Project 3: Visualizing DHQ bibliography

Reviewer 1:

1. forethought of using pruning to remove incomplete data, but would that not skew the final results?
2. I like the clear disambiguation of culture with respect to DH
3. well documented process, but could have been part of the appendix.
4. a lot of effort was put into data collection. The times wasted on data collection while unfortunate ensures a more thorough data set.
5. interesting insights and problems (and their respective solutions)
6. very interesting "keywords derived from the titles and abstracts may provide more insight into the values and ideas represented in individual cultures, than the thematic classification provided by the client"
7. bibliographic coupling and co-authorship networks were confusing, as i wasn't sure what I was looking at. The color scheme while nice failed to illustrate difference between them. (This may have been due to the images being cut-off on the right)
8. I like the word cloud as a representative (or indicators) of culture within different clusters. These representations actually make more sense to me than a random network arrangement.
9. "Although we believe our visualizations provide insight into how authors and papers relate to one another within DHQ, it is difficult to say whether we can describe this as insight into cultures" then how does the visualization help?

Reviewer 2

1. Quality of data selection, cleaning, preparation and documentation

Very thorough description of the effort involved in mining and cleaning the data, as well as reconciling and correlating data using other sources such as scraping web sites.

Good introduction of what the data is, and what the client's visualization goals are. I also found it helpful that the authors took the time to define the business domain of the data, including terms such as Digital Humanities and "big tent".

My only complaint with the documentation is that it is too long. The project instructions clearly stated it should fit on 4 standard pages, but the document is 17 pages long with very tiny font I found difficult to read. It needs some serious editing to cut it down.

2. Appropriate selection of tools, algorithms, workflows, and parameter values

It was very helpful to the reader that they described the tools they used (R, Sci2, OpenRefine) to mine and clean the data. I was also very impressed with the detailed documentation of workflow

steps, which would make the process repeatable to anyone else who wished to try to achieve similar results.

It was great that they described how they used R to generate the word clouds.

3. Quality of data analysis, visualization results, and discussion if insights gained

Excellent data analysis and descriptions of missing data, and what steps they did to fill in the missing gaps.

The included visualizations make good use of color, and the legends are very helpful to understand the networks.

Extremely thorough list of insights gained from the extracted co-author and co-citation networks and word clouds.

4. Completeness and quality of validation and redesign

The team went above and beyond in describing their data validation issues and what they did to address them. They even included such detail as specific duplicate author names with variations in spelling, and how they were unified to make a consistent dataset. This is an impressive level of detail and thoughtfulness in documentation.

5. Overall quality of content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas using text and visualizations, and the appropriateness of references to/attribution(s) for the work of others

This is an excellent paper, the team deserves an A+.

The authors did a very good job in describing the data The document is very clear and well written.

Extremely thorough use of references throughout the documentation. And very detailed discussion of the work of others with respect to academic cultures.

Reviewer 3

Quality of data selection, cleaning, preparation, and documentation:

The group understood the issues with the data they were given. They did a good job cleaning and enhancing the information to do better analysis.

Appropriate selection of tools, algorithms, workflows, and parameter values:

The tools selected were appropriate for the task. Maybe more could be covered with what was done in R and OpenRefine.

Quality of data analysis, visualization results, and discussion if insights gained:

The network graphs are looking good. Instead of a word cloud, did you try to do anything with topic burst or topic over time, or at least some graph with a scale?

Insights the group gathered from the visualizations were well covered.

Completeness and quality of validation and redesign:

The redesign looks good. I'm not sure if any nodes can be trimmed (top 25 nodes, etc.)from the Bibliographic coupling one to make it clearer.

Overall quality of content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas using text and visualizations, and the appropriateness of references to/attribution(s) for the work of others:

It was a very well written paper and a pleasure to read. As a reader with no prior knowledge, the background to the projects was covered enough so I came away understanding more.

They understand the limitations of the data the client brought in trying to achieve the project goals.

