Chapter 6

Visualizing the topical coverage of an institutional repository using VOSviewer

Introduction

Using text mining and visualization to identify, display, and analyze the topical coverage of large text corpora is increasingly common in a number of academic disciplines. This process, sometimes called bibliometric mapping, is fairly common in the field of library and information science. While its practical application in academic libraries is fairly new, it is conceivable that librarians could use these methods for a variety of purposes. This chapter will demonstrate the potential use of term co-occurrence maps, visualizations that demonstrate the relationships between highly occurring terms in a set of documents, as a means to understanding the scholarship archived in a library-run institutional repository. In these maps, terms are placed in a two-dimensional space so that terms that appear more often in combination with other terms are placed closer together. This process causes these frequently co-occurring terms to cluster together, and these clusters are interpreted as representing research areas present in this body of text. It is important to note that the computer simply recognizes rates of occurrence and co-occurrence, clustering terms together. It is incumbent on the person viewing the map to assign the meaning to these clusters. Nonetheless, these data visualization techniques provide a useful way to explore a set of documents, uncover latent patterns, and pose new questions to further analyze using additional methods.

As the push for open access to scholarship continues, libraries invest significant resources in setting up and maintaining institutional repositories. Term co-occurrence maps provide an opportunity to evaluate these services, beyond traditional metrics like download counts. Generating these maps from the titles and abstracts of items in a repository visually demonstrates how research clusters around specific areas across the sciences, social sciences, and humanities. This kind of analysis can help librarians determine whether a repository's content accurately represents the research output of an institution as a whole or if it is lacking in some key area. For example, a librarian might know that his or her institution is highly regarded for its active research in sociology, but upon analyzing the library's repository this librarian could find an absence of terms that indicate the presence of sociological research. It should be noted that the analysis of these maps is difficult and often requires consultation with subject-matter experts (Peters and van Raan 1993). However, this data-driven approach can complement what librarians already know about their repository, and combined with the input from subject matter experts, can provide insight into the way research happens at their institution. Armed with knowledge of the institutional research landscape, librarians can better perform outreach to faculty and communicate the value of institutional repositories as a key research service.

This chapter will outline the process for generating term co-occurrence maps from the titles and abstracts of items in ScholarWorks, the institutional repository at Indiana University-Purdue University Indianapolis (IUPUI). Term co-occurrence maps are created using VOSviewer, a freely-available tool for generating bibliometric maps (N. J. Van Eck and Waltman 2010). The resulting visualizations show clustering of relevant terms, representing the major research areas present in the repository's scholarship. An overview of how to export the necessary metadata from the repository, clean and prepare the data, along with a step-by-step guide to the visualization workflow is provided. The chapter will conclude with some discussion on interpreting term maps and specific ways librarians can use these maps to understand the research environment of their institution. The raw data used in this project, the R script used in cleaning and preparing the data, and the graph modeling language (GML) files for the resulting maps are all available on IUPUI DataWorks[1].

Background

Term co-occurrence maps (sometimes referred to as co-word maps or term maps) have a rich history in bibliometrics, a subfield of library and information science, which uses various methods to quantitatively analyze scholarly literature. Most often this analysis focuses on a specific domain in order to understand both its current state and its evolution over time. Term co-occurrence maps attempt to show the dominant themes in a set of documents by connecting terms that occur together in a single document. A document can be a paragraph, abstract, title, or the full text of an article. Term co-occurrences in a body of text are organized into a matrix, which is interpreted as a network, where terms are nodes connected by links based on their co-occurrence in a document. These maps are typically displayed in two-dimensions using a variety of techniques. Term maps date back to the early 1980s, with Callon et al.'s (1983) landmark study involving a co-word analysis of keywords from 172 scientific articles on dietary fibers. When mapped, terms are placed in a vertical fashion with more frequently occurring terms appearing at the top and co-occurrence represented by links connecting terms. Not all co-occurrences are represented in the map. In order to simplify the maps and reduce term density, a term must appear at least three times in association with one other term in the data to meet the threshold for inclusion in the map (Callon et al. 1983).

Subsequent term maps emphasized the strength of co-occurrence by using weighted links to connect the terms. The more frequently two terms co-occur, the thicker the link connecting the terms appears in the map. In their article, Rip and Courtial (1984) show the connections between keywords from articles published over a 10-year period in *Biotechnology and Bioengineering*, a core journal in biotechnology. Both circular and vertical maps are used to visualize the data. Similarity between terms is measured using the Jaccard Index and shown through weighted links (Rip and Courtial 1984). The circular maps used facilitate interpretation by placing the most highly occurring terns at the center of the map.

One of the major drawbacks of early term co-occurrence maps is the lack of objectivity regarding term placement on the map. Terms are situated in two-dimensional space in an ad hoc manner simply to facilitate ease of reading (Rip and Courtial 1984). The, arguably, intuitive assumption that distance between terms in the map corresponds to their similarity does not hold true. To address this shortcoming, multidimensional scaling (MDS), a method from spatial-data analysis, was introduced as a method for creating term maps. Using this approach, maps are generated where terms are automatically placed using computer software so the distance between terms reflects the rate of co-occurrence, resulting in highly co-occurring terms being placed in close proximity, forming clusters of similar terms (Tijssen and Van Raan 1989). Ultimately this approach yields maps that are more intuitive that previous term co-occurrence maps. However, map readability, especially for larger term maps, still proves challenging due to overlapping term labels and link density.

More recently, computer programs like VOSviewer enable the analysis of much larger bodies of text and increase map readability simultaneously through improvements in term placement. At the heart of the tool is a mapping technique referred to as visualization of similarities (VOS), which differs from prior methods for term placement. The VOS method improves on multidimensional scaling by locating terms closer to their ideal coordinates on the map and by giving weight to indirect similarities (N. van Eck and Waltman 2007). Additionally, previous tools for visualizing term co-occurrence maps, such as SPSS or Pajek, suffer from problems of label overlapping labels and lack ways to explore small portions of the map in any detail (N. J. Van Eck and Waltman 2010). The VOSviewer program is highly flexible. The tool can read data directly from Web of Science or Scopus, allowing users to generate term maps from article abstracts, or it can read text files, allowing for the creation of term maps from any text. Users can employ the VOS mapping method to create maps from a dataset in the tool itself, or view maps created using multidimensional scaling in other programs such as SPSS (N. J. Van Eck and Waltman 2010). Once maps are created, either natively or in another tool, VOSviewer provides two ways to visualize the data: the network visualization view or the density visualization view. In the network visualization view terms are presented by labels on top of circles. The size of the label and circle corresponds to the overall frequency in the dataset. The color of the circle corresponds to the cluster to which the term has been assigned. In the density view, terms are represented by labels which, again, correspond to frequency in the dataset. The color in the density view ranges from blue (lowest density) to red (highest density). These color values are determined by the number of nearest terms in the area around a point and the weight, or relative frequency in the case of term co-occurrence maps, in the dataset (N. J. Van Eck and Waltman 2015). Each view offers users a unique way to uncover patterns in the data. Additionally, users can view small portions of the map by using a zoom and scroll functionality. Finally, the tool also offers the ability for screenshots of maps and the ability to save both image and map files in variety of formats.

