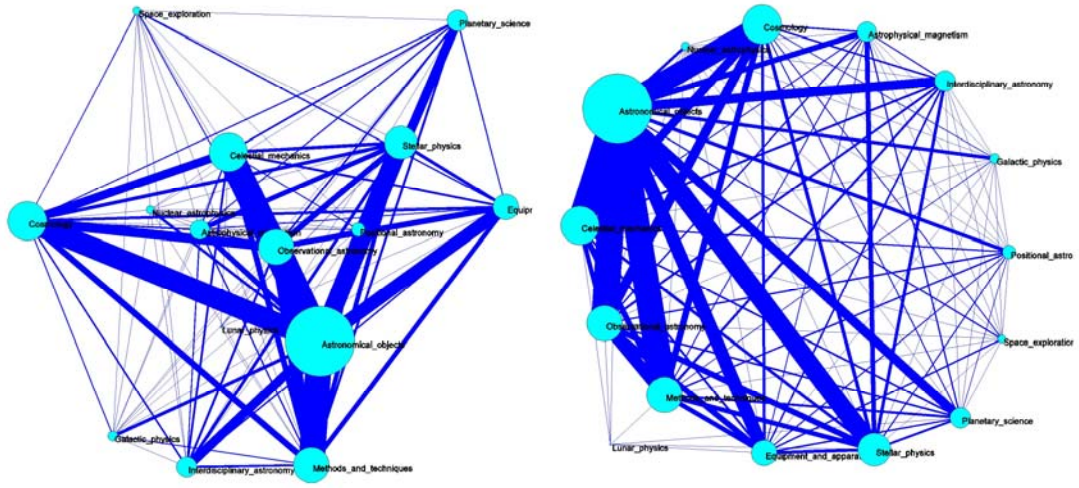

Figure 2. DrL and Pajek visualizations of UAT first-level terms using all similarities. Node sizes reflect the numbers of papers (summed fractional counts) associated with each term. Edge thicknesses show relative similarities between pairs of terms.

These two visualizations lead to some interesting observations. First, 'Astronomical objects' – the largest node, see Table 1 – has strong similarities to most of the other terms. Although a circle would have provided an ideal visualization template, it is clear that these data cannot be represented naturally and unambiguously by a circle. Second, the DrL visualization tends toward a wheel-like structure (central hub with spokes), suggesting that 'Astronomical objects' might work well as a central node.

Based on these observations, we created another visualization where instead of using all similarities, we only used the top two similarity values and edges for each first-level term. We also deleted the term 'Lunar physics' from the remaining calculations due to its very small overall weight (see Table 1). The resulting visualization (using Pajek) was nearly circular with 'Astronomical objects' in the middle. Given the near circularity of the visualization, and given that 'Astronomical objects' naturally ended up in the center of the visualization, we chose to use a wheel-like (hub and spokes) type of formation with 'Astronomical objects' as the hub, and the remaining 13 terms evenly distributed around the perimeter of the circle as shown in Figure 3. This visualization was then used as the visual template upon which cluster solutions were overlaid.

We note that there are some edges that cross one another in this visualization. This is unavoidable given the data. Nevertheless, the edge crossings are minimized by this representation, and the visualization largely preserves the dominant similarities between first-level terms around the perimeter of the circle.

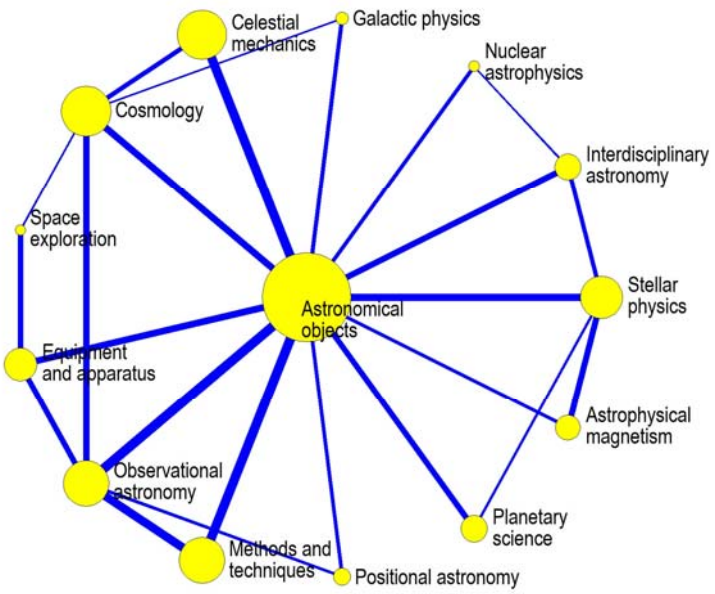

Figure 3. Hub and spokes visualization of UAT first-level terms with 'Astronomical objects' as a central node.

