Moving Beyond Borders with Business Information:  
One Book’s Transformation from  
Print to Network Visualization  

Carissa M. Phillips, MBA, MSLIS  
Associate Professor and Business and Finance Information Librarian  
University of Illinois at Urbana-Champaign  

Abstract  

Digital collections, such as those available through the Hathi Trust and Internet Archive, offer tremendous opportunities for scholars to access digitized text which can then, with specialized software, be transformed into statistics and visualizations, revealing new characteristics of the content. As a pilot effort for a larger planned project to explore and analyze digitized business information sources, the author created a dataset from, and subsequently a network analysis of, the “Directory of Directors in the City of Chicago” for 1900-1901. This process yielded preliminary insights into the content, organization and subject matter of the book, and the transformation of the text into analyzable formats now makes further research into those areas possible. Through experimentation with techniques and software tools to break down the barriers to examining the book’s contents, this pilot effort also revealed best practices to be implemented and refined over subsequent phases of this project and in other future projects.  

Introduction  

Within the past decade, millions of books, journals, newsletters and other materials have been scanned, digitized, and made freely available online by organizations such as the Hathi Trust (http://www.hathitrust.org/) and Internet Archive (https://archive.org/index.php). This new digital format has eliminated barriers created by distance, since physical materials that are located thousands of miles away from potential users can now be accessed online from any internet connection, and sets and series for which constituent volumes or issues are scattered across multiple institutions can now be reunited online. Further, with these information sources now in digitized form, the contents can be manipulated and analyzed in ways that would have been time-prohibitive just a few years ago, if they were possible at all. Text can be transferred to word processing programs which can, at minimum, count the number of words and the frequency of specific occurrences, and depending on the software, can create elaborate visualizations that are as beautiful as they are informative. Tables of numbers can be transferred into spreadsheets, where the numbers can be subjected to mathematical analysis, and can even be graphed to reveal differences between categories or trends or over time. Photographs and images can be extracted for closer examination of their composition.
With so many materials now accessible online, and so many new options for analyzing these materials, some librarians are taking on new roles to work directly with digitized materials, with software tools, and with the researchers who use them. In a piece titled “We Are All Digital Humanists Now,” Little (2011, 353-354) asks of academic librarians: “will we be able to build skills, workflows, competencies, and infrastructures to help humanities faculty and students create and sustain digital projects?” Two years later, among the emerging roles that Little (2013, 123) saw for academic librarians were knowing how to “utilize new experimental technologies in order to curate content, plan and manage digitization projects and workflows, and bring users and new technologies together.” As these comments demonstrate, supporting humanities research and, more generally, supporting digitization projects, seem to be the contribution areas most frequently mentioned for librarians.

However, humanities subjects are clearly not the only ones present in digital repositories and at this point in time relatively little discussion—either formally in the literature or conference presentations, or informally among practitioners—seems to have been devoted to the emerging research possibilities for social scientists, particularly for business researchers, brought about by the digitization of historical resources. Even a cursory examination of the Hathi Trust, Internet Archive, and other online repositories reveals thousands of historical business books, magazines, newsletters, almanacs, and yearbooks, each begging consideration of the new details, characteristics, and perspectives that may be awaiting discovery. While humanities researchers seem to be well on their way to identifying best practices for analyzing textual content, those techniques are relevant for business information only when it appears as narrative text, for example in investors’ reports, company annual reports, or almanacs or yearbooks. More often than not, the information in historical business sources is in tables, in structured lists, or otherwise formatted to emphasize relationships between names (people or organizations) and important characteristics of those entities. These different formats require different considerations and approaches to effectively extract, manipulate, and analyze the business data.

In addition, relatively little discussion seems to have been devoted to the importance of developing an in-depth understanding of a resource’s content, so as to better advise researchers in its use. While many of these newly digitized resources may have been sitting unused on library shelves for years, digitization has given them new life and now encourages their use, making it easier to incorporate them into research projects. Just as it is becoming increasingly important for librarians to develop technical skills to assist researchers in extracting information from digitized resources, first and foremost it remains important for librarians to understand what resources are available in digitized form, the contents of those resources, and how the resources contrast with and complement other resources in what they offer to researchers.