While VOSviewer was initially designed to create bibliometric maps like journal citation maps, it performs well as a text-mining tool for creating term co-occurrence maps, easily ingesting large amounts of text. Creating a term co-occurrence map in VOSviewer involves four steps. In the first step, the tool identifies noun phrases, which are word sequences consisting of only nouns and adjectives, via part of speech tagging using the Apache OpenNLP toolkit (N. J. Van Eck and Waltman 2011). In the second step, VOSviewer identifies relevant terms, which ultimately reduces clutter in the resulting map. In order to determine a term's relevance, the tool filters out more general noun phrases by comparing certain noun phrases that co-occur only with a limited set of other noun phrases versus those noun phrases that co-occur with many different noun phrases (Waltman, van Raan, and Smart 2014). The third step involves mapping and clustering the terms using the VOS mapping technique combined with a modified modularity-based clustering approach (Waltman, van Eck, and Noyons 2010). Finally, the map is displayed in both the network visualization view and the density visualization view.

VOSviewer has recently gained popularity for its ease of use, the intuitive maps in generates, and its scalability. The tool has been used this to study the evolution of scholarship in academic domains as diverse as land use and urban planning (Gobster 2014) to computer and information ethics (Heersmink et al. 2011). The tool is also adept at illuminating connections between research areas in highly interdisciplinary fields, such as the interface between engineering and physical sciences with health and life sciences (Waltman, van Raan, and Smart 2014). Due to VOSviewer's easy-to-use interface, its ability to ingest large volumes of text, and its utility in showing connections in highly interdisciplinary areas, it is a good tool for analyzing the topical coverage of an institutional repository.

Example Project

Project Background

This project began in early 2015 as a way to understand the current state of IUPUI's institutional repository, ScholarWorks. The first item was deposited in the repository, which at the time was named IDeA (IUPUI Digital Archive), in August 2003 (Odell 2014). The first instance of IUPUI's repository ran on the first version of DSpace, which was release the year before. Early adopters on campus included the School of Medicine, University Library, and Herron School of Art and Design (Staum and Halverson 2004). Over the years the repository has grown and been organized into different communities, with some of the original communities subsumed as collections into larger communities. At the time of this study, ScholarWorks archives over 4,000 unique items and hosts 25 different communities, spanning the sciences, social sciences and humanities (see table 6.1).

Table 6.1 ScholarWorks communities and number of items

ScholarWorks Community	Number of Items
Theses, Dissertations, and Doctoral Papers	1255
School of Medicine	1136
Faculty Articles	858
University Library	772
School of Liberal Arts	467
Office of the Vice Chancellor for Research	286
School of Informatics and Computing	241
Robert H. McKinney School of Law	214
Lilly Family School of Philanthropy	175
School of Education	142
School of Public and Environmental Affairs	78
School of Science	70
Herron School of Art and Design	64
School of Engineering and Technology	55
School of Dentistry	49
Richard M. Fairbanks School of Public Health	41
Moi University/IUPUI Partnership	38
School of Nursing	37
Kelly School of Business – Indianapolis	26
Indiana University-Purdue University Columbus	23
Center for Service Learning	17
School of Rehabilitation Sciences	12
School of Physical Education & Tourism Management	11
School of Social Work	8
Alumni Works	5

Initially the project was undertaken as a proof of concept, but it was also done with an eye toward the future. One of the goals of this project is to serve as a baseline against which to assess the evolution and growth of ScholarWorks as a repository. This study proves timely due to the recent passing of a campus-level open access policy. In October 2014, the IUPUI Faculty Council passed an open access policy, encouraging faculty and researchers to make their scholarship as openly available as possible ("Open Access Policy, IUPUI Faculty Council (October 7, 2014) | Open Access @ IUPUI" 2015). While self-archiving is not mandated by the policy (researchers are able to opt out on an article-by-article basis), a significant component of the work involved in implementing the policy centers on an aggressive outreach program aimed at helping faculty and researchers self-archive their journal articles in ScholarWorks. Due to an increase in this work, the number of submissions to the repository is expected to expand its coverage significantly in the coming years. Thus, studying the dominant research themes of items archived in the repository at this point is an important first step in assessing future expansion of repository coverage.

Obtaining and Cleaning the Data

This project analyzes the abstracts and titles of items in the repository. Each title and abstract are considered to be distinct documents in this corpus. Using titles and abstracts as the units of analysis is preferable to using keywords or subject terms due to the higher prevalence of titles and abstracts in the data. Submitting an item to the repository involves filling out a series of web forms, which populate Dublin Core Metadata fields on the repository backend. In order allow for flexibility in the submission process the only metadata requirements are the provision of a date and a title. Additionally, records cannot be created in the repository without a file. The flexibility in submission process is useful, but results in incomplete metadata for many items.

However, the fact that item title is a required field in the submission process ensures that at least some text is associated with each item in the repository. Ultimately, using titles and abstract for topical analysis results in a more complete dataset than using keywords or subject terms.

In ScholarWorks, metadata files are available for export at the community level to users with administrative privileges. A comma separated value (CSV) file for each of ScholarWorks' 25 communities is exported. Each community CSV file contains the standard Dublin Core elements, using various properties, to describe community content. Obviously, the abstract and title are needed for this analysis, but the item ID is also used for de-duplication, as an item's membership in a ScholarWorks community is not mutually exclusive (more on the de-duplication process later). Each CSV file is opened in Microsoft Excel to check data integrity. It is immediately apparent that the level of specificity used to describe an item varied greatly both within and across communities. This variation stems from the submission process where users have lists of options for describing an item via dropdown menus. For example, when selecting the language for an item, users can select *English* or *English (US)*. Ultimately, these differences result in varying levels of consistency in metadata both across and within repository communities, resulting in the element dc.description.abstract[en] being used to describe one item, while dc.description.abstract[en_US] is used to describe another. A similar problem occurs with the titles for items as well. To address this inconsistency, the Excel concatenate function is used to combine the columns across which abstracts and titles are spread into a new column in each file titled *abstract.combined* and *title.combined*. After combing abstracts and titles into one column, each file is saved as a separate CSV file.