With this template now in place, overlays of nine different cluster solutions on the circular UAT first-level visual template were created, as shown in Figure 4. Each cluster in each solution is represented as a purple circle where size reflects cluster size. The nine solutions have a similar look and feel, meaning that in general the clusters are in the same areas of the circle. Clusters near the central node ('Astronomical objects') have strong representation of that branch of the thesaurus in their index terms. Clusters near the upper left are related to 'Cosmology' and 'Celestial mechanics', while clusters closer to the lower right are more related to 'Astrophysical magnetism' and 'Planetary science'.

Differences are also reflected in the numbers of clusters. For example, from top to bottom, solutions show varying levels of detail – two of the solutions on the top row (EBC and EHY) have 13 and 11 clusters, while the solutions on the bottom row have over 100 clusters each. As one moves from top (few clusters) to bottom (many clusters), in most cases the number of clusters in each quadrant of the map increases with the total number of clusters. Also, for solutions with larger numbers of clusters, more of those clusters move nearer to the edge of the circle, thus indicating clusters that are more focused within first-level thesaurus branches. This is particularly true for the STS Lev1 solution (bottom middle), which contains a large number of small, highly differentiated clusters.

In summary, Figures 3 and 4 show that it was possible to create a simple visualization template of high-level thesaurus terms based on the full set of index terms within the branches of the thesaurus, and that this template provides a way to visually compare cluster solutions. One could go further. For example, Figure 4 suggests particular pairs of clusters from different solutions that could be chosen for detailed comparison based on their being in similar positions in the visualization. The lower right clusters from the EBC and EHY solutions are one example of clusters that may be similar in content based on their positions.

