CHAPTER 5

Creating an Influencer-Relationship Model to Locate Actors in Environmental Communications

David Rheams

Introduction

This chapter describes a method for locating actors in a corpus of disconnected texts by creating an archive of newspaper articles. The archive can be searched and modeled to find relationships between people who influence the production of public knowledge. My area of focus is an environmental communications project concerning groundwater debates in Texas. Groundwater is a valuable but hidden resource in Texas, often contested and yet little understood. An acute drought in 2011 intensified public interest in groundwater availability, usage, and regulations. News stories about drought, rainfall, and groundwater were a familiar sight in local newspapers, as public officials debated ways to mitigate the drought's effect. Though there was much discussion of groundwater during the drought, the agencies, politicians, and laws that manage groundwater resources remained opaque. I wanted to find out

D. Rheams (✉)
The University of Texas at Dallas, Richardson, TX, USA

l. levenberg et al. (eds.), *Research Methods for the Digital Humanities*,
https://doi.org/10.1007/978-3-319-96713-4_5

how groundwater knowledge was produced and who influenced public knowledge about this essential environmental resource.

I started the project by reading newspaper articles and highlighting the names of relevant actors and places. After reading through a few dozen articles it quickly became evident that the study needed a more sophisticated approach. The relationship between the politicians, corporations, myriad state water agencies, and others was impossible to discern without creating a searchable archive of the relevant newspaper articles. The archive was intended to be a model of groundwater communications that allows a researcher to realize patterns within the texts. This chapter describes the method to create and model this archive. The humanities and social sciences are familiar sites of quantitative textual analysis. Franco Moretti's concept of "distant reading" describes a process of capturing a corpus of texts to find cultural trends through thousands of books.[1] Media Studies scholars use sentiment analysis and other techniques to discover patterns in public pronouncements. The method described in this chapter is similar in that it relies on quantitative analysis. However, the quantification of keywords or the model created from the archive is not the final outcome of the research; it is where one can begin to see the object of inquiry and begin to formulate a hypothesis. Questions are drawn from the model, rather than conclusions.

The method is applicable outside of environmental communications topics. Political, cultural, and social questions may benefit from this approach. I offer a detailed description of this method as a practice of methodically describing my research, but also in hopes that other researchers will continue to refine and improve the processes in this chapter. The chapter discusses each stage of the project in the sections. I conclude with a few thoughts for possible improvement to the method and ways to approach textual analysis critically.

The stages for creating an influencer-relationship model are best summarized by the C.A.G.E. method. The method requires the four following steps:

1. **C**onceptualize the model
2. **A**ssemble the model
3. **G**roup the actors
4. **E**valuate the results

[1] Franco Moretti, *Distant Reading* (London and New York: Verso, 2013).

Table 5.1 Software

Software/Technology	*Description*
Plot.ly	An online data visualization platform
Microsoft Excel	A spreadsheet program that can be substituted with Google sheets
MySQL Workbench	A free database management software suite
Import.io	An online web scraping application
MySQL database	A standard MySQL database hosted online or on a local machine
Text Parsing and Analytics tool	A small application written for this research project to parse and analyze text documents

The model is designed during the conceptualization phase as a manual prototype of the more extensive project. The next step, assembling the model, is the process of collecting and storing data in a searchable manner. The researcher can find patterns among different groups by coding the initial research results. For example, this project separated *politicians* from *aquifers* to assist in searching the database. The final stage, evaluating the results, renders the data into a usable format. Each of these steps required software to aid in data collection and management. While this list of applications in Table 5.1 is not meant as a recommendation or endorsement, they serve to illustrate the type of software available to a researcher. Software changes quickly, and there are always new methods and platforms to explore.

The concluding section of this chapter discusses alternatives to these applications. The technology used to perform the method is somewhat interchangeable. The platform required to assist with research is at the researcher's discretion, and should be chosen based on the requirements of the research design. The conceptual model determines the requirements for the research project.

The Conceptual Model

The conceptual model is the planning stage of creating an influencer-relationship model. During this phase, the constraints of the project are identified, along with potential sources of texts, and a method to analyze the texts. Essentially, the conceptual model is the prototype for the larger project. This phase also allows the researcher to validate the method by manually collecting a small amount of data and seeing if the study design

accomplishes the desired results. I approach this stage by asking a series of questions:

1. **What texts are likely to contain influencers?** There is a range of texts where actors exert influence over public opinion: newspapers, blogs, government documents, speeches, or online videos. The medium of communication will, in part, determine the next two questions.
2. **How should the texts be collected?** When building the archive, texts need to be collected and stored in specific ways to ensure the researcher can ask questions of the data. This problem should help the researcher find the best method of collection either by using online repositories of information, manually collecting files stored online, or other locations. The key to this question rests on finding a way to save computer-readable text. A MySQL database is easy to use, but has limitations with large datasets. There are many types of databases, and I encourage the researcher to find the one that fits the research design criteria. There is no reason to invest thousands of dollars on a platform if the source material is only a few thousand words.
3. **How should the data be evaluated?** This question determines the way a researcher will interact with the data. Different data types and databases allow for different kinds of queries and visualizations, so it is essential to understand the strengths and limitations of the software platforms before creating the archive. The conceptual model is flexible enough to allow a researcher to make mistakes and start over before investing too much time (or money) into a particular technical approach.