This paper reports the preliminary findings and outcomes of a project, still in progress, to explore a specific digitized historical business resource, with the goal of gaining an in-depth understanding of its presentation, content, format, and to some extent, its subject matter. At the same time, the project has enabled the author to begin developing the skills to extract data from the resource and derive meaningful information on which to base future research.
Background

In order to better comprehend the nature and scope of the historical business resources which have been digitized, in early 2013 the author began exploring, documenting, and gathering business-related resources in the Hathi Trust Digital Library and Internet Archive. In particular, within the Hathi Trust Digital Library, the author began using the “Collection Builder” to create online “collections” or groupings of digitized resources by topic and type, such as “Directories by State,” “Investment Analysis,” and “Money Trust Investigation.” The Money Trust Investigation collection consists of the digitized transcripts of congressional hearings from 1913 in which “interlocking” directorates—the person-to-person, person-to-organization, and organization-to-organization relationships created by directors with multiple board memberships—were a primary concern, evidenced by the many occurrences of the word “interlocking” within the volumes. As a complement to the Money Trust Investigation collection, the author began creating a collection of resources that contain the names and organizational affiliations of directors during the time of the hearings and in the years before and after.

Director relationships have long been a subject of study, at least since the 1913 hearings, and continuing to the present day. As a demonstration, a simple search conducted on April 29, 2014 in ProQuest’s ABI/INFORM Complete database for interlock* and director* within the abstracts of articles yielded 539 results, including one article from 1914, 238 articles between 2000-2009, and 140 articles from 2010 to present. In addition, at least as far back as 2001, websites such as “They Rule” (http://www.theyrule.net/) have demonstrated that insights can be gained when director relationships between individuals and organizations are visualized and their network structures are revealed.

The author’s initial investigations led to a number of observations, most notably that several historical resources about directorates have been digitized, and that the topic of interlocking directorates clearly has both historical and current interest. Given the potential uses of these digitized resources in future research, it seemed important to gain an understanding of them in order to advise researchers about their content. Further, given the demonstrated value of visualizing directorates to extract insights, a network analysis seemed as though it would be as relevant to gaining understanding of the resources as would an analysis of the text.

As a first effort to conduct text analysis and network analysis using directorate resources, the series “Directory of Directors in the City of Chicago” (the Directory), compiled by the Audit Company of New York, was identified. This six-volume set, produced between 1900 and 1906, is held as physical volumes at a handful of locations, mostly within Illinois and neighboring states. Prior to September 2013, when this project began, the 1900-1901, 1902, and 1906 volumes were the only ones in the series which had been digitized and made available online, specifically through the Hathi Trust Digital Library and Internet Archive.

The contents of each of the Directory’s volumes are arranged in two sections. The first and largest section lists individuals’ names, alphabetized by last name, followed by the names of the organizations which they served as Directors, and any additional roles the individuals held with those organizations. See Figure 1 for an example taken from the first page of the alphabetical listing in the 1900-1901 volume. While the individuals all had Chicago addresses,
the organizations with which they were affiliated may have been located anywhere. The appendix of each book is devoted to a “selected” list of prominent organizations, from Chicago and elsewhere, with a complete listing of the officers and directors of each, regardless of where the individuals themselves were located.

![DIRECTORY OF DIRECTORS IN THE CITY OF CHICAGO.](image)

Figure 1. A sample of entries from the alphabetical listing of the 1900-1901 Directory. (Audit Company of New York 1900, 1)

**Methodology**

As the first and smallest in the series, the 1900-1901 volume of the Directory was the best candidate for a pilot project to quantitatively analyze and visualize the contents. The alphabetical listing, ranging from pages 1 to 222 in the volume, was used in this pilot project. The appendix was excluded from the analysis presented in this paper, because that section contains only a partial set of the individuals and organizations in the alphabetical listing, as well as others not in the alphabetical listing. Further, to manage the scope of the project, using only the alphabetical listing made it possible to focus the analysis on a single geographic location.