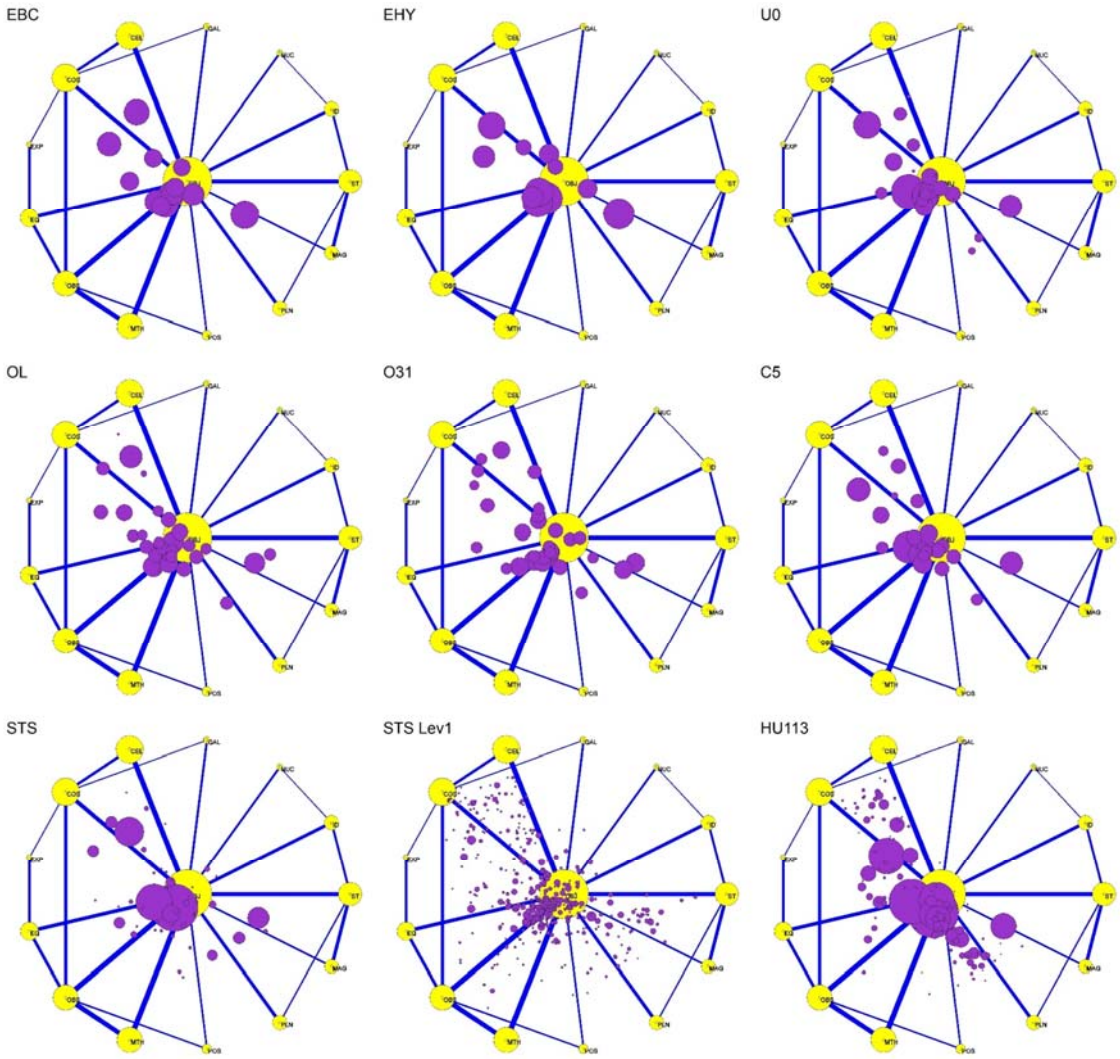

Figure 4. Overlays of nine different cluster solutions on the UAT first-level visual template.

Creation of an overlay thesaurus visualization based on global mapping

We now move to a second type of visualization that can be created using thesaurus terms. This case had a very different aim. Instead of comparing cluster solutions, the purpose of this visualization was to create a similarity-based visual template of the PLOS thesaurus (see Figure 1) that could be used to show overlays of the time-dependent portfolios of different journals and subsets of papers. The tree and radial tree graphs of Figure 1, although simple to generate, are not suitable for this task because journals and sets of papers are best represented using a template that is similarity-based rather than alphabetically ordered.

We used the version of the PLOS thesaurus that was current as of August 2013. It contained 11 first-level terms, and 10,551 unique terms occurring in 15,164 locations at a maximum depth of 8 levels. The thesaurus has undergone a number of changes since that time. The first and second levels of the PLOS thesaurus correspond directly to PLOS subject areas (http://www.plosone.org/taxonomy). Term and branch weights for first-level terms are given in Table 2.

This is a preprint version of a paper that has been accepted to appear in *Scientometrics*, 2/8/17.

Table 2. Comparison of level 1 PLOS term and branch weights in the PLOS corpus.

Level 1 term	term weight	branch weight
Biology and life sciences	2.18	42000.81
Computer and information sciences	2.27	949.55
Earth sciences	0.30	1004.50
Ecology and environmental sciences	1.26	1295.95
Engineering and technology	39.41	848.00
Medicine and health sciences	43.42	18573.60
People and places	0	1859.27
Physical sciences	0.65	5469.01
Research and analysis methods	0.03	7397.61
Science policy	5.92	160.03
Social sciences	3.67	1519.68

It is well known that visualizations with a large number of terms often become unintelligible, and that visualizing sets of terms whose weights vary over many orders of magnitude can also be problematic. These are some of the reasons that Leydesdorff et al. (2012) chose to visualize only the top two levels of the MeSH thesaurus rather than including more levels in their visualizations. Accordingly, we restricted our first attempt at visualizing the PLOS thesaurus to the 296 unique terms from the second level. A data set comprised of 642,166 terms associated with 81,078 PLOS documents (2003 through mid-2013), previously indexed using MAI (Dupuich & Carr, 2013), was made available to us for this purpose. We used a co-term methodology very similar to that used by Leydesdorff et al. (2012). However, the number of second level terms indexed directly was relatively small (only 22,600 out of 642k terms). Thus, we chose to use all terms (except first-level terms) and sum the counts and co-occurrences by second level branch. For example, if a single document had assigned to it a level-4 term and a level-6 term, the co-occurrence was calculated to occur between their respective second level terms. Term-term similarities were calculated using the full aggregated data using fractional counts, similar to the method used with the UAT first-level terms in our first example.