My conceptual model consisted of arranging newspaper articles on a whiteboard. To begin this project, I read ten articles about groundwater and highlighted the names of all the actors quoted in the article. Journalists use a person's full name the first time they quote them in a story, so looking for two or more consecutive words with capital letters provided a list of the actors quoted in stories about drought and groundwater. The approach also captures geographic places and the names of state agencies, as most of these are multiple words. Next, I put these articles on a whiteboard and drew lines to the different news stories that quoted the same actors. These lines highlighted the relationship between

actors. After I validated that the model was capable of producing the desired result, I focused on specific texts and a method of analysis. The process of viewing articles on a whiteboard helped to clarify the next steps of the research, which was to build an archive of newspaper articles.[2]

Why Newspapers?

Newspapers are the basis for this study's communications model, because they play a significant role in shaping public opinion, disseminating information, and providing "knowledge claims" to the public.[3] Even with social media, citizens still turn to local newspapers for information about regional politics and other topics specific to the regional community.[4] Newspapers rank just above televised news programs regarding viewership within a local community in 2011.[5] However, the articles are not limited to print; the readership study includes digital versions of the news stories. Though the media landscape has changed from 2011 with digital distribution gaining prominence both in readership and newspaper revenue, both print and digital newspaper articles are a critical source of local information according to a 2017 Pew Research Center "Newspaper Fact Sheet."

Choosing the sources of the articles was one of the first steps in designing this study. Ideally, every item from every newspaper in Texas would be freely accessible, but this is not the case. I limited the archive by making a list of all newspapers in operation in Texas between 2010 and 2014, during the acutest drought years in the past few decades. I removed any paper catering to suburbs or small towns as they would likely be reproducing stories from larger papers and would be more

[2] I repeated the process of creating a conceptual model many times to arrive at a workable method and found it helpful to document each question and approach. First, research practices should be transparent, and it can be easy to forget to record critical choices. Second, the documentation can help to clarify the results later in the research process.

[3] Mats Ekström, "Epistemologies of TV Journalism," *Journalism: Theory, Practice & Criticism* 3, no. 3 (2002), 259–282.

[4] Tom Rosenstiel, Amy Mitchell, Kristen Purcell, and Lee Rainier, "How People Learn About Their Local Community," Pew Research Center, 2011; Maxwell T. Boykoff and Jules M. Boykoff, "Balance as Bias: Global Warming and the US Prestige Press," *Global Environmental Change* 14, no. 2 (2004), 125–136.

[5] Tom Rosenstiel, Amy Mitchell, Kristen Purcell, and Lee Rainier, "How People Learn About Their Local Community," Pew Research Center, 2011.

difficult to access. Therefore, my study limits the papers to only those printed in cities with a population over 50,000. I further reduced my list of sources to include only newspapers that allowed full-text access to articles either on their website or from Westlaw, LexisNexis, or ProQuest. I settled on nine newspapers and two monthly magazines that met these criteria. The research sample represented metro areas (e.g., the Dallas-Fort Worth metro area) and smaller cities (e.g., Midland and El Paso), coastal areas, and towns in both east and west Texas. The two nationally recognized state magazines, the *Texas Monthly* and the *Texas Observer*, concentrate on issues specific to Texas. I selected these sources to capture a cross section of the state: rural and urban articles, as well as articles written in the different ecosystems and economies across Texas.

Qualitative Content Analysis

The mapping process is similar to creating a citation map. In a typical citation map, the researcher knows the actors they are searching for in advance. However, this technique is designed to uncover previously *unknown* actors and networks. Content analysis was a natural choice for a research method because it lends itself to projects that require examining large volumes of text and allowed a view of the conversations about hydraulic fracturing, agricultural groundwater, and domestic water conflicts. Additionally, this method is both predictable and repeatable.[6]

The output of this analysis provides a list of actors and articles to investigate further, and the results are quantifiable based on the number of articles that contain the actor. While there are well-documented issues with word frequency counts, this study mitigates the risk of connecting frequency to importance by combining quantitative identification methods with qualitative analysis.[7] Statistics is not the basis of observations found in this research project; instead, the quantitative results become a guide for further investigation. Krippendorff describes this method as

[6] Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed (Thousand Oaks, Calif: Sage, 2004); J. Macnamara, "Media Content Analysis: Its Uses, Benefits and Best Practice Methodology," *Asia Pacific Public Relations Journal* 6, no. 1 (2005), 1–34; Bernard Berelson and Paul Lazarsfeld, *Content Analysis in Communications Research* (New York: Free Press, 1946).