The next stage in preprocessing involves loading the data into R for further cleanup. Using a simple R script is seen as preferable to performing the rest of the cleanup in Excel due to the size of some of the files and the fact that scripting the cumbersome cleanup process reduces the chance for human error. Using the script, each CSV file is loaded into R. Then the IDs, combined abstracts, and combined titles are extracted from each file and saved as vectors. These vectors are then combined into subsets of the original files. Each subset is then combined into one data frame containing the IDs, abstracts, and titles from all 25 CSV files. The item ID is used to de-duplicate the dataset and the unique titles and abstracts are saved as character vectors. Finally, the character vectors are written to two separate text files, one containing unique abstracts and the other containing unique titles. These files are then manually checked and combined into one file using a text editor. At this stage the file is ready for visualization using VOSviewer. The next section provides a step-by-step overview of the visualization process.

The Visualization Workflow

Creating term maps in VOSviewer is a relatively easy process. The first step is to download and install the tool, which is freely available from http://www.vosviewer.com/.

1. Launch the program and select ***create*** from the action panel menu on the left of the tool. A popup will appear, select ***Create a map based on a text corpus***.
2. Choose the text file with the abstracts and titles. Load that file as a ***VOSviewer corpus file***. It is not necessary to use a ***VOSviewer scores file***.
3. Set counting method to ***binary***. This is preferred over full counting, especially for larger bodies of text. Full counting uses every instance of a term in a document to assess its similarity to others, while binary counting only uses the presence of the term. This prevents the maps from being skewed by a single term appearing frequently within one document.

4. Ignore the ***thesaurus file***. This file will eliminate certain noun phrases from the final map. Terms can always be deselected at a later state, but supplying a thesaurus at this step can be helpful in eliminating potentially non-meaningful terms, such as *results* or *methodology*, from the resulting map.

5. Set the ***minimum occurrence threshold***. By default, VOSviewer uses a threshold of 10, which works well for fairly large datasets. The total number of terms in the ScholarWorks dataset is 75,134 terms. Using a minimum occurrence threshold of 10, the dataset is pared down to 1801 terms.

6. VOSviewer assigns relevance scores to each term. The distribution of second-order co-occurrences of a single noun phrase over all noun phrases is compared with the overall distribution of noun phrases over all noun phrases, and the greater the difference between these two distributions the more relevant the term is considered to be (N. J. Van Eck and Waltman 2011). This significantly reduces the number of terms to 60% of the terms above the selected threshold. For the ScholarWorks data, reducing the terms to the most relevant 60% results in 1081 terms.

7. Verify selected terms and de-select any non-meaningful terms outside the scope of analysis. Clicking on the column heading for ***Occurrences*** or ***Relevance*** allows for the sorting of these terms in either ascending or descending order. Sorting by the most frequently occurring terms facilitates the removal of non-meaningful terms from the map. For example, frequently occurring terms like *article* could be removed from the analysis. This ultimately makes the map easier to read and highlights meaningful

relationships between the terms. Generally, term deselection is done in an ad hoc fashion and will vary depending on the data and goals of the project. For the initial exploratory analysis of the ScholarWorks data, no terms were deselected.

8. Click ***finish*** and VOSviewer performs mapping and clustering. Term co-occurrence maps created from text files are available to view in either the ***Network Visualization*** or ***Density View***. To change between views, click on the tabs at the top of the main panel in the center of the tool.

9. Changing the ***clustering resolution*** increases or decreases the number of clusters in the map, which can help uncover patterns in the data. To change this parameter, click on the ***Map*** tab in the action panel on the left of the tool. By default the clustering resolution is set to 1.0. Increasing this number produces more clusters in the map and decreasing reduces the number of clusters.

Results

The initial map shows six clusters of terms in the *Network Visualization* view (see figure 6.1). The red cluster to the left of the map includes terms associated with social science and humanities disciplines, the green and blue clusters to the right include science-related terms, and the yellow cluster that connects the two areas has many public health-related terms (see table 6.2). These four clusters will be examined in detail, later. However, it is worth analyzing the remaining two clusters. The purple-colored cluster in the upper right of the map contains terms that could not easily be assigned to one of the other clusters. This occurs for two reasons. First, general terms, such as *period*, appear in many titles and abstracts, but do not co-occur frequently enough with any other specific terms to be assigned to either of the other clusters. Second, terms in this cluster, such as *attorney general* and *opinion*, are highly specific to a set of items within the repository. In the case of *attorney general*, *opinion*, and *official opinion,* these terms refer to a historical set of digitized opinions from the Indiana Attorney General. Other terms, such as *digital aerial photography, county, accuracy*, and *report* are all associated with a set of county horizontal accuracy reports, which provide aerial photographs of Indiana counties. Due to the uniformity of the titles and lack of additional text that might associate them with their respective disciplines, law and geography, these items are clustered together.

Table 6.2 Top five most frequently occurring terms from each cluster

Term	Occurrences	Cluster	Color
student	568	Social Science & Humanities	Red
cell	441	Molecular Biology & Genetics	Green
function	412	Molecular Biology & Genetics	Green
experience	376	Social Science & Humanities	Red
program	371	Social Science & Humanities	Red
library	353	Social Science & Humanities	Red
community	334	Social Science & Humanities	Red
mechanism	329	Molecular Biology & Genetics	Green
protein	327	Molecular Biology & Genetics	Green
expression	292	Molecular Biology & Genetics	Green
property	198	Other Sciences & Dentistry	Blue
concentration	169	Other Sciences & Dentistry	Blue
teeth	166	Other Sciences & Dentistry	Blue
score	165	Public Health	Yellow
agent	144	Other Sciences & Dentistry	Blue
surface	143	Other Sciences & Dentistry	Blue
diabetes	99	Public Health	Yellow
predictor	92	Public Health	Yellow
reliability	89	Public Health	Yellow
item	86	Public Health	Yellow