The entire 1900-1901 volume of the Directory was downloaded from the Hathi Trust Digital Library website in portable document format (PDF). Because only the alphabetical listing was to be used in this analysis, a new PDF document containing only the alphabetical listing was generated from the old. Next, the author used ABBYY FineReader optical character recognition (OCR) software to read the edited PDF file and output the text into document processing software. Much trial-and-error showed that, for this particular case, outputting the text from ABBYY FineReader into document processing software, and then copying the text to spreadsheet software (specifically, Microsoft Excel) would eliminate most formatting and word-wrapping issues, leaving relatively easy-to-read entries in the spreadsheet.

With the 1900-1901 Directory’s alphabetical entries in Microsoft Excel, the author visually examined each line of the transferred text against the scanned volume which had been downloaded from the Hathi Trust website. In addition to the reasons outlined earlier, the Directory had been selected for this project because the type is clear, the formatting is consistent, and the scanned images are easily readable by both humans and by OCR software. These were desirable characteristics to minimize the number of manual corrections. After minor corrections
were made to ensure that the entries in the spreadsheet were identical to the entries in the scanned version of the Directory, the lines of text were separated and repositioned in Excel to appear as in Figure 2, with each row representing a unique relationship between a person and an organization which he or she served as a director. Columns of information about personal address and other roles with each organization (not shown below) were also added.

![Figure 2. A portion of the 1900-1901 Directory, entered in a spreadsheet.](image)

Since the network visualization was to be based on the person-to-person and person-to-organization relationships stated or implied in the Directory, a crucial first step in analyzing the data was to standardize the way that the names of the individuals and organizations were presented, so that those relationships could be correctly and consistently identified. Because the names of the individuals were presented alphabetically, distinct individuals were readily identifiable by name, or in the eleven cases in which personal names were duplicated (for example, there were two people named “Louis Benjamin”), by also considering their unique personal address. Initially, after the scanned document had been downloaded from the Hathi Trust Digital Library, a visual count performed on the Directory’s alphabetical listing had revealed 5,470 unique individuals. With the entries now in Excel, the count of 5,470 people was confirmed by using Excel’s “pivot table” feature.

Identifying unique organizations presented a greater challenge. Because organization names were scattered throughout the 222 pages of the alphabetical listing, there was some inconsistency from occurrence to occurrence, which was revealed when the data in the spreadsheet was sorted by organization name. To establish some consistency, a column for “Organization Name (Standardized)” was added to the spreadsheet, and can be seen in Figure 2. Much of the needed standardization was achieved simply by matching up variances in punctuation, spacing, and capitalization, or adding “The” to those occurrences without it.

Other standardization efforts required more judgment on the part of the author. The appendix of the 1900-1901 Directory was used to corroborate (when possible) the director relationships in the alphabetical listing, but because the appendix is only a subset of the alphabetical listing, it was only helpful in a handful of cases. Many of the discrepancies arose between railroad companies and similarly-named railway companies, for example, the Chicago & Alton Railroad Company versus the Chicago & Alton Railway Company. In this case, these were treated as distinct organizations, since three individuals were shown as directors for both organizations, implying that these were two separate, simultaneously-existing organizations with separate boards. In all cases, the decision was made to not merge the entries for seemingly similar organizations unless there was compelling evidence to do so, meaning that most of the approximately 20 ambiguous cases encountered were left as separate organizations. Because of this small number, the impact of these decisions on the network analysis is believed to be minimal. Further investigation will be conducted and any necessary corrections will be made.
when the 1902-1906 volumes of the Directory and other supporting materials are reviewed in the upcoming phases of this project. For the 1900-1901 volume, as a result of this standardization exercise, 2,306 distinct organizations were identified.

The software chosen for the network analysis was NodeXL, in large part because it is free add-in software that works within an Excel spreadsheet, where the data already resided. In addition, the author had gained experience with NodeXL through a semester-long network analysis course. To generate a network analysis of people connected to other people through organizations, the data required a particular format to be entered into NodeXL. To begin the formatting, the author sorted the spreadsheet by the “Organization Name (Standardized)” column, to identify all of the Chicago-based directors who appeared in the 1900-1901 Directory for each organization. The author next generated the “projected relationships” from person-to-person afforded by their common directorships. For example, if Adam, Bob, and Charles are the three directors for a single organization, then it is implied, or “projected,” that Adam and Bob have a relationship, interact, or are “connected” through the organization, and Adam and Charles are similarly connected, as are Bob and Charles, for a total of three projected relationships. Generating these relationships is a simple exercise when the number of board members is low, but quickly becomes more complicated as the number of members grows. For the Federal Life Insurance Company and Germania Building & Loan Association, both of which were shown to have 19 Chicago-based board members, the number of projected (person-to-person) relationships generated through each organization was 171. Figure 3 shows what these projected relationships look like in a spreadsheet, for example, the relationship between Jacob Aaron and Lafayette M. Haase as the sole Chicago-based board members of the Aaron Electric Company.