[7] Steve Stemler, "An Overview of Content Analysis," *Practical Assessment, Research & Evaluation* 7, no. 17 (2001).

qualitative content analysis, where "samples may not be drawn according to statistical guidelines, but the quotes and examples that the qualitative researcher present to their readers has the same function as the use of samples".[8] Content analysis separates the actors from the articles to render the texts abstract, but searchable.

The people who produced knowledge claims became clear once the articles became abstract. The process of grouping actors together revealed observations and insights into when, where, and to whom the public looks for information about groundwater. For example, this method identified each article where actors overlap, allowing a researcher to generate a list of *politician*s quoted in articles about the *Edwards Aquifer*. Another query located state senators' names in articles that contained the key-phrases *ExxonMobile*, *DuPont*, *TWDB*, or *Texas Railroad Commission*. Each of these queries was combined with the metadata.[9]

Assembling the Model

I collected 4474 articles from nine Texas news publications written between 2010 and 2014 using online newspaper repositories and collecting articles directly from newspapers' websites. Each method accompanied different technological challenges. The first data collection method was to search LexisNexis, ProQuest, and Westlaw for newspapers with the option to download full articles. Four newspapers met these criteria: *The Austin American-Statesman*, *The Dallas Morning News*, *The Texas Observer*, and the *El Paso Times*. The search queried the full text of articles published between 2010 and 2014 and returned results for all articles containing the words *drought* or *groundwater*. The search terms were kept deliberately broad, allowing the software to capture possibly irrelevant information. Irrelevant articles were filtered before downloading by using negative keywords to remove sports-related articles.[10]

[8] Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed (Thousand Oaks, Calif: Sage, 2004).

[9] The metadata for a text document contains the articles publication name, city, date, author, word count, and other identifying information.

[10] A negative keyword is any word that should not return results; it is used to narrow a search. For example, sports terms were made negative. One of the unintended and unofficial findings of this project is that "drought" is more common when describing a basketball team's win/loss record rather than a meteorological condition in local papers.

The repositories exported the news articles into a single text document containing 500 articles. While this format is acceptable for a human reader, the files must be converted from text into a table for purposes of digital content analysis. Each article needs to be separated into columns that allow for analysis; one column contains the author, another includes the publication date, another consists of the source, another contains the text of the article, and so on, to allow for content analysis. Otherwise, it is impossible to separate one result from another or to group articles that share common attributes together.

To overcome this challenge, I needed a data-parsing web application to convert the text file into rows within a table. The program had three requirements: first, the tool must upload the text file to a database; second, the application must run a *regular expression* to separate the text into rows based on pre-designated columns; and third, the application needs to write the results in a standard format (CSV).[11]

I was unable to find any software that performed this task, so I asked a colleague for help designing a small application to help separate the text files into a usable archive. I worked with that colleague, Austin Meyers, to design a lightweight PHP application, the Text Parsing and Analysis Tool, to build and analyze the archive.[12] The program converts output from LexisNexis, ProQuest, and Westlaw, from large text files into a MySQL table. The table columns contain the full text of a single article, links to images in the story, and other useful metadata. The second function of the Text Parsing and Analysis tool was to analyze the stories within the database. The next section discusses this feature in detail.

[11]A regular expression is a sequence of characters used to find patterns within strings of text. They are commonly used in online forms to ensure the fields are correctly filled out. For example, a programmer may use a regular expression to confirm a cell has the correct format for a phone number or address. If the user does not use the proper syntax, an error is returned. There are numerous online tutorials to help people write a regular expression. I used regex101.com and regexr.com to help write the expressions needed for this project.

[12]Austin Meyers, founder of AK5A.com, wrote the PHP scripts used in the application. Austin and I have collaborated over the past 15 years on numerous technical projects and applications. The process of web scraping is a method for extracting objects or text from HTML websites. There are many software companies producing applications to assist in mining website data. Researchers may also choose to build a web scraper for specialized research projects.

Table 5.2 Articles table

article_id	author	headline	publication	date	city	length	article_text	url

If newspapers were not available from the databases, I accessed them directly from the newspaper's website. However, I did not want to copy and paste each article into a database because manual processes are time-consuming and error-prone. Instead, I used a web scraping application[13] to collect the articles from newspaper websites. After experimenting with several scrapers, I decided on Import.io to automate text and image scraping from a group of websites.[14] Import.io had a simple user interface and did not require writing any complicated code. Import.io outputs the data in a comma separated value file (CSV) that is readable by any spreadsheet program. The table created by Import.io had the same column names as the tables created by Text Parsing and Analysis Tool, which avoided confusion and reduced the amount of time needed to match the sources.

Group the Actors

Table 5.2 shows the organization of articles into a single table with columns for the publisher, the full text of the article, the publication date, and other identifying information. The table allows a user to search through all nine sources for specific phrases or patterns in the text and find relationships between the publications.