A Fruchterman-Reingold map of the PLOS second level terms creating using the top three edges per term is shown in Figure 5. Although this map provides a reasonable representation of the biomedical sciences, we were not comfortable with it because the natural sciences (physics, chemistry, mathematics, etc.) and engineering are placed in most cases within the biomedical sciences rather than as separate fields. The map of Figure 5 bears little resemblance to consensus maps of science (Klavans & Boyack, 2009). This does not indicate a mistake in the methodology, but rather reflects the fact that PLOS journals are strongly focused in biomedicine, and that non-biomedical science will appear within that context rather than within an "all of science" context if the map is based solely on PLOS content. Given that this map does not resemble consensus maps, and given perceptions that PLOS will expand their content into non-biomedical fields, we decided against using this map, and went another direction.

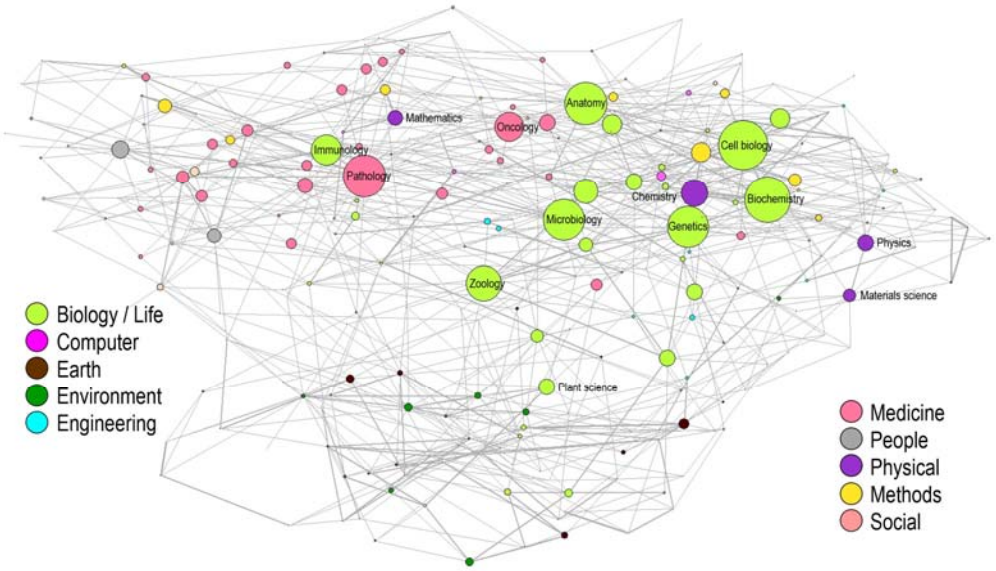

Figure 5. Mapping of second level terms from the PLOS thesaurus. Node sizes reflect the numbers of papers (summed fractional counts) associated with each term.

Figure 1 shows that the PLOS thesaurus covers all areas of science. However, the thesaurus is not evenly distributed – there are far more terms and more detailed terms in the biology and medicine branches than in branches representing the natural sciences and engineering. Nevertheless, the coverage does extend to all of science. We thus determined that a map of PLOS thesaurus terms should be based on a much broader basis than just PLOS content, which is primarily (90%+) biomedical. Accordingly, we indexed five years of titles and abstracts from Scopus (2007-2011) using the MAI software and PLOS rule base – the same software and rule based used to generate the data used to create the map of Figure 5. This generated a data set comprised of 57,935,114 index terms associated with 11,934 unique thesaurus terms across 10,402,493 documents from all of science.

What does one do with this much data? How does one create an intelligible and useful map using these data? A co-term analysis would quickly become unwieldy, particularly given that the counts per term span four orders of magnitude, and would likely result in a map that would be difficult to interpret. We settled on a much simpler approach that is facilitated by the fact that we already had a detailed map of Scopus content based on citation analysis.