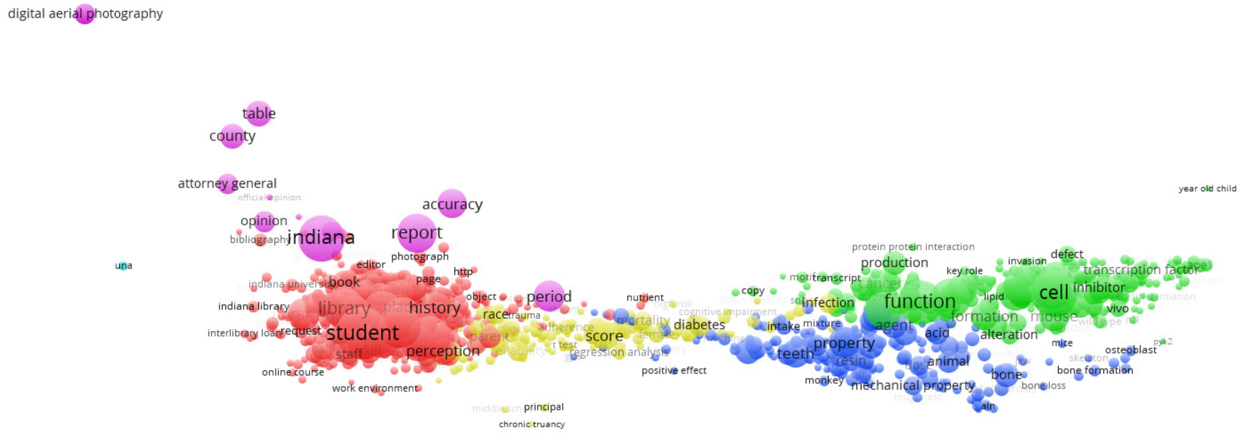

Figure 6.1. ScholarWorks term map with six term clusters.

The light blue cluster consisting of two terms, *una* and *cultura*, represents a small number of Spanish language items in the repository, all of which are found in the Theses, Dissertations, and Doctoral Papers community. VOSviewer is designed for data in English and cannot perform part of speech tagging on other languages, which is why the article *una* made it through to the map and was not excluded during stopword removal. However, the presence and clustering of these terms suggests some possibility for a basic language-based map for multilingual repositories. Due to the limited number of foreign-language materials in ScholarWorks, this type of analysis is beyond the scope of this study.

The largest cluster is the humanities and social science cluster at the left of the map, including 478 terms (see figure 6.1). Upon initial review, the terms that stand out the most include *student*, *program*, *experience*, and *library*. It is not really surprising that library-related terms figure so prominently in this cluster. The University Library community is the fourth largest in ScholarWorks, which is likely due to the fact that librarians are more aware of this service and are often advocates for open access. However, it is interesting that despite its relatively small size, especially when compared to the School of Medicine and Theses, Dissertations, and Doctoral Papers communities (see table 6.1), that terms from this community dominate the map. This suggests the presence of a large amount of library-related research in the repository, or that these items use similar language to describe the research.

Switching to the density visualization view provides more information on the overall structure of the map (see figure 6.2). It is immediately apparent that the highest term density occurs at the center of the social science and humanities cluster. The highest density area centers on the term *student*, which makes sense given that it is the most frequently occurring term in the dataset. The next two highest areas of term density occur in the science clusters, centered on the terms *cell* and *function.* The area connecting the science clusters with the social science and humanities clusters, containing public health terms, has a relatively low term density compared to the rest of the map.

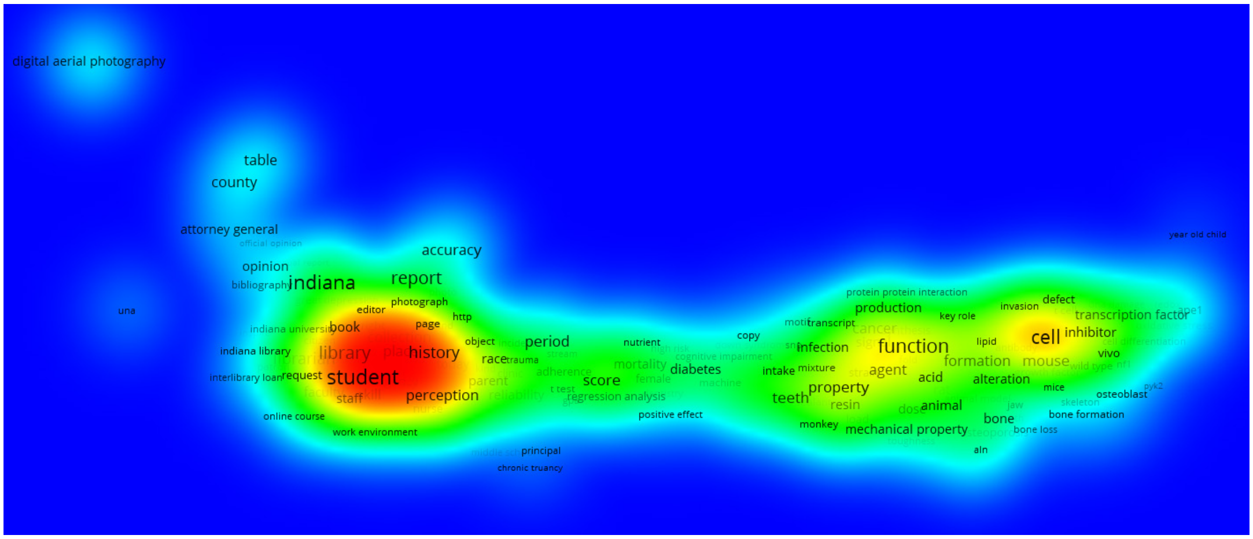

Figure 6.2. ScholarWorks term map in Density Visualization view.

To examine the social science and humanities cluster more closely, the clustering resolution is increased in VOSviewer to provide a more granular view. The default clustering resolution of 1.0 does not provide much detail (see figure 6.3). However, changing this parameter to 2.0 yields a map with sufficient granularity to see different research areas (see figure 6.4).