Text Parsing and Analysis Tool

Once the data existed in the table, another problem presented itself. Even though I had identified the textual patterns within the articles, there was not an efficient way to sort through each article using MySQL

[13]Web-scraping is a technique to transfer the content of a webpage to another format.

[14]Import.io is a free web scraping application that converts a webpage into a table. It can be automated to run against multiple websites or used to search within large sites. For example, the Brownsville Herald website has over 118,000 pages (as of December 19, 2017) and it would be impractical to search the entire site and copy and paste individual articles. The website accompanying this book has a video on how Import.io can be used to gather newspaper articles into a database.

```
1 ([A-Z][a-zA-Z0-9-])([\s][A-Z][a-zA-Z0-9-])+
```

Fig. 5.1 Regular expression for consecutive capitalized words

queries. The classification of two million words in the archive needed completion on both an individual level and within the context of other words.

The Text Parsing and Analysis Tool is designed to locate patterns within the syntax of the text, what Krippendorff calls syntactic distinctions, to identify unknown actors.[15] The syntactical distinction searched for in this scenario was any string of text where two or more consecutive, capitalized words were found. In effect, this method provided a list of proper nouns found in the dataset. For example, a search using this criterion will identify *Allan Ritter* (a Texas State Representative) or *Texas Railroad Commission* because there are two or more consecutive capitalized words in each phrase. The search could identify any phrase that contained two or more consecutive capitalized words, using the regular expression given in Fig. 5.1.

The process is similar to a Boolean phrase match[16] used in search engines; but rather than locating a distinct expression, the search determines proper nouns. A series of actions happen once the tool identifies a proper noun: the application records the phrase in a table, it applies a unique ID number to the words automatically, it records the article ID number, and the 60 characters preceding the text. An identification number was attached to each of the key phrases to assist in writing queries to find patterns between the articles.

The regular expression searched the articles and returned 244,000 instances of consecutive capitalized words (i.e., key-phrases). The majority of these phrases were not relevant or only appeared once or twice within the text. Any key-phrase found in less than 20 articles was removed from the list to reduce it to a manageable and meaningful size. While the frequency of a particular key phrase does not necessarily connote importance in groundwater conflict, it helps a researcher prioritize the list. I completed this process by sorting a frequency list in the

[15] Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed (Thousand Oaks, Calif: Sage, 2004).

[16] A Boolean phrase match allows a user to search for phrases using operators such as AND, OR, NOT to refine searches.

Table 5.3 Key phrase table structure

article_id	keyword	keyword_id	preceding_60_characters	following_60_characters

database and manually deleting the irrelevant results. The new table had the column defined in Table 5.3.

The Article ID column connects each key-phrase to the relevant article. The search located 362 relevant key-phrases. The table has considerably more rows than the Articles Table, because there are multiple key-phrases per article. The key-phrase table makes the data searchable, but the unwieldy size of the table makes it difficult to parse. The key phrases need to be grouped to help a researcher see the patterns of relationships between actors. The process of adding the group is another way to *code* the data.

Coding the Key Phrases

The code for the key-phrases is descriptive words that follow five broad groups: people, places, agencies, industries, and activist groups. For example, the phrase *Allan Ritter*[17] appeared in 50 articles and was coded as a *politician*, thus grouping *Allan Ritter* with the other state politicians found throughout the articles. The Edwards Aquifer was coded as *aquifer,* and Ladybird Lake was coded as *lake*. The 27 codes were mutually exclusive and included:

- State Government Body
- Lake
- City Government Body
- River
- National Government Body
- Politician
- Aquifer Name
- Activist Group

[17]Allen Ritter is the Texas State Representative for District 21 and current chairman of the Texas House Committee on Natural Resources.

Though the codes for this project were relatively simple, this stage of research is critical. Different coding criteria changes the output and the way the researcher interacts with the data. There are risks in reliability problems that occur from "the ambiguity of word meanings, category definitions," and these risks are compounded when more than one person applies the codes.[18] The way to reduce reliability problems is to provide definitions of what each code means and ensure that the codes used are mutually exclusive. For this project, I wrote definitions of each code before assigning the codes. There are a number of textbooks on how to correctly code data for qualitative research methods, though I found Krippendorf's work to fit this project.[19]

I completed the task of coding by manually adding one of the 27 codes to each of the 362 key phrases (i.e., actors) on an Excel spreadsheet. Once the spreadsheet was complete, I uploaded the CSV as a new table in the MySQL database. This approach worked because I had relatively few key phrases and codes. Online security and performance were not issues because the database was not public[20] and relatively small. The upload added a column to the key phrases table, changing the structure to the following (Table 5.4).

Once coded this way, the actors are cross-indexed by time, location, mutual key phrases, or other variables. Every keyword phrase or keyword code was assigned a numerical identification number to assist with creating an accurate index and keep the MySQL queries short. The identification numbers also helped to verify the correctness of the query by not relying on text searches of the database. However, the techniques of creating the archive are less important than the questions a researcher

[18] Robert Philip Weber, *Basic Content Analysis*, 2nd ed. Sage University Papers Series, no. 07-049 (Newbury Park, CA: Sage, 1990).