<table>
<thead>
<tr>
<th>Director 1</th>
<th>Shared Organization</th>
<th>Director 2</th>
</tr>
</thead>
</table>
| AARON, JACOB      | Aaron Electric Co.               | HAASE, LAFAYETTE M.
| ABBOTT, C. A.     | Abbott Alkaloidal Co., The       | ABBOTT, WALLACE C.
| ABBOTT, C. A.     | Abbott Alkaloidal Co., The       | RANSON, JAMES W.
| ABBOTT, C. A.     | Abbott Alkaloidal Co., The       | SCOVILLE, LOUIS P.
| ABBOTT, EDWIN F.  | Charles W. Shonk Co.             | SHONK, CHARLES W.
| ABBOTT, W. L.     | Stockholm Manufacturing Co.       | BURLING, WILLIAM S.
| ABBOTT, W. L.     | Stockholm Manufacturing Co.       | NICKELSON, SAMUEL T.
| ABBOTT, W. L.     | Stockholm Manufacturing Co.       | SMITH, W. R.
| ABBOTT, WALLACE C.| Abbott Alkaloidal Co., The        | RANSON, JAMES W.
| ABBOTT, WALLACE C.| Abbott Alkaloidal Co., The        | SCOVILLE, LOUIS P.
| ABBOTT, WALLACE C.| Abbott Alkaloidal Co., The        | WAUGH, W. F.
| ABBOTT, WALLACE C.| Clinic Publishing Co., The        | CURRIB, GEORGE H.

Figure 3. “Projected relationships” for a sample of the directors.

Once the projected relationships were generated, it was easy to identify and remove individuals and organizations that stood alone and would not contribute to a network structure. This determination was based on the number of board memberships that an individual held as well as the size of those boards (or specifically, the number of Chicago-based members) since both aspects are crucial in creating the connections that make a network. There were 159 people who were only on boards for which they were the sole Chicago-based board member. These 159 people were designated as “isolates,” and both they and the 172 organizations they represented were removed from the analysis, since they had no connections with other individuals or organizations in the Directory. As a next step, 382 “orphan” organizations with only one Chicago-based board member were removed, because they would not generate a personal connection with anyone else in Chicago for their lone board member (the individuals themselves remained in the dataset, however, since their memberships on other boards would connect them
with other people). After these people and organizations were removed, the dataset for the network analysis consisted of 5,311 people (5,470 in the alphabetical listing of the 1900-1901 Directory, less 159 isolates) and 1,752 organizations (2,306 in the alphabetical listing of the 1900-1901 Directory, less 172 organizations associated with the isolates and 382 orphans).

With the 5,311 people paired according to their projected relationships through board memberships for 1,752 organizations, NodeXL was used to generate an “undirected” network, so named because in this case there is no direction for the interactions among individuals who simply belonged to the same board of directors at the same time. The analysis method selected in NodeXL was Harel-Koren Fast Multiscale, because it resulted in a visualization that best grouped the individuals into clusters that reflected their board affiliations, as will be seen later.

Findings and Discussion

Before conducting the network analysis, it was helpful to learn more about the data that would be used. The 1,752 organizations are represented in Figure 4 below, categorized by the number of Chicago-based members on their boards of directors. The lowest number of board members represented in Figure 4 is two, because as explained in the Methodology section, the 554 (172 + 382) organizations with only one Chicago-based board member were removed from further consideration in the network analysis. Figure 4 shows that, according to the 1900-1901 Directory, approximately 40% of the 1,752 organizations had three board members in Chicago, making that the most common size. These may have simply been small boards, or perhaps these individuals were actually part of larger boards alongside members from other cities who were not mentioned in this Directory because of its focus on Chicago directors.