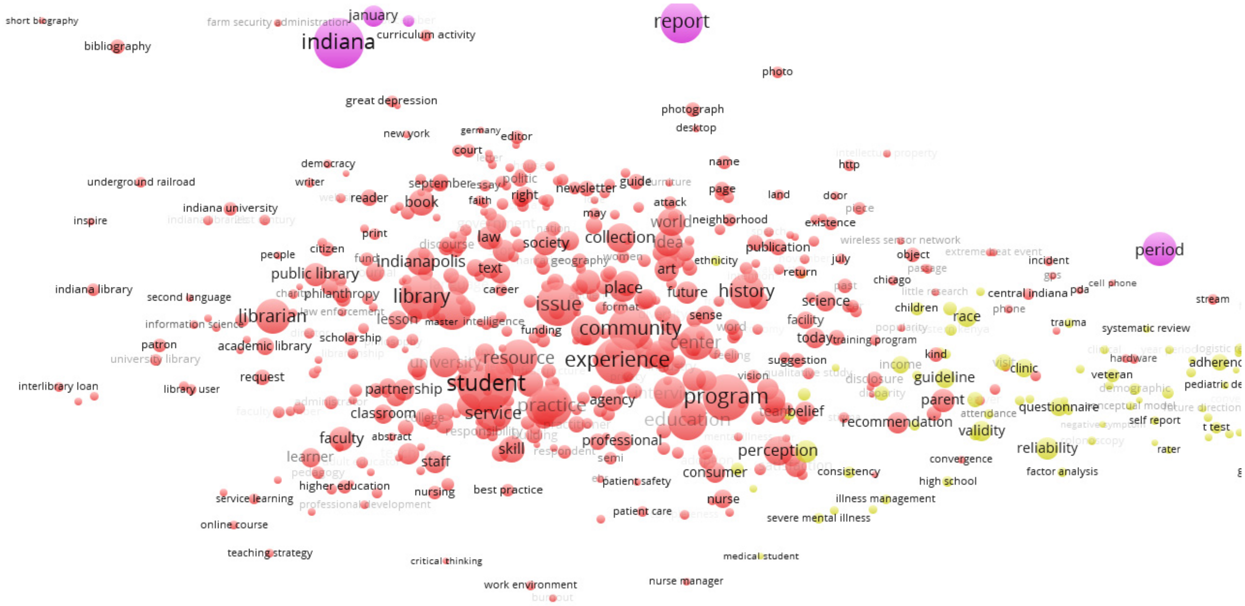

Figure 6.3. Social science and humanities terms at 1.00 cluster resolution

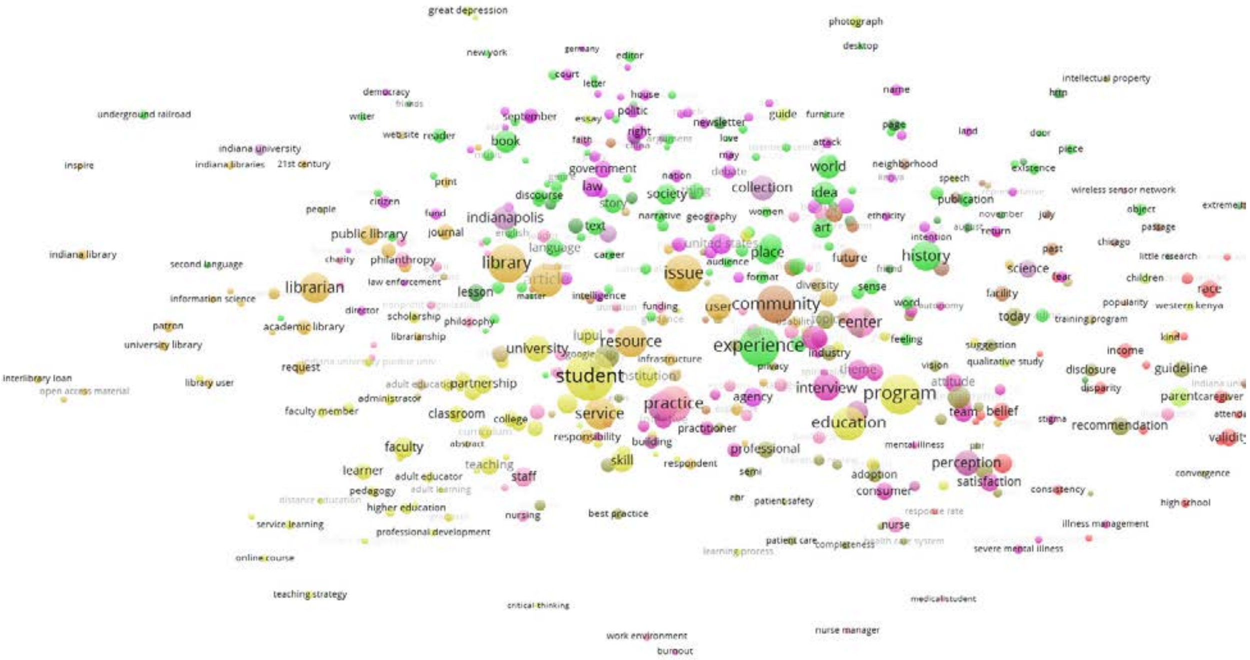

Figure 6.4 Social science and humanities terms at 2.00 cluster resolution

There are now four prominent sub-clusters present. The largest of these sub-clusters is the arts and humanities (green) and is spread across the upper portion of the map. Within this sub-cluster the most frequently occurring terms are *experience, history, place, world,* and *idea.* It is important to note, that while the terms *experience* and *history* appear in this sub-cluster, they are centrally located on the map, suggesting their use as terms in a variety of items across the social sciences and humanities and providing an example of how VOSviewer handles indirect similarities. The next biggest sub-cluster includes terms that are related to the scholarship of education (yellow) in the lower left of the social science and humanities cluster. The most frequently occurring terms in this cluster include *student, program, education, opportunity,* and *university.* It is interesting to note the overlap between this sub-cluster and the adjacent library research sub-cluster (gold) above the scholarship of education sub-cluster. In fact, the term *information literacy*, which is too small to appear in figure 6.4 but can be seen in figure 6.5, spans the boundary between these two sub-clusters. The library research cluster is dominated by terms that include *article, resource,* and *service.* The last sub-cluster within the social science and humanities cluster is *government, public policy*, and *law,* which can be seen in purple at the top of the social science and humanities cluster. The most frequently occurring terms in this cluster include *United States, law, opinion, government,* and *right.*