We will not detail creation of the Scopus map here (see Figure 6a), but simply note that it is similar to the map of Boyack & Klavans (2014a), and contains over 20 million documents from Scopus. The 10.4 million documents indexed using the PLOS rule base are included in this map, and each is assigned to a cluster whose position on the map is known. Thus, the positions of all of these papers are known, and the positions of each of the 11,934 thesaurus terms can be calculated (triangulated) as the average position of the papers to which the term is indexed, as shown in Figure 6b. Now that the position of each thesaurus term is fixed, the Scopus map can be taken away (Figure 6c), and the PLOS thesaurus map stands on its own as a basemap that can be used as a template for overlays. Figure 6d shows the positions of all of the terms in the *Biology and life sciences* branch of the PLOS thesaurus. It is interesting to note, and of value to the thesaurus manager, that some of these terms appear in different parts of the map than expected. For example, the terms ‘circuit models’ and ‘electron transport chain’ are located in the engineering and physics sections of the map, even though they are in the biology section of

the thesaurus. This points out the fact that while these terms have a specific meaning in biomedicine, within the context of all of science these terms are overwhelmingly triggered by papers in engineering and physics. As PLOS context expands into these fields, the PLOS thesaurus managers may (or may not) choose to split the terms and create more specific rules designed to trigger different terms in different fields. The point is not if they will do this or not, but rather that a map such as that shown in Figure 6 gives the thesaurus manager information to enable decisions about future management of the thesaurus.