[19] Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed (Thousand Oaks, Calif: Sage, 2004); Johnny Saldaña, *The Coding Manual for Qualitative Researchers*, 3rd ed (Los Angeles, CA and London, New Delhi, Singapore, Washington DC: Sage, 2016); Sharan B. Merriam and Elizabeth J. Tisdell, *Qualitative Research: A Guide to Design and Implementation*, 4th ed (San Francisco, CA: Jossey-Bass, 2016).

[20] A local server is a MySQL database hosted on the user's computer rather than hosted by a provider. Running the database on a local machine, as opposed to online, reduces risks allowing the user to experiment without worrying about security or performance issues. Instructions, best practices, and links to help get you started with a MySQL database on are on the website accompanying this book.

Table 5.4 Key phrase table structure with codes

article_id	keyword	keyword_id	preceding_60_characters	following_60_characters	keyword_code	keyword_code_id

asks of the archive. The technology enabled the ability to search, but did not dictate the search.

Once the phrases were identified and the tables existed in the database, I was able to query the database as needed to answer the question at hand. The answer to one question usually led to additional queries. For example, I wrote queries to elicit which politicians were most likely to be quoted in the same article that discussed aquifers. Another query created a matrix of each actor listed alongside all other key phrases and the article ID where the pair was located.

Use Case: Querying the Database

Each text analysis project will require some method of querying the database. One of the requirements for the influencer-relationship model is to locate places where two or more actors are quoted in the same article. The following query is an example of the method I used to create a table of news stories with both the phrase *Railroad Commission* and *Environmental Protection Agency* (shown as *keyword_id* 263 in the query). The following snippet of MySQL has been simplified (pseudo code) to show the method rather than the actual query.[21]

The query in Fig. 5.2 tells the database to combine two tables (the Articles table and the keywords table) and find all the articles with both keywords. There are also two comments in the query to remind the user where to input variables. I often include similar comments when using queries multiple times. These comments also help clarify the way the query functions. The query outputs a table with each of the objects under the SELECT statement (lines two through eight) as columns.

The query produces the table by using a *match against* expression to *match* the text in quotation marks *against* the article column and returns

[21]There are many ways to construct a MySQL query. I used a query similar to the one in Fig. 5.2 because it fit into my workflow; it was easy for me to find the keyword_id that associated with the keywords I was interested in. However, another researcher may have rewritten the query differently, but still arrive the same output.

```
SELECT
      article_id,
      keyword_id,
      keyword_code_id,
      keyword_code,
      keyword,
      author
      publication
FROM
            articles
      JOIN
            keywords ON articles.article_id = keywords.article_id

-- ---------------------------------INSERT KEYWORD HERE -------------------
where
      match(article)
      against(' "KEYWORD" ' IN BOOLEAN MODE)
-- ---------------------------------INSERT KEYWORD ID HERE -------------------

and keyword_id = 12345
group by article_id
```

Fig. 5.2 MySQL query example

Table 5.5 Results of the query

article_id	*keyword_id*	*keyword_code_id*	*keyword_code*	*keyword*	*publication*	*author*
1	287	22	Environmental Activist Group	Railroad Commission	*Austin-American Statesman*	Price
82	369	45	State Government Body	Environmental Protection Agency	*Brownsville Herald*	Smith

all of the articles with the phrase *Railroad Commission*. All rows are filtered using the *keyword_id* field within the search. I kept a table of each keyword ID as a reference when writing queries (Table 5.5).[22]

The result is displayed with both index numbers and their English language referents. Both are presented to assist with verification of the result. Querying a database is often a process of trial-and-error. I wrote many queries which did not work, or returned inaccurate results, while working through the data. Each new query was a refinement of a previous one and often lead to more questions. The output can be viewed in multiple ways. Either as tables or as charts, graphs, or other data visualizations.

[22]There are other ways to accomplish the same goal using MySQL, the only requirement for this type of project are that the results are accurate. Using identification numbers rather than the text searchers sped up the verification process.

Evaluate the Results

Evaluating the results from the queries gives the data shape, and helps to render it into a usable format. I used data visualizations as a way to show the relationship between actors within newspaper articles by allowing a broad view of the database. All of the materials under review can be viewed at the same time, which is helpful when trying to determine the scope and impact of particular actors. There is a rich history of data visualization literature[23] and there are numerous methods of rendering data into a coherent image. One of the benefits of the Digital Humanities is that these practices are conducted in conjunction with considerations about visualizing complex datasets. Most critiques recognize that a visualization is not an exact representation of the object being studied, but a particular perspective of a specific database.[24] The visualization is not necessarily reality, but a guide to additional questions.