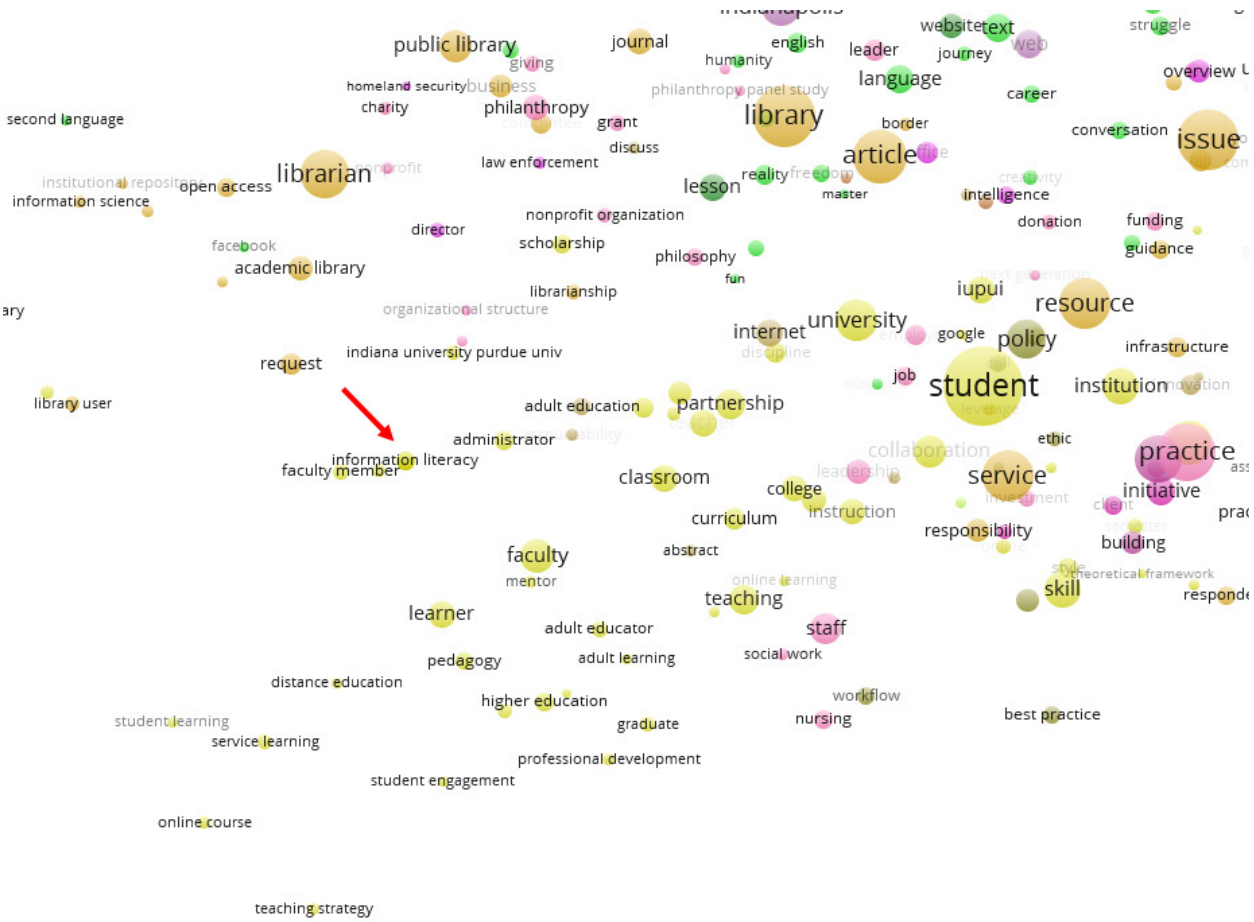

Figure 6.5. Information literacy term appears at the boundary between education-related research and library-related research.

The right side of the term map is dominated by the two science clusters, which include the biophysics and dentistry cluster (blue) and the molecular biology and genetics cluster (green). Examining the structure of the two clusters yields nothing unexpected. For example, the term *mechanical property* appears toward the bottom of the biophysics and dentistry cluster, far away from terms such as *protein protein interaction*, which occurs at the top of the molecular biology and genetics cluster due to a high level of dissimilarity (see figure 6.1). Conversely, highly-similar terms such as *disease* and *resistance* occur at the boundary between these two clusters. To identify further patterns, the clustering resolution is changed. Increasing the clustering resolution parameter to just 1.5 results in a clearer distinction between the dentistry-related terms (purple) and biophysics terms (light blue) to their right, which include mostly bone-related research (see figure 6.6). To confirm the relative large amount of bon-related research, a quick keyword search is done in ScholarWorks for the term *bone*, returning 761 results.

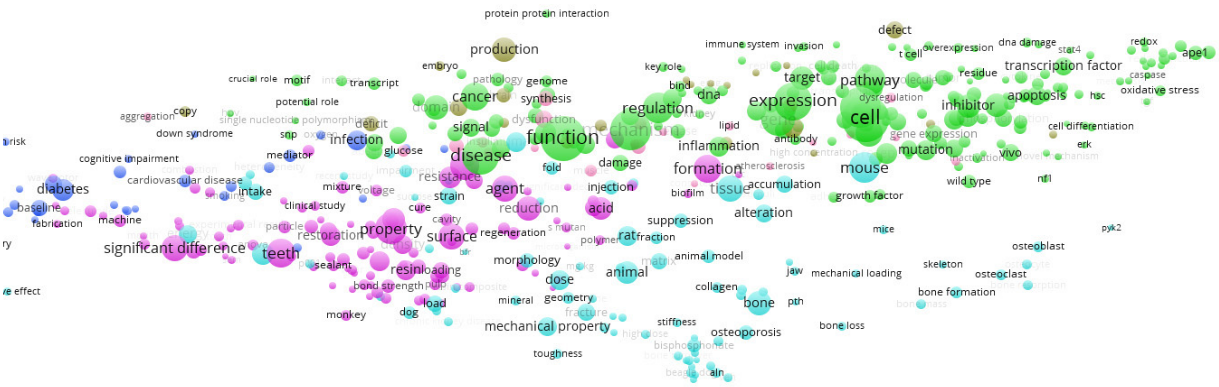

Figure 6.6 Science-related terms with clustering resolution of 1.5

Even at this level of clustering, all the molecular biology and genetics terms appear clustered together, represented by the green colored terms (see figure 6.6). Increasing the clustering resolution to 2.0 produces higher granularity, but without validation by a subject matter expert it is difficult to identify any meaningful sub-clusters or patterns in the data (see figure 6.7). However, even with expert input, this research area could still lack any easily identifiable clusters of terms, due either to the relatively small amount of data or the diversity of research in this area.