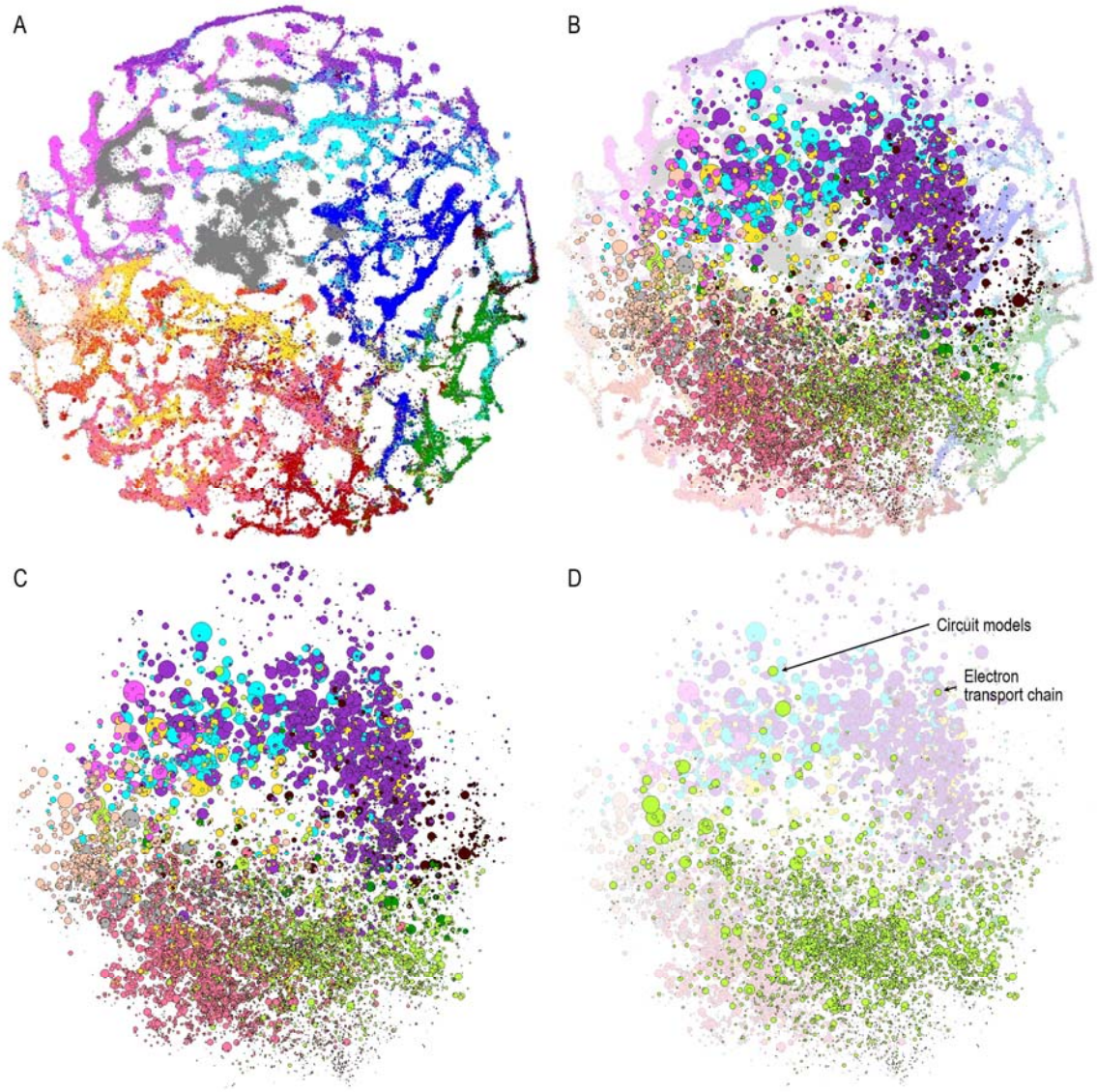

Figure 6. Creating a thesaurus map using a global map of science. A) Scopus map, B) overlay of PLOS thesaurus terms on the Scopus map, C) PLOS thesaurus terms become a basemap, D) overlay of biology terms on the PLOS thesaurus map. Node sizes reflect the numbers of Scopus papers (summed fractional counts) associated with each term. The color legend for the PLOS maps is the same as that used in Figure 5.

Note that the node sizes in the PLoS thesaurus map reflect the number of papers to which each term is assigned. Thus, the map can also be used to suggest terms that can be split to provide better differentiation of terms in the thesaurus.

Figure 7 shows another use for this thesaurus map. Two of the PLOS journals are compared – PLOS Biology and PLOS One. From these maps it is easy to see the differences in size, coverage and breadth of these two journals. It is also clear that PLOS One, although it is promoted as a multidisciplinary journal, is still primarily a broad biomedical journal. This thesaurus map could also be used to show growth in particular journals over time, or it could be linked to various metrics. For instance, by linking PLOS altmetrics (http://article-level-metrics.plos.org/alt-metrics/) to these terms at the article level, one could show the highest impact topics using an overlay of the terms associated with the top 1% tweeted or downloaded PLOS papers.

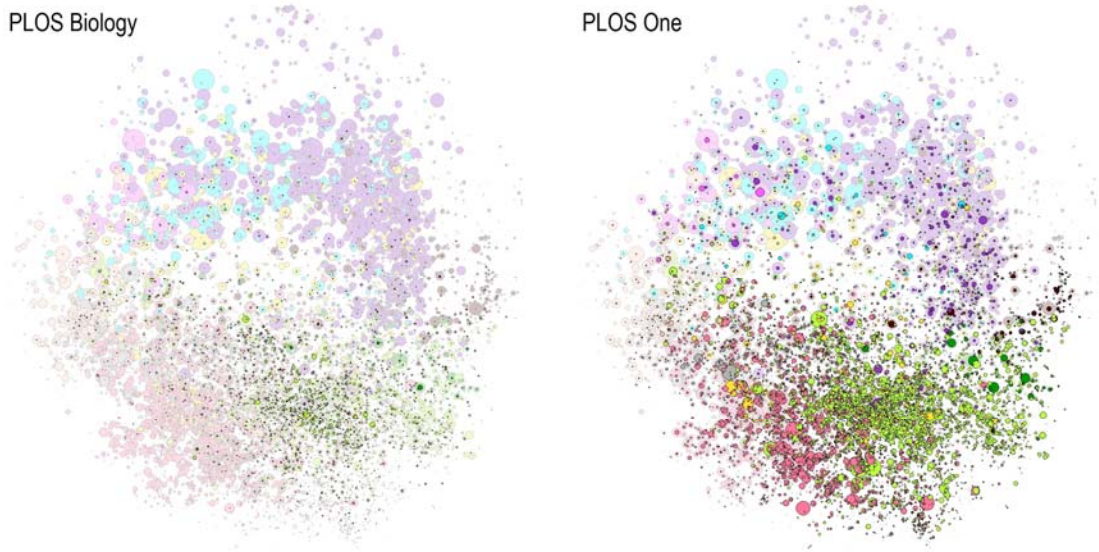

Figure 7. Overlays of PLOS Biology and PLOS One on the PLOS thesaurus map. The color legend is the same as that used in Figure 5.

