Metadata Management: Visualization of Data Relationships

Jessica J. Schuler
Department of Resource Analysis, Saint Mary's University of Minnesota, Minneapolis, MN 55404

Keywords: Geographic Information System (GIS), Metadata, Spatial Data Relationships, Visualization

Abstract

As spatial data resources increase, metadata management becomes increasingly important. Many data resources hold dependent relationships with other data resources in a large infrastructure. Modern visualization techniques may be utilized to graphically display relationships between spatial data resources. This project explores the usability of three different graphical visualization methods for web-based applications. A survey containing the three visualization methods was administered to volunteers and measured the usability of each method based on volunteer response.

Introduction

As spatial data becomes more open, the map resources collected each year are increasing in number (Gui, Cao, Liu, Cheng, and Wu, 2016). The Minnesota Department of Natural Resources (DNR) is a large state agency with over 2,000 users who create spatial data. DNR data resource administrators manage hundreds of spatial data resources created by users within the DNR. Many resources rely upon one another, and changes to these resources may break data connections between resources. Geospatial data is characterized by complex multifile formats that need to be tied together, bundling data with associated metadata and ancillary documentation (Morris, 2006).

Management of DNR spatial data resources is cumbersome when utilizing manual methods, and relationships between resources may be missed. Interactivity and intuitive information visualization methods are critical factors of any framework that needs widespread and voluntary adoption (Gui, Yang, Xia, Liu, Xu, Li, and Lostritto, 2013).

Visualization of spatial data resources, and the relationships between those resources, would help to identify resources containing links to malfunctioning data. Data which has malfunctioned has a cascading effect upon all other resources linked to it. With relationships between resources easily visualized, changes to data resources may be avoided if it will be the cause of data to malfunction. Effective visualizations clarify data; they transform abstract collections of numbers into shapes and forms that viewers quickly grasp and understand (Thomas, 2015). The ability to quickly access and correctly interpret data is important to DNR data resource administrators.

Background

Staff at the DNR created an open source application called "Geospatial Information

Jessica J. Schuler. 2017. Metadata Management: Visualization of Data Relationships. Volume 20, Papers in Resource Analysis. 10 pp. Saint Mary's University of Minnesota University Central Services Press. Winona, MN. Retrieved (date) http://www.gis.smumn.edu

Tracker (GI Tracker)" to assist in managing organizational geospatial data. Data is collected via a Python script run daily which iterates through all agency server instances specified, collects metadata about all data resources, including relationships, and stores this information in a Mongo database instance.

Data is interacted with through a web application. The current design of the application is a traditional-style file listing of resources by type, with color coding to show status. The next step in development of this application is to show visualizations of data relationships in a graphical way which will increase comprehension.

This study explores the usability of different data visualization methods to recognize relationships between data types. Spatial data relationships may be displayed in multiple ways, such as server and database, database and service, or service and application. Three visualization methods including collapsible tree, network graph, and left tree are further explored to determine the usability of the visualization method. These three methods were chosen for their popularity, ease of use, and help documentation available.

The ISO 9241-11 standard defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (World Wide Web Consortium, 2002). For this study, usability will be measured by time, comprehension, and user preference.

Methods

A web-based survey was utilized to target a wide range of volunteers. Online questionnaires are good for reaching many people quickly if they are geographically dispersed (Rogers, Sharp, and Preece, 2013). The survey focused on displaying data relationships graphically to volunteers to assess their comprehension of the visualization presented.

Website Design

The main framework of the survey website utilized an open source toolkit called Bootstrap. Bootstrap is one of the most popular front-end frameworks for building responsive websites with hypertext markup language (HTML), cascading style sheets (CSS), and JavaScript (JS) (Bootstrap, 2017). The JavaScript library Data-Driven Documents (D3) was utilized to create the different data visualizations. The D3 library requires minimal overhead and supports large datasets with dynamic behaviors for interaction and animation through a diverse collection of modules (Data-Driven Documents, 2017).

The first visualization presented was the collapsible tree (Figure 1).

Figure 1. Collapsible tree visualization.

Thomas (2015) states tree maps represent numeric data with two-dimensional areas, and they indicate hierarchies by nesting subordinate (children node) areas within their parents (parent node). A parent node may be clicked on to reveal any children nodes related to it or clicked to hide any children nodes. The presentation of this diagram was vertical with children nodes indented under parent nodes. Nodes were represented by rectangle boxes containing a label for which data resource the box represented. Visually, nodes were a darker color when they contained children which were not shown.