Johanna Drucker points to some of the limitations of humanities visualizations, warning that their uncritical use may serve as an intellectual "Trojan horse".[25] She argues that research in the humanities has long resisted the temptation to reduce all phenomena to data and that humanists need to recognize uncertainty and complexity. Lisa Otty and Tara Thomson explain how digital humanists alleviate some of the potential errors.[26] I stand in agreement with Otty and Thomson, that we must communicate the exact steps undertaken to make the design choices. The steps required to create the visualizations are "crucial components" of the visualization and the research project as a whole.[27] They cite Dörk

[23] Edward Tufte has written extensively on data visualization beginning in the 1970s (Tufte 2001, 2006). Ben Fry has not only written on the subject (Fry 2008) but has also developed a programing language around data visualization called Processing.

[24] Marian Dörk, Christopher Collins, Patrick Feng, and Sheelagh Carpendale, "Critical InfoVis: Exploring the Politics of Visualization," In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, edited by Wendy E. Mackay and Association for Computing Machinery (New York: ACM, 2013).

[25] Johanna Drucker, "Humanities Approaches to Graphical Display," *Digital Humanities Quarterly* 5, no. 1 (2011).

[26] Lisa Otty and Tara Thomson, "Data Visualization in the Humanities," In *Research Methods for Creating and Curating Data in the Digital Humanities*, edited by Matt Hayler and Gabriele Griffin. Research Methods for the Arts and Humanities (Edinburgh: Edinburgh University Press, 2016).

[27] Ibid.

et al.'s four principles of visualization defined in their article "Critical Info-Viz: Exploring the Politics of Visualization' as a way to visualize data in the Digital Humanities.[28] Dörk et al., Otty and Thomson, and Drucker each state the requirement to think critically about the potential ways visualizations shape knowledge. Visualizations should prevent preconceived ideas rather than entrench them. Models are representations of the data and should allow a researcher to explore the object of inquiry fully, without guiding the viewer to a specific conclusion.

Researchers can ensure models are open to interpretation by thinking through their methodology. Digital Humanities, like other disciplines, tend to reflect on their research to determine how the researcher impacts their results. For example, had I limited my model to show only the relationship between corporations and aquifers, my analysis would have been skewed towards corporations. The visualizations would not be open to exploration because that sort of approach limits the study to a single conclusion. To prevent this scenario, Dörk et al. recommend four principles to follow when creating visualizations: *disclosure*, *plurality*, *contingency*, and *empowerment*. I have found that these serve as a useful guide for most any project. A researcher must *disclose* the exact steps taken to create visualization to allow a viewer to understand the presentation of data. *Plurality* asks that a visualization be holistic, answering multiple different questions so as not to present only one side of the data. The contingency principle "acknowledges the situation of the viewer in relation to the phenomenon being represented" by allowing a viewer to explore the visualization, on their terms. For a map-based visualization this can mean orienting the map towards their location. The final principal, *empowerment*, acknowledges the way a viewer can interact with the visualization by leaving comments or making updates available for all viewers. This research project followed these principles as closely as possible, and in doing so, proved an early hypothesis regarding the influence of corporations on groundwater knowledge claims wrong.

During the research project, I created numerous tables to help sort the data in different ways. The visualization helped to provide a different perspective of the entire dataset. The first visualization I produced is

[28] Marian Dörk, Christopher Collins, Patrick Feng, and Sheelagh Carpendale, "Critical InfoVis: Exploring the Politics of Visualization," In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, edited by Wendy E. Mackay and Association for Computing Machinery (New York: ACM, 2013).