Reviewer 4

1. The quality of the data selection appears to be high. The group appears to have grappled with the linking of authors and institutions in a straight-forward manner. Author/institution data seems to be the linchpin of the project as the group stated that, "we are interested in cultures, which is best represented by people rather than articles."
2. Tool usage appears to be appropriate to the undertaking and the group has taken steps to identify tools and routines that bridge identified gaps in analysis needs.
3. The data analysis itself appears to encompass the objectives by identifying co-author networks where the flow of ideas can be traced. I particularly liked the section that identified 'document pairs that have been cited together most frequently'. This is useful in identify what I would call a keystone paper – one that appears to be at the center of a movement and influencing future related work.
4. Data validation appears to have been incisive : the members of the group have decided to deviate from the 'thematic' nature of the cultural analysis and instead, chose to focus on the keywords from the articles analyzed:

"We believe that the keywords derived from the titles and abstracts may provide more insight into the values and ideas represented in individual cultures, than the thematic "classification provided by the client.

This appears to resonate with the idea of analyzing a culture; due to how close relationship between language and culture (ie: ontologies).

5. Overall Quality: On a Likert scale 1- 5; 1 being low quality and 5 being high quality. I rate the overall quality a 4. The only thing I would say would enhance this analysis would be to correlate the networks of co-authors based on keywords to where on the maps those keywords are dominant and networks are forming.

I am not sure if the group found that it was less helpful – but there was an explicit statement about analyzing the 'geographic nature' of the development of the culture. It could be that this might be irrelevant. That would be my only question for the group.

Reviewer 5

1. Data selected appears to support their needs. Meta-data descriptions of the field Auditors/AuditUser would be helpful to understand how this field functions. I assume that this is main field to be analyzed as this field appears to log access to the database.
2. The proposed visualization appears to support their goals. A circular tree seems to be a logical choice for this type of visualization.
3. The data analysis itself appears still be in progress. What is meant by 'Understanding the data objects and aligning it to the requirements of this project"?
4. The group appears to have identified the Javascript library D3 as the solution to working with a large dataset. Does D3 process data or visualize the data? What will they use to process the JSON file?
5. I do have a good sense of what they are attempting to visualize: departmental access to database objects over time. Not enough information to give an overall assessment though. Need a little more information about the processing of the xls file to JSON.

Reviewer 6

First and foremost, I am really impressed with their document. It's really impressive. All little details and every aspect are very well documented. The dataset that they have and their goals mentioned is also quite interesting. They do mention how exactly do they carry out data cleaning and deal with the missing fields. The team is aware of the possible challenges and suggests solutions for the same; that is really awesome. The related work in the field have been properly cited and explained. The visualizations and the insights gathered are really interesting. I especially liked when they gave out different insights for all the individual visualizations that they worked with. I think the only thing I found missing in this document is the mention of the various tools that they used. I am guessing they used SCi2 for the visualization from the background, but not too sure about it. Overall, I found their work to be complete. They have mentioned some sort of future work which I didn't come across in the other two projects that I reviewed. So kudos for that!!

Reviewer 7

1. Quality of data selection, cleaning, preparation, and documentation

There appears to be much work required for data hunting as the data seems to be inadequate for the research. There were inconsistencies in the data. They were quick to identify that the thematic classification would not work for their analysis and approached the client for alternate routes. The validation was well documented.

2.Appropriate selection of tools, algorithms, workflows, and parameter values

The tools were good. The team faced much difficulty with the shortcoming of the data. So, the algorithm used to generate clusters and prepare data was good.

3.Quality of data analysis, visualization results, and discussion if insights gained

The data analysis needs more work. The team faced challenges defining the use of cultures and using it in their analysis. So in order to paint a better picture, maybe we need more visualizations.

4.Completeness and quality of validation and redesign

There needs more work in the redesign aspect. The validation could be improved.

5.Overall quality of content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas using text and visualizations, and the appropriateness of references to/attribution(s) for the work of others

The visualizations presented were very good. It is a sigh that there was no more data. Otherwise, we could have got some interesting insight. The team managed well with the data they had and adapted to the change. But I feel there needs more work in the redesign.

Reviewer 8

Rubric	Rating (1-5) 5=Great	Feedback
1. Quality of data selection, cleaning, preparation, and documentation	4	Data selection is good but more data cleaning/preparation need to be discussed?
2. Appropriate selection of tools, algorithms, workflows, and parameter values	5	Very good data statistics overview with sufficient network descriptive parameters.
3. Quality of data analysis, visualization results, and discussion if insights gained	5	Very good network graph
4. Completeness and quality of validation and redesign	5	It's good that you get to communicate their data choice and design with the client.
5. Overall quality of content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas using text and visualizations, and the appropriateness of references to/attribution(s) for the work of others	5	Insight is good but the network graph will also need some sort of verbal description, right?