The second visualization presented was the network graph (Figure 2). Network graphs represent objects, generally known as points (nodes) with lines (edges) connecting points to show relationships (Thomas, 2015). For this visualization the nodes were labeled for what data resource they represented. When a node was clicked on, all related nodes and lines changed colors to show a direct relationship. This allowed the volunteer to further explore relationships within the diagram.

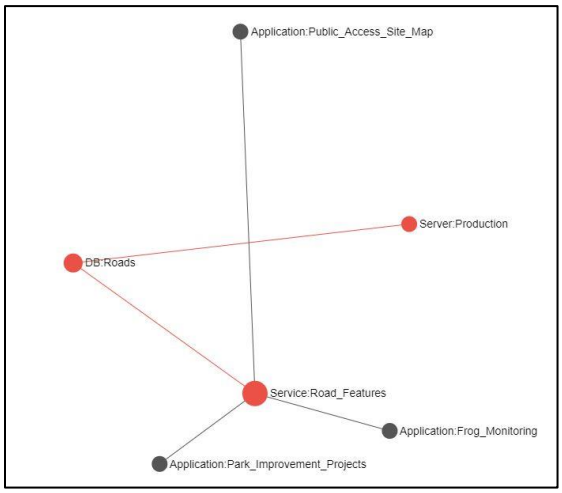

Figure 2. Network graph visualization.

The last visualization presented was the left tree (Figure 3). Like the first visualization presented, this is also a tree map. For this visualization the nodes were represented by circles with each node labeled with the data resource it represents. The presentation of this diagram was horizontal with children nodes appearing lined up to the right of the parent node. Relationships between the parent and children nodes were represented by a line connecting them. Nodes were represented as a darker color when they had hidden children nodes.

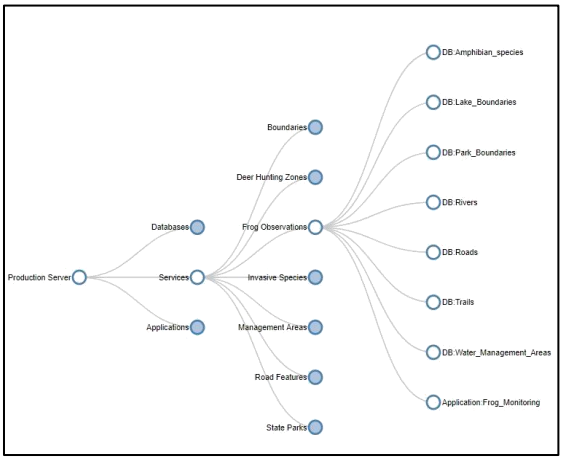

Figure 3. Left tree visualization.

Volunteers

A list of potential volunteers was generated from the researchers' professional contacts which included individuals from the DNR and other places. Contacts were sent an email inviting them to participate in an anonymous survey where they would not be compensated. If a contact was interested, they were directed to the survey website. Contacts were also encouraged to share the survey website with others to increase survey responses. Saint Mary's University of Minnesota Research Review Board (RRB) approval was gained so this study could utilize data gathered from human volunteers.

Survey Design

Once volunteers agreed to participate, they were presented with a questionnaire to gather background demographic data. Questionnaires are a well-established technique for collecting demographic data and users' opinions (Rogers *et al*., 2013). Volunteers were asked about their age, gender, education level, field of work, and computer usage in the form of multiple choice questions. Multiple choice answers provided standardized responses for quantitative analysis of the results.

With each visualization presented, a series of three multiple choice questions were asked to assess the volunteers' comprehension of the visualization. The questions tried to mimic real world scenarios which these visualizations may be used for. The three questions were as follows:

- You find out the Service named <service name> is down. How many Applications are affected?
- A user would like the schema changed on the Database named <database name>. How many Services may be affected?
- How many Services does the Application <application name> utilize?

For each section of the survey a timestamp with the date and time was generated to measure the time spent on the section. The volunteer was asked if they were interrupted while viewing the page and, if so, to select a time in minutes for which they were interrupted for to account for long page viewing times. Before moving to the next visualization, the volunteer was asked to provide any comments about the visualization in a text field.

Upon completion of all three visualizations they were asked two multiple choice questions as to which visualization was the easiest to use and which visualization was the most difficult to use. The volunteer was then given the opportunity to provide any final comments in an open text field. On submit, survey results were submitted and stored in a MySQL database.