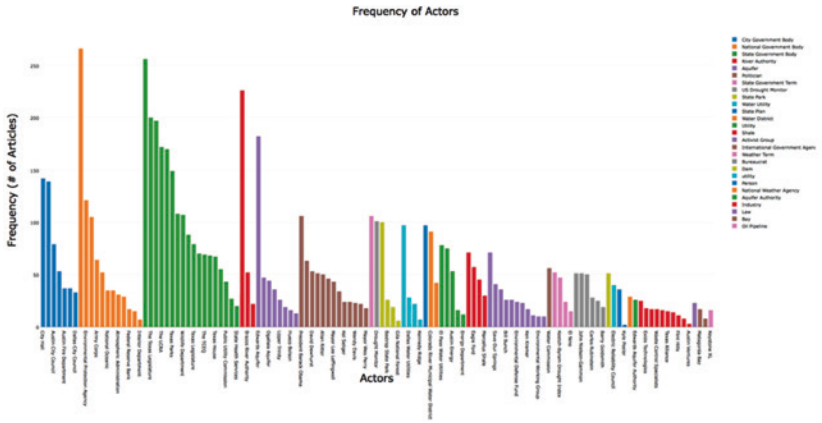

Fig. 5.3 Frequency of actors

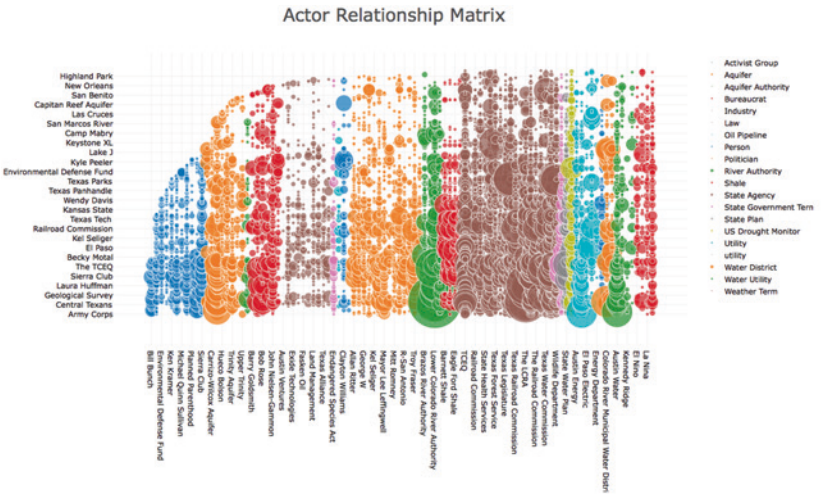

Fig. 5.4 Actor relationship matrix

a bar chart of the frequency of each actor as they appear in the dataset. The bars are grouped by color to denote the category of the actor; these categories were based on the codes I applied to the keyword phrases. I used data visualization software, Plot.ly, to create Figs. 5.3 and 5.4.

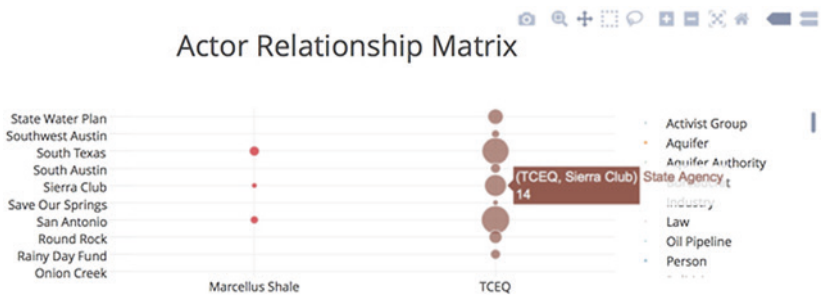

Fig. 5.5 Magnified view of the actor relationship matrix

Figure 5.3 does not prove anything on its own; however, in context, it is a useful way to visualize the collection of articles and allows the viewer to explore the data on their own terms. The blue, orange, and green bars represent the city, federal, and state agencies. A quick look at the chart shows the frequency of quotes and stories from official sources, which shows where journalists are getting much of their information. To validate this finding, I sent emails to Texas journalists writing about environmental and water issues.

To address the relationship between actors, I created the Actor Relationship Matrix chart shown in Fig. 5.4. Thinking back to Dörk et al.'s, principles, this chart helps to answer questions about *plurality* and *contingency*. The size of the sphere represents the frequency of stories that contain both actors. The colors represent the groups of key phrases to make the chart easier to parse. When viewed in its entirety, as it is in Fig. 5.4, the chart is similar to the visualization in Fig. 5.3. However, once the chart is magnified, the viewer can explore the data via the relationship between actors within the articles. Figure 5.5 is a magnified view of the Actor Relationship Matrix that allows a user to examine the database and the relationship between actors.

The magnified view shows that 14 articles listed both the Texas Commissions on Environmental Quality and the Sierra Club. The journalistic practice of creating balanced stories by quoting opposing viewpoints within the same story is evident when looking at this chart.[29]

[29] Maxwell T. Boykoff and Jules M. Boykoff, "Balance as Bias: Global Warming and the US Prestige press," *Global Environmental Change* 14, no. 2 (2004), 125–136.

The practice of indexing that Lance Bennett described is also evident in this dataset, along with other takeaways.[30] The data in this matrix is designed for a user to explore and help to see the articles in a productive way. While I used environmental topics for this paper, any relevant topic or dataset could be presented in a similar matrix.

I used the free version of Plot.ly to create these visualizations.[31] While Plot.ly allows a user to connect the visualization program directly with a MySQL database, it was easier for me to export the results of a query as a CSV and upload the CSV to Plot.ly. I chose Plot.ly as the visualization platform because it was accurate, has an accessible user interface, and is free for basic users. There was no need here to invest in expensive software options. Especially for scholars who are new to this type of software, it is often easiest to start simple and explore. I created quite a few visualizations before I found ones that were useful. Again, the process is trial and error, with each iteration being a refinement of the one before.

Outcomes and Next Steps

Another way to think of the output of this project is to think of it as a small-scale search engine, for a particular group of articles that would have been unavailable in any other context. For this project, the quantitative analysis leads to more rigorous qualitative research. I used the analysis from the visualizations to locate actors to interview, articles to examine, agency websites to explore and environmental laws to review. In other words, the broad view of the data enabled and clarified further research. I used the model as a reference while conducting interviews with state hydrologists, water district managers, and environmentalists. For example, the influencer-relationship model showed that journalists relied heavily on state agencies, especially the Texas Water Development Board, for information about groundwater, but that they did not cite spokespeople from industries which use groundwater. This result led to an interview with a prominent Texas environmental journalist and became the genesis of a conversation about how journalists located sources for groundwater stories.