![Organizations by Board Size](image)

Figure 4. Organizations in the network analysis, categorized by the number of Chicago board members.

Figure 5 shows the 5,311 individuals with board memberships that connected them to other people in Chicago, categorized by the number of director positions they held. Over 85% of
the people to be included in the network analysis had only one board membership; these individuals were assuredly connected to at least one other person in Chicago through that single director role, because everyone who was not connected to anyone (the 159 people designated as “isolates”) had previously been excluded from further consideration in the network analysis. The overwhelming number of individuals with only one directorship actually provides an important clue to what the network visualization will look like. These individuals by definition did not unite multiple organizations through their single director role, which in turn impacted the extent to which boards in the network interlocked with each other.

Figure 5. Individuals in the network analysis, categorized by the number of directorships held.

The visualization of the network analysis performed on the 1900-1901 Directory appears in Figure 6, which requires some explanation to better understand what is shown. Essentially, NodeXL determined that the 5,311 directors and 1,752 organizations belonged not to a single network, but rather to 1,019 separate, independently functioning networks or “components,” each represented in its own box in Figure 6. The large component on the left side of Figure 6 contains 1,494 people (or 27.3% of all names in the alphabetical listing of the 1900-1901 Directory) represented as dots, connected through their directorships on 521 organizations, which are represented by lines. In the smaller boxes, the lines representing the organizations are more difficult to see, as they are obscured by the dots, circles, boxes, squares, diamonds, triangles, and other shapes that represent the individual directors. The smallest boxes in the lower right corner of Figure 6 represent the 263 components of 2 people. Progressing to the left and upward, the 444 components of 3 people can be seen in the middle, with components of 4 to 33 people shown vertically and horizontally at the top and far left. The earlier finding that over 85% of the network participants had only one board membership is clarified here, too, since those are the individuals who make up the bulk of the 1,019 components, particularly the smallest groups of two, three, and four members. Note that even in those smallest groups, however, there are occasions in which more than one company is represented, implying that that small group of people acted together as directors of multiple companies, but did not interact as directors with
any other person or organization in Chicago. As the later volumes of the Directory of Directors in the City of Chicago are examined, it will be informative to see how the number and sizes of these components change from year-to-year.

Figure 6. Network representation of the Directory, with isolates and orphans removed. The people (points) are connected through organizations (lines).

The most interesting feature of Figure 6 is arguably the single component containing 1,494 people. Within this component, you can find not only some of the best-known organizations and business people in Chicago in 1900-1901, but also hundreds of organizations and people whose names may have been relatively unknown even at that time. This analysis reveals the extent to which small businesses may have participated in the same network with major companies and organizations. Again using NodeXL, statistics about the component and its members were calculated to show that, among these 1,494 people, the maximum “distance” from any one person to another was thirteen “steps,” with an average of five “steps.” In other words, in theory, any one of the 1,494 people could have made the acquaintance of another by sequentially leveraging first their own personal director contacts, and then the director contacts of other directors, meeting up to thirteen people to reach their final “goal” acquaintance, and on average, meeting only five people. The actuality of small business owners interacting with major figures in the Chicago business world in this way cannot be determined from this analysis alone of course, but it does raise an interesting possibility which may encourage further research.

NodeXL was again used to calculate measures of influence and power attributable to the 1,494 people themselves. In particular, three metrics—degree centrality, betweenness centrality, and eigenvector centrality—were calculated. In this application, degree centrality is a function of the number of board memberships that an individual held, combined with the size of those boards. So individuals on several boards and/or on large boards would have high degree
centrality, which presumably equates to more power and influence. Betweenness centrality is a measure of the potential control that one director might have had on the interactions between any two other directors (Wasserman and Faust 1997, 188-189), because he may have been called to act as an intermediary in any communications or interactions, or simply to make introductions. Finally, eigenvector centrality is fundamentally a measure of how closely affiliated any individual was with others who were in positions of power or influence in the network.