Results

The survey was open to responses for approximately two weeks. A total of 44 individuals responded with 77.27% who identified as female, 20.45% as male, and 2.27% who preferred not to identify. Of those who volunteered, only 33 completed the survey fully. Reported ages were normally distributed with four age groups represented: 38.64% aged 25 to 34, 20.45% aged 35 to 44, 15.91% aged 45 to 54, and 25% aged 55 to 64 (Appendix A). Education level of volunteers was normally distributed with 70.45% holding a Bachelor's degree or higher. Highest reported occupations were "Other" at 29.55%, "Geographic Information Science Professional" at 27.27%, and "Business / Clerical" at 15.91%.

Time

For the 33 volunteers (n = 33) who completed the survey fully, total time to complete the survey was not normally distributed with a skewness of 1.875 (SE = 0.409) and kurtosis of 3.528 (SE = 0.798). Average time to complete the survey was 23.15 minutes, with a median of 13.15 minutes, and standard deviation of 23.29. Average times for each visualization were 10.12, 3.91, and 7.99 minutes for collapsible tree, network graph, and left tree, respectively.

The ten volunteers who failed to complete the survey spent an average time of 17.77 minutes on the survey before

quitting. Volunteers quit during each section of the survey, with the most quitting during the network graph visualization (Figure 4).

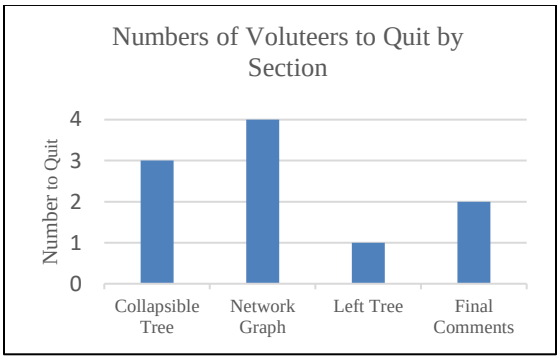

Figure 4. Volunteers to quit by each survey section.

Since average time to complete the survey was found to not be normally distributed, nonparametric tests were utilized to determine any correlations between survey variables. Spearman's rank-order correlation was utilized to determine the relationships between the following: age and time, education and time, and occupation and time. All results were found to not be statistically significant. Age and time showed a weak correlation ($r_s(31) = 0.324$, p = 0.066, two-tailed). Education and time showed a weak negative correlation ($r_s(31) = -0.261$, p = 0.142, two-tailed). Occupation and time showed a very weak negative correlation ($r_s(31) = -0.024$, p = 0.893, two-tailed).

Comprehension

Comprehension scores for all three tests combined were normally distributed with a skewness of -0.946 (SE = 0.409) and kurtosis of -0.272 (SE = 0.798). The average total score was 69.02%, with a median of 66.67%, and standard deviation of 33.99. Highest comprehension rates occurred with the collapsible tree visualization with an average score of 76.77%. The left tree visualization was close behind with an average score of 74.75%. Lowest comprehension occurred with the network graph with an average score of 55.56%.

Comprehension scores between each of the survey questions for a given visualization varied the most with the collapsible tree with a range of 11.36%. Comprehension was equally consistent with both the network graph and left tree visualizations with a range of 2.27% (Figure 5).

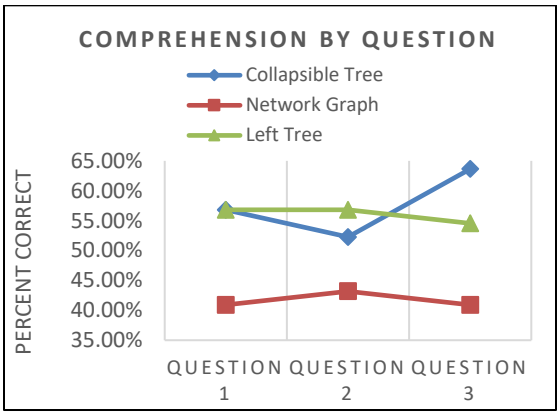

Figure 5. Comprehension scores by question.

Spearman's rank-order correlation was run to determine the relationships between volunteers' total time and overall comprehension score (Figure 6). Results were found to not be statistically significant and show a very weak negative correlation between time and score, ($r_s(31) = -0.181$, p = 0.313, two-tailed).