[30] See Lance Bennett's (2016) *The Politics of Illusion* (10th edition) by the University of Chicago Press.

[31] The specific steps required to create these charts are available on the website accompanying this book.

All the choices made during this project had an impact on its results. It is helpful to be aware of each of one's design decisions, and the impact that these decisions have on the outcome when thinking critically about Digital Humanities projects. If I had chosen to use a different database schema, or a slightly different regular expression, or if I had visualized the output differently, it may have altered the results. The technology used to create the influencer-relationship model is not static, however, and should be expanded to include new platforms.

If I were to rebuild the Text Parsing and Analysis Tool, I would consider using Python, rather than PHP with regular expressions. Python's Natural Language Toolkit is designed to assist with language analysis.[32] However, I had to choose whether to take time off to learn a new language or to work with the tools readily available. I chose the latter because it was enough to arrive at the model I needed, and the improvements were not evident until we had completed the first version of the application. Alternatively, if the project required modeling millions of texts, one could use a NoSQL database such as MongoDB, and/or employ Dgraph to query the archive. However, the technologies used are less important than the methodology. This process can be used with any technique or technological tool. The key aspect of the method presented here is the ability to locate influential actors across multiple sources of texts. The specific patterns and techniques are decided by the research question. I considered this project a success because I was able to ask questions of the data, and it helped shape the next phase of research, which included interviews with actors, close reading of the articles, and examinations of new documents that were alluded to within the news stories. Like software, research is an iterative process that gets more accurate with each new phase. Each new look at the object offers new opportunities to refine methods in the Digital Humanities.

References

Berelson, Bernard, and Paul Lazarsfeld. *Content Analysis in Communications Research.* New York: Free Press, 1946.

[32] Python is a general-purpose high-level programming language and is ideal for small-scale applications. Python's Natural Language Toolkit (NLTK) is a platform that allows Python application to work with English language data. The NLTK is a free, open-source project used by researchers across disciplines.

Boykoff, Maxwell T., and Jules M. Boykoff. "Balance as Bias: Global Warming and the US Prestige Press." *Global Environmental Change* 14, no. 2 (2004): 125–136. https://doi.org/10.1016/j.gloenvcha.2003.10.001.

Dörk, Marian, Christopher Collins, Patrick Feng, and Sheelagh Carpendale. "Critical InfoVis: Exploring the Politics of Visualization." In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, edited by Wendy E. Mackay and Association for Computing Machinery. New York: ACM, 2013. http://mariandoerk.de/criticalinfovis/altchi2013.pdf.

Drucker, Johanna. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5, no. 1 (2011). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html#p4.

Ekström, Mats. Epistemologies of TV Journalism. *Journalism: Theory, Practice & Criticism* 3, no. 3 (2002): 259–282.

Fry, Ben. *Visualizing Data*. Sebastopol, CA: O'Reilly Media, Inc., 2008.

Krippendorff, Klaus. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, Calif: Sage, 2004.

Macnamara, J. "Media Content Analysis: Its Uses, Benefits and Best Practice Methodology." *Asia Pacific Public Relations Journal* 6, no. 1 (2005): 1–34.

Merriam, Sharan B., and Elizabeth J. Tisdell. *Qualitative Research: A Guide to Design and Implementation*. 4th ed. San Francisco, CA: Jossey-Bass, 2016.

Moretti, Franco. *Distant Reading*. London, New York: Verso, 2013.

Otty, Lisa, and Tara Thomson. "Data Visualization in the Humanities." In *Research Methods for Creating and Curating Data in the Digital Humanities*, edited by Matt Hayler and Gabriele Griffin. Research Methods for the Arts and Humanities. Edinburgh: Edinburgh University Press, 2016.

Rosenstiel, Tom, Amy Mitchell, Kristen Purcell, and Lee Rainier. "How People Learn About Their Local Community." Pew Research Center, 2011. http://www.journalism.org/2011/09/26/local-news/.

Saldaña, Johnny. *The Coding Manual for Qualitative Researchers*. 3rd ed. Los Angeles, CA and London, New Delhi, Singapore, Washington DC: Sage, 2016.

Stemler, Steve. "An Overview of Content Analysis." *Practical Assessment, Research & Evaluation* 7, no. 17 (2001). http://PAREonline.net/getvn.asp?v=7&n=17.

Tufte, Edward R. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press, 2001.

———. *Beautiful Evidence*. Cheshire, CT: Graphics Press, 2006.

———. *Envisioning Information*. Fourteenth printing. Cheshire, CT: Graphics Press, 2013.

Weber, Robert Philip. *Basic Content Analysis*. 2nd ed. Sage University Papers Series, no. 07-049. Newbury Park, CA: Sage, 1990.