In this group, out of all 1,494 individuals, one person ranked first in all three centrality measures: Charles L. Hutchinson. The visualization of his closest connections appears below, in Figure 7. Mr. Hutchinson appears at the center of the network, and the “clusters” of dots and lines encircling the center roughly correspond to the nine boards of which Mr. Hutchinson was a member. The lines emanating from the point in the center reflect Mr. Hutchinson’s degree centrality score of 68. The larger points represented by black boxes are Mr. Hutchinson’s 16 direct contacts who each ranked in the top 10% (out of 1,494 people) in terms of degree, betweenness, and eigenvector centrality. Such a concentration of other very powerful people in direct connection with Mr. Hutchinson corroborates his high eigenvector centrality score.

![Network of Charles L. Hutchinson](Figure 7: Network of Charles L. Hutchinson)

Finally, as a test of the ways in which network analysis could be conducted on a subset of information from the Directory, a list of companies in the rail industry was compiled, based solely on a review of organization names containing either “railroad” or “railway.” In this way, 78 companies were identified, represented by 216 directors in Chicago. In the network visualization shown in Figure 8, NodeXL determined that there were 19 distinct components of Chicago directors connected through their common affiliations with railroad or railway companies. The largest of these components, shown on the left side of Figure 8, contained 54
companies and 146 Chicago directors. Because of the relatively fewer companies in this subset, the details of the component structures, and in particular the intricate configurations of relationships created by rail companies at the time, can be much more clearly seen here than in Figure 6. This visualization also gives an excellent example of betweenness centrality. The large square in the middle of the graph represents Norman Ream, who had the highest betweenness score in this group, corroborated visually by his position as the sole connection between the 35 individuals in the upper left corner and the 110 individuals in the lower left corner.

Figure 8. Chicago directors (points) connected through directorships on railroad or railway companies (lines).

Future Directions

Since this was a pilot project, there were several limitations in its scope. Most notably, while this analysis has yielded valuable preliminary quantitative and visual information about the contents of a single volume, the accuracy, reliability and completeness the contents themselves are completely unknown at this point, and in fact there is evidence to question them. A cursory examination of the alphabetical listing in the 1900-1901 Directory reveals that the entries are almost exclusively men’s names, while the appendix shows that there were women in Chicago serving on the boards of directors for some of the same companies, which at the very least challenges the completeness of the alphabetical listing and the editorial consistency of the 1900-1901 volume. Further, examination of the Directory’s volumes for 1902 through 1906 shows substantial growth in the overall size of the alphabetical listings in each year, with the Directory more than doubling in size between 1900-1901 and 1906. This significant expansion over such a short time creates reservations about the completeness of the first volume. These and similar questions will be addressed in upcoming phases of this project, as the 1902 through 1906
volumes are examined in depth and as multiple other resources are consulted in an attempt to independently corroborate at least some of the information contained in the Directory.

More challenging questions for future investigation also relate to the accuracy, reliability and completeness the Directory’s entries. One glaring unanswered question pertains to how the information for the Directory’s entries was collected. While it seems likely that the information was solicited directly from the individuals themselves, there may be patterns that emerge in later analysis to reveal common characteristics linking the people or the organizations together, making them more likely than others to be candidates for inclusion in the Directory.

Conclusions

As a pilot project for a larger planned research effort, creating a dataset from the 1900-1901 Directory of Directors was an important first step in breaking down barriers that limited the accessibility and usability of the contents for research. Quantitative information provided new insights into the content and organization of the Directory, as well as potential characteristics of the Chicago business environment at that time, which can now be compared and contrasted with quantitative information to be generated from the 1902 through 1906 Directories. Creating a network visualization and analytical metrics from the dataset helped to highlight key individuals and their relationships in 1900-1901, which lays the groundwork for similar analysis for the 1902-1906 volumes. Once completed, these project outputs are intended to provide researchers with an in-depth understanding of the Directory as a series, to aid them in their own research.

Through this project, early best practices for generating data from non-narrative structured text have been identified to implement and refine through analysis of the 1902 through 1906 Directories and similar projects. The clarity of the original text, its structure and format on the page, and the crispness of the scan were found to be important considerations when planning a project that requires OCR software to convert text into data. Selecting analysis methods that will yield meaningful quantitative, qualitative, and visual information, and utilizing the software tools that make such analysis possible, are skills to be developed over time, resulting in new ways to work with digitized resources and support other researchers as they do the same.

References