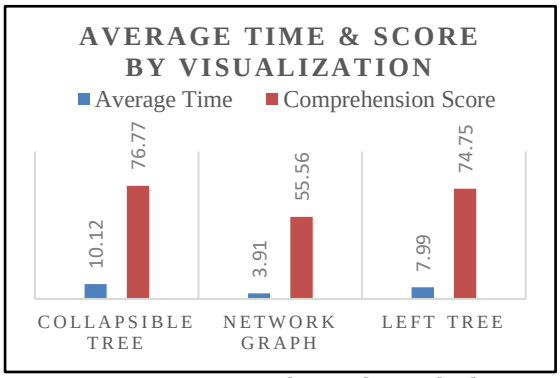

Figure 6. Average time and score by method.

Additional Spearman's rank-order correlations were utilized to determine relationships between each test score and age, education, and occupation. Results were shown to not be statistically significant with weak correlations (Appendix B).

User Preference

Overall, 51.52% of volunteers preferred the left tree visualization with 42.42% of volunteers preferring the collapsible tree. Over half of the volunteers, 60.61%, reported the network graph visualization was the hardest to utilize.

Discussion

Usability for this study was measured by time, comprehension, and user preference. A scoring matrix was constructed by assigning a score of 3 for highest and 1 for lowest scores in each measurement area (Table 1). The lowest average time, highest comprehension score, and highest preference receive the highest scores.

Table 1. Usability scoring matrix.

	Collapsible Tree	Network Graph	Left Tree
Time	1	3	2
Comprehension	3	1	2
Preference	2	1	3
Total Score	6	5	7

The results of the scoring matrix show the left tree has the highest usability score of seven with the network graph showing the lowest score of five. This usability score is supported, especially by the volunteers, where six volunteers specifically noted how easy the left tree visualization is to use.

Errors

Volunteers took the survey independently without in-person instruction or intervention which may be the cause for user error. Written instructions orienting the volunteer were presented with each visualization, and volunteers should be considered novice users because they have not used these visualizations previously. Time calculations started with the collapsible tree visualization. Times for this visualization were the highest and most likely due to users orienting themselves to the survey. Further instruction, including a short video demonstration, about the visualization prior to the timer starting may have reduced time for this section. Additionally, visualizations could have been presented in a random order to further reduce this orientation time bias.

Volunteers particularly struggled with the network graph visualization. Through user comments, it appears at least six volunteers did not read the instructions which informed them as to how to use the visualization. As a result, those volunteers were unable to interact with the graph at all, reducing effectiveness of this visualization. Additional reduction of usability also occurred when the display of the network graph did not always appear correctly. For particularly large displays the nodes sometimes overlapped, and users could not see all the data relationships.

Errors with technology also reduce effectiveness of usability. While the survey site was well tested prior to deployment, some users reported an SQL error when submitting their survey. This error occurred when volunteers utilized single or double quotes in the open comment fields provided for each visualization. Error handling for these

fields did not include an escape for these characters causing the affected text fields to not be transferred correctly to the database and the user would see an error appear on the screen. Some volunteers were able to revise their comments to remove any quotation characters and then re-submit to complete the section. Others were not sure what to do and quit the survey which would account for volunteers quitting at all stages of the survey. Due to finding this error during the open survey period, no changes were made to correct the issue, so the user experience would remain the same and results would be consistent.

Future Development

The colors utilized in the data visualizations, which included blues, grey, and reds, may also have lowered usability. Bresciani and Eppler (2008) state the meaning of symbols and colors are not universal, and some graphic representations may be misinterpreted in other cultural contexts. Therefore, the colors displayed and the symbols utilized may have affected the volunteers' interpretation of a visualization. One volunteer specifically commented that different colors may have increased satisfaction by making visualizations easier to read. Further development should explore the use of different color pallets as well as different shapes for symbolization.

With this survey, all three visualization method usability scores were close, ranging from five to seven. Usability changes as users become more experienced. Further development should include multiple usability tests given to the same users over time in random order. This would provide a more accurate usability score as time and comprehension scores would be further refined. Additionally, over time, as users become more experienced, their preference for each visualization method may change.

Conclusion

As spatial data resources increase, it is important to manage these resources effectively. Relationships between data sources may be easier recognized by using graphical visualization methods explored by this study, such as the left tree visualization method. Visualization methods incorporated into data management systems help spatial data administrators manage data more effectively.

Acknowledgements

Many individuals assisted with the development of this research project from the beginning through the end. I would especially like to thank the entire MNIT GIS Operations staff at the DNR including Hal Watson, Zeb Thomas, Chris Pouliout, Mike Tronrud, and Kitty Hurley. I would also like to thank my Saint Mary's University of Minnesota graduate advisor, Greta Poser.

References

Bootstrap. 2017. Bootstrap Homepage. Retrieved October 1, 2017 from https://getbootstrap.com/.