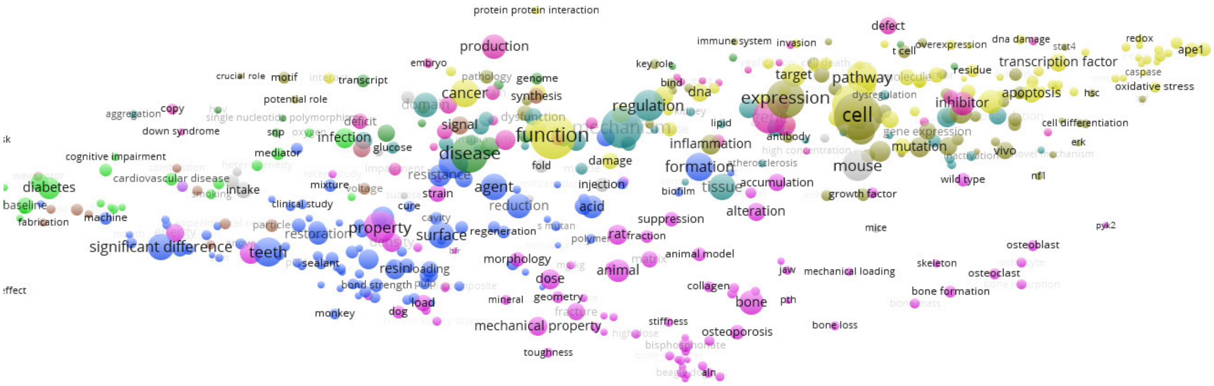

Figure 6.7 Science-related terms with clustering resolution of 2.0

Perhaps the most interesting feature of the map is the cluster that connects the three clusters of social science and humanities, biophysics and dentistry, and molecular biology and genetics. The yellow cluster that bridges the sciences with the social sciences and humanities contains many public health-related research terms. This cluster is the most widely dispersed in the map, with terms scattered among the social science and humanities cluster, and the two sciences cluster. In total, the public health cluster contains 145 terms, which includes frequently occurring terms such as *diabetes*, *predictor*, *mortality*, *depression*, and *race*. There are also a number of terms that indicate the heavy use of surveys as a data collection method, such as *score*, *item*, and *questionnaire*.

Probably the most interesting feature of the public health cluster is where it intersects with the other clusters on the map. As an interdisciplinary field, there is a lot of overlap between public health and other areas. At the intersection of the public health cluster with the social science and humanities cluster, terms that indicate health economics research, such as *consumer*, *patient care*, and *health care system*, are found. Additionally, terms such as *race*, *income*, and *disparity* are found at this edge of the public health cluster and the social science cluster, indicating the presence of sociological and public policy health-related research. On the opposite side of the public health cluster, terms that are more often associated with health-related research in the sciences are found. Terms such as *smoking*, *cardiovascular disease*, and *infection* intermingle with the terms in the two science clusters.

Discussion

The distribution of term densities across the map is interesting and somewhat unexpected. The relative high density of terms in the social science and humanities cluster was surprising, given that the majority of research at IUPUI is happening in medicine and health sciences. When the two science clusters are combined, they total 442 terms, which is roughly similar in size to the social science and humanities cluster, with 478 terms. However, the density of terms appears far greater in the social science and humanities cluster. This raises interesting questions about the research that is archived in these areas. Perhaps research in the social sciences and humanities has a more limited set of terms with which to describe the research being done. Or perhaps the research archived in ScholarWorks in the social sciences and humanities is more on similar topics like student engagement. Whatever the case, it appears that the research in the sciences that is archived ScholarWorks is more diverse than the research in the social sciences and humanities, at least based on the terms used to describe this research. This difference represents an area where ScholarWorks may not accurately reflect the research landscape of the institution and is something librarians should give consideration. Those librarians serving faculty in the social sciences and humanities should take steps to ensure the full range of research happening in their departments is accurately reflected, if possible.

The overall structure of the map provides further insight into the connections between major research areas. As mentioned earlier, IUPUI is a campus with a strong emphasis on the health sciences, and as such it is unsurprising to see so many health-related terms scattered throughout the map. In this way, the term map serves as an apt metaphor for campus, with researchers focusing on health-related issues physically spread across campus in various departments. Furthermore, it is interesting to see how distinctly the public health cluster bridges the gap between the social science and humanities cluster with the two science clusters, providing evidence for the highly interdisciplinary nature of public health research. However, one of the major challenges in this project reveals itself in the structure of the map. The small collections of specific items, usually with uniform titles such as the Opinions of the Attorney General of Indiana collection in the Robert H. McKinney School of Law community, create separate clusters not connected to the rest of the map and interpreting the map difficult. If the viewer is unaware of these collections and their uniform titles that increase the frequency of certain words, he or she might lend too much weight to the importance of these clusters. While these clusters do provide important insight into the contents of the repository, they distract from the more interesting relationships between the researches areas that are depicted in the rest of the map. Therefore, librarians engaged creating these types of term maps should have some basic level of familiarity with the contents of their repository, and, as should always be the case, approach the resulting maps with a critical eye. Another challenge related to the structure of the map and cluster formation, pertains to the way bodies of text containing many different research areas do not always form coherent clusters. While VOSviewer can show the connections between interdisciplinary areas of research, it relies of sufficient high-quality data. The ScholarWorks dataset needs to be larger in order to more accurately see the relationship the research areas present.

Despite the relatively small amount of data, there are many groups of terms in the clusters that point to easily identifiable research areas. Some of the more prominent terms provide clues about institutional values, or at least the values of those actively engaged in supporting the repository. For example, terms related to student engagement and educational research figure prominently in the social science and humanities cluster. Much of this research is archived in the Center for Service Learning community. However, it is interesting to compare the prevalence of these terms with the relatively small size of the community, suggesting that these are terms used throughout the social science and humanities cluster. This pattern meshes well with many of IUPUI's institutional values, which prize student engagement and student learning as key values. Similarly, the health-related research across the disciplines and not just in the health sciences is strongly indicative of IUPUI's culture. Programs such as Medical Humanities & Health Studies[2] and new degrees such as the Ph.D. in Health Communication[3] mean that health-research terms show up in unexpected places, as evidenced by the many health-related terms at the bottom left of the social science and humanities cluster. However, these terms do not form into any easily identifiable clusters, due in equal parts to the small number of items in these research areas and the difficulty in clustering interdisciplinary research. One of the limitations of using term co-occurrence maps to draw conclusions about the nature of research archived in an institutional repository is how susceptible they are to individual researchers with many items on the same topic. For example, much of the bone-related research in the biophysics sub-cluster (see figure 6.7) is attributable to one researcher at the university. The 'repeat customer' phenomenon can make it seem as though there is a lot of research being done institutionally in a particular area, when in reality there are 10 articles from one researcher on a single topic. Again, accurate interpretation of these maps relies heavily on a knowledge of the repository's contents.