Conclusion

This paper introduces two methods for creating maps of thesaurus terms that can then be used as templates or basemaps on which to display the contents of publication sets. The first example maps first-level terms from the Unified Astronomy Thesaurus (UAT) into a circular configuration, which can then be used to show relative positions of clusters of astronomy papers from different cluster solutions based on the thesaurus terms assigned to the papers in the clusters. The second example triangulates the entire thesaurus from the Public Library of Science (PLOS) onto a global map of science, and then uses the resulting map of thesaurus terms as the basis for an overlay map. This map can be used for several purposes, including mapping of subsets of PLOS content, and the identification of thesaurus terms whose rule bases may need to be changed.

We note that both visualizations are somewhat circular in construction. This is by design, because our experience is that circles are intuitively understood by most individuals. Using a clock metaphor, each hour on the clock roughly corresponds to a particular segment of the underlying data, and it is this high-level correspondence with data that makes such maps useful as templates. Despite the differences in the way these two maps were constructed, the main practical difference is in their level of detail (14 nodes vs. over 10,000 nodes). Each map is appropriate for its use case – the level of analysis needed (i.e., in terms of detail) typically

prescribes the level of detail needed in a visualization. We suggest that the methods described here should be generally applicable to other thesauri and use cases, and that care should be taken to make sure the level of visualization matches the level of detail required for the desired analysis.

Acknowledgments

Access Innovations, Inc. is gratefully acknowledged for indexing the Astro-set using the UAT thesaurus. Mapping of the PLOS thesaurus was done as part of work done for PLOS jointly with Access Innovations, Inc. Also, this paper benefitted greatly from reviews by Andrea Scharnhorst and two anonymous reviewers.

References

Akdag Salah, A., Gao, C., Suchecki, K., & Scharnhorst, A. (2011). Generating ambiguities: Mapping category names of Wikipedia to UDC class numbers. In G. Lovink & N. Tkacz (Eds.), *Critical point of view: A Wikipedia reader* (pp. 63-77). Amsterdam: Institute of Network Cultures.

Börner, K. (2015). *Atlas of Knowledge: Anyone can Map* (2nd ed.). Cambridge: MIT Press.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464.

Börner, K., Ma, N., Biberstine, J. R., Wagner, R. M., Berhane, R., Jiang, H., et al. (2010). *Introducing the Science of Science (Sci2) Tool to the Reporting Branch, Office of Extramural Research/Office of the Director, National Institutes of Health.* Paper presented at the Science of Science Measurement Workshop. Retrieved from http://www.nsf.gov/sbe/sosp/social/wagner-borner.pdf

Boyack, K. W., & Klavans, R. (2008). Measuring science-technology interaction using rare inventor-author names. *Journal of Informetrics, 2*, 173-182.

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 non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics, 8*, 569-580.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks - an introduction to co-word analysis. *Social Science Information, 22*(2), 191-235.

Dupuich, J., & Carr, G. (2013). Case study: Developing the PLOS thesaurus. *Bulletin of the American Society for Information Science and Technology, 39*(2), 22-25.

Hlava, M. M. K. (2015). *The Taxobook: Principles and Practices of Taxonomy Construction*: Morgan & Claypool.

Kay, L., Newman, N., Youtie, J., Porter, A., & Rafols, I. (2014). Patent overlay mapping: Visualizing technological distance. *Journal of the American Society for Information Science and Technology, in press*.

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. (2010). Toward an objective, reliable and accurate method for measuring research leadership. *Scientometrics, 82*(3), 539-553.

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics, 94*, 589-593.

Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics, 98*, 1583-1599.

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from Web-of-Science data. *Journal of Informetrics, 6*, 318-332.

Leydesdorff, L., Rafols, I., & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal-journal citations. *Journal of the American Society for Information Science and Technology, 64*(12), 2573-2586.

Leydesdorff, L., Rotolo, D., & Rafols, I. (2012). Bibliometric perspectives on medical innovation using the Medical Subject Headings of PubMed. *Journal of the American Society for Information Science and Technology, 63*(11), 2239-2253.

Petersen, A. M., Rotolo, D., & Leydesdorff, L. (2016). A triple helix model of medical innovation: Supply, demand, and technological capabilities in terms of Medical Subject Headings. *Research Policy, 45*(3), 666-681.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871-1887.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538.