Bresciani, S., and Eppler, M. 2008. The Risks of Visualization: A Classification of Disadvantages Associated with Graphic Representations of Information. *Institute for Corporate Communication*, ICA Working Paper # 1/2008.

Data-Driven Documents. 2017. Data-Driven Documents Homepage. Retrieved October 1, 2017 from https://d3js.org/.

Gui, Z., Yang, C., Xia, J., Liu, K., Xu, C., Li, J., and Lostritto, P. 2013. A Performance, Semantic and Service Quality-Enhanced Distributed Search Engine for Improving Geospatial Resource Discovery. *International Journal of Geographical Information Science, 27(6),* 1109-1132.

Gui, Z., Cao, J., Liu, X., Cheng, X., and Wu, H. 2016. Global-Scale Resource Survey and Performance Monitoring of Public OGC Web Map Services. *International Journal of Geo-Information,* 88-112.

Morris, S. 2006. Geospatial Web Services and Geoarchiving: New Opportunities and Challenges in Geographic Information Services. *Library Trends, 55(2),* 285-303.

Rogers, Y., Sharp, H., and Preece, J. 2013. *Interaction Design: Beyond Human-computer Interaction 3rd edition.* United Kingdom: John Wiley & Sons Ltd.

Thomas, S. 2015. *Data Visualization with JavaScript.* San Francisco, CA: No Starch Press, Inc.

World Wide Web Consortium. 2002. Usability – ISO 9241 definition. Retrieved September 17, 2017 from https://www.w3.org/2002/Talks/0104-usabilityprocess/slide3-0.html.

Appendix A. Summary results by age.

		Age Group			
		25-34	**35-44**	**45-54**	**55-64**
	Count of Volunteers	17	9	7	11
	# Incomplete Surveys	4	0	3	4
Sex	Female	13	7	7	7
	Male	4	2	0	3
	Prefer not to answer	0	0	0	1
Education Level	no answer	0	0	0	1
	High school	0	1	0	1
	Some College	1	0	0	1
	Associates Degree	1	4	1	2
	Bachelors Degree	8	1	1	4
	Masters degree	7	3	5	2
Employment	no answer	0	0	0	0
	Business / Clerical	3	2	1	1
	GIS Professional	6	3	3	0
	Industrial / Manufacturing	0	0	0	2
	IT Professional	0	1	1	2
	Non-IT related field	1	1	0	0
	Project Management	1	0	0	0
	Student	1	0	0	1
	Website Development	1	0	0	0
	Other	4	2	2	5
Computer Usage (per day)	1-3 Hours	2	0	2	1
	4-6 Hours	0	2	0	2
	7-9 Hours	13	4	3	6
	10-13 Hours	2	3	2	1
	14+ Hours	0	0	0	1
Data Creation	Yes	14	7	5	8
	No	3	2	2	3
Visualization Comprehension (% of correct answers)					
Collapsible Tree	Question 1	75.0%	78.0%	100.0%	60.0%
	Question 2	62.5%	77.8%	66.7%	60.0%
	Question 3	81.3%	77.8%	83.3%	80.0%
	Average	**72.9%**	**77.9%**	**83.3%**	**66.7%**
Network Graph	Question 1	43.0%	67.0%	17.0%	6.0%
	Question 2	43.0%	78.0%	33.0%	7.0%
	Question 3	43.0%	67.0%	33.0%	6.0%
	Average	**43.0%**	**70.7%**	**27.7%**	**6.3%**
Left Tree	Question 1	77.0%	67.0%	100.0%	75.0%
	Question 2	77.0%	67.0%	100.0%	75.0%
	Question 3	69.0%	67.0%	83.0%	88.0%
	Average	**74.3%**	**67.0%**	**94.3%**	**79.3%**

Appendix B. Spearman's correlations.

		Age	Education	Occupation
Collapsible Tree	Correlation Coefficient	-0.049	-0.078	-0.230
	Sig. (2-tailed)	0.786	0.667	0.198
	N	33	33	33
Network Graph	Correlation Coefficient	0.060	0.118	-0.115
	Sig. (2-tailed)	0.739	0.513	0.523
	N	33	33	33
Left Tree	Correlation Coefficient	0.079	-0.03	-0.298
	Sig. (2-tailed)	0.664	0.868	0.092
	N	33	33	33
Total Score	Correlation Coefficient	-0.020	0.153	-0.273
	Sig. (2-tailed)	0.91	0.396	0.124
	N	33	33	33