There are a number of areas noticeably absent from the ScholarWorks term map. Given the strong presence of an engineering program on campus, it is surprising to see the lack of an engineering cluster or at least a significant number of engineering-relate terms. Another gap in the map is in the area of physics. These gaps are confirmed by consulting the repository. There is only one item archived in the Physics collection within the School of Science community and the School of Engineering and Technology community only has 55 items. Further gaps include math, chemistry, and chemical biology. The lack of chemistry-related research is unsurprising due to issues around research-related patents and trepidation towards open access. Despite the lack of some areas in the map, there are small clusters of terms that suggest emerging areas in the repository. Identifying a potential emerging areas requires a general knowledge of the institution and its research. One potential emerging area at IUPUI is in Philanthropy, with the recent founding of the Philanthropic Studies program in the Lilly School of Philanthropy. Terms related to this emerging area appear in the social science and humanities cluster just above the library-related terms, including *philanthropy, giving, grant*, *fund,* and *nonprofit organization.*

Conclusion

This chapter demonstrates how librarians can visually represent the research archived in library-run institutional repositories using term co-occurrence maps. Specifically, these maps demonstrate how different research clusters around themes in the sciences, social sciences, and humanities. Somewhat unexpectedly, the highest density of terms appears in the social sciences and humanities, followed by the sciences. These two sections of the map are connected by public health. This map serves as a valuable resource to subject librarians in two primary ways. First, the map charts the research landscape of the institution, showing connections that while obvious to some, are new to others. For example, some librarians may be unaware just how pervasive health-related research is on IUPUI’s campus, showing up in social science and humanities research, as well as in the sciences. Second, the map identifies gaps in the repository’s coverage.

One prominent example, is the relatively small amount of scientific research outside of the health sciences. Many of these gaps are evident when looking directly at the numbers of items in the collections that make up the ScholarWorks communities, but visualizing the entire repository as one term map brings these gaps into context.

The two biggest limitations of these term maps are the relatively small dataset and the necessary reliance on subject-matter expert input for interpretation. These maps are made with the titles and abstracts from 4346 items, which is a relatively small amount of data for this type of large scale textual analysis. Furthermore, the relatively small amount of data makes these term maps susceptible to being skewed by small special collections with uniform titles, such as the Opinions of the Indiana Attorney General, and single researchers who have a number of articles on the same topic. However, as the repository expands in size it will be less vulnerable to being skewed and more accurately reflect the institution's research landscape. Additionally, input from subject matter experts will result in a more comprehensive analysis. Many librarians lack the specialized knowledge to connect clusters of terms with the research areas these terms potentially represent. For the ScholarWorks term map, this is especially true in the sciences where a lack of expert knowledge only allows for the general classification of clusters as dentistry, biophysics, and molecular biology and genetics.

Future iterations of this project will need to include an interpretation and validation phase that involves input from faculty or other subject-matter experts on cluster identification. This input will facilitate librarians' understanding of the map and improve everyone's understanding of the research landscape at IUPUI. Furthermore, a much larger high quality dataset will improve the resulting map. As more time passes since the implementation of the campus-level open access policy and librarians work to mediate submissions of faculty research, the amount of text in the repository for analyzing will only continue to grow. Replicating these term maps in a year or two years will yield a much fuller picture of the research landscape and potentially provide insight into new and emerging research areas on campus. Despite the drawbacks of the ScholarWorks term maps, they are still useful for librarians planning outreach around the open access policy. With these term maps in mind, librarians should focus on increasing the diversity of social science research beyond library and education research and increase the repository's holdings in scientific research beyond the health sciences. Lastly, these maps have the potential for helping librarians, particularly those new to campus, to begin to chart the research and intellectual landscape at their institutions.

References

Callon, Michel, Jean-Pierre Courtial, William A. Turner, and Serge Bauin. 1983. "From Translations to Problematic Networks: An Introduction to Co-Word Analysis." *Social Science Information* 22 (2): 191–235. doi:10.1177/053901883022002003.

Gobster, Paul H. 2014. "(Text) Mining the LANDscape: Themes and Trends over 40 Years of Landscape and Urban Planning." *Landscape and Urban Planning* 126 (June): 21–30. doi:10.1016/j.landurbplan.2014.02.025.

Heersmink, Richard, Jeroen van den Hoven, Nees Jan van Eck, and Jan van den Berg. 2011. "Bibliometric Mapping of Computer and Information Ethics." *Ethics and Information Technology* 13 (3): 241–49. doi:http://dx.doi.org/10.1007/s10676-011-9273-7.

Odell, Jere. 2014. "Building, Growing and Maintaining Institutional Repositories." presented at the Michiana Scholarly Communication Librarianship Conference, IUSB, South Bend, IN, October 20.

"Open Access Policy, IUPUI Faculty Council (October 7, 2014) | Open Access @ IUPUI." 2015. Accessed May 20. https://openaccess.iupui.edu/policy.

Peters, H.P.F., and A.F.J. van Raan. 1993. "Co-Word-Based Science Maps of Chemical Engineering. Part I: Representations by Direct Multidimensional Scaling." *Research Policy* 22 (1): 23–45. doi:10.1016/0048-7333(93)90031-C.

Rip, Arie, and J. Courtial. 1984. "Co-Word Maps of Biotechnology: An Example of Cognitive Scientometrics." *Scientometrics* 6 (6): 381–400.

Staum, Sonja, and Randall Halverson. 2004. "IDEA: Sharing Scholarly Digital Resources." IUPUI, Indianapolis, IN, February 27.

Tijssen, R., and A. Van Raan. 1989. "Mapping Co-Word Structures: A Comparison of Multidimensional Scaling and LEXIMAPPE." *Scientometrics* 15 (3-4): 283–95.

van Eck, NeesJan, and Ludo Waltman. 2007. “VOS: A New Method for Visualizing Similarities Between Objects.” In *Advances in Data Analysis*, edited by Reinhold Decker and Hans-J. Lenz, 299–306. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-70981-7_34.

Van Eck, Nees Jan, and Ludo Waltman. 2010. “Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping.” *Scientometrics* 84 (2): 523–38. doi:10.1007/s11192-009-0146-3.

———. 2011. “Text Mining and Visualization Using VOSviewer.” *arXiv:1109.2058 [cs]*, September. http://arxiv.org/abs/1109.2058.

———. 2015. “VOSviewer Manual (Version 1.6.0).”

Waltman, Ludo, Nees Jan van Eck, and Ed C.M. Noyons. 2010. “A Unified Approach to Mapping and Clustering of Bibliometric Networks.” *Journal of Informetrics* 4 (4): 629–35. doi:10.1016/j.joi.2010.07.002.

Waltman, Ludo, Anthony F. J. van Raan, and Sue Smart. 2014. “Exploring the Relationship between the Engineering and Physical Sciences and the Health and Life Sciences by Advanced Bibliometric Methods.” *PLoS ONE* 9 (10): e111530. doi:10.1371/journal.pone.0111530.

[1] http://hdl.handle.net/11243/9

[2] http://liberalarts.iupui.edu/mhhs/

[3] http://liberalarts.iupui.edu/comm/