Reviewer 9

1. Quality of data selection, cleaning, preparation, and documentation
    1. Good description of the data
    2. Significant effort spent on cleanup, ie scraping the DHQ site to obtain missing data
2. Appropriate selection of tools, algorithms, workflows, and parameter values
    1. A large variety of tools used: R, Sci2, Openrefine
    2. A good set of additional variables created
3. Quality of data analysis, visualization results, and discussion if insights gained

    1. In discussion of related work you can show some similar visualisations, as discussed in the course and/or other publications
    2. Very well done co-author network. What criteria is used to (not) show author name besides node?
    3. co-citation network looks ok but there is room for improvement:
        1. meaning of the color of the nodes? (relate to word clouds)

        2. overall graph looks rather chaotic, perhaps rearrange
        3. numbers at edges, mostly 1.0, can be removed, or use as thickness/color of edge
        4. Legend missing with full reference (name)
    4. Word clouds for co-citation network: looks nice aesthetically, but what are the valuable insights gained from it?
2. Completeness and quality of validation and redesign
    1. very extensive documented stats and method, perhaps a bit too extensive, a summary might be sufficient, or complete log could be included as appendix
    2. Disciplinary identification of authors: you identified 24 unique cultures in the datasets, anyway to represent these in a graph?
    3. Redesigns look very promising, could be enhanced with some author names and/or specific characteristics of the clusters
3. Overall quality of content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas using text and visualizations, and the appropriateness of references to/attribution(s) for the work of others
    1. High quality of content and very well structured!
    2. Review the document on duplicated info, some of the information is duplicated in several places, can be cleaned up for the final version

Reviewer 10

1. *Quality of data selection, cleaning, preparation, and documentation*

The process of data selection is clearly described. Encountered problems in the data is mentioned as well. The strategy of how to address these issues is mentioned too.

2. Appropriate selection of tools, algorithms, workflows, and parameter values

Data analysis and algorithms are documented. The taken steps of the analysis are described in detail. The selected tools are appropriate and the outcome has been described clearly too.

3. Quality of data analysis, visualization results, and discussion if insights gained

Data analysis is well executed: appropriate statistics of the used datasets have been provided and are documented. Also, the workflow of data analysis is described thoroughly.

4. Completeness and quality of validation and redesign

Encountered problems in the datafiles, like missing articles and duplicate authors have been fixed accordingly. You appear to have taken much effort to deliver a quality dataset. Excellent approach: a dataset should be of as high quality as possible!

Concerning redesign, this section could have been described more in detail. Visualisations of this redesign are provided, but they lack additional explanation. For example, what are the differences compared to the original design and why is it better?

5. Overall quality of content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas using text and visualizations, and the appropriateness of references to/attribution(s) for the work of others.

Overall quality of content is excellent: information appears to be complete, which includes many references which seem to be appropriate. Your ideas are well visualised and described by providing necessary hand sketches and accompanied descriptions. The section which mentions future directions seems to be just a listing of some opportunities. I suggest to describe this points a bit more in detail, if possible.

Reviewer 11

1. **Quality of data selection, cleaning, preparation, and documentation**

Clearly an extraordinary amount of effort went into filling data gaps. One key piece of information I was missing that wasn't mentioned up front was what time period the data cover. Also, was there not some other database (e.g. Web of Science) that contained all parameters you needed for every article published in the journal?

2. **Appropriate selection of tools, algorithms, workflows, and parameter values**

The workflows were very clearly articulated and seemed appropriate. I wonder if you'd considered using a topic modeler on the abstract text instead of/in addition to SLM community detection? There's a free one available that's very straightforward to use here: https://code.google.com/p/topic-modeling-tool/

3. **Quality of data analysis, visualization results, and discussion of insights gained**

All the relevant statistics seemed to be listed and the initial visualizations looked good, but I think you could say more about what the statistics and visualizations together really *mean*. For instance, what conclusions can you draw (or start to draw) about culture?

4. **Completeness and quality of validation and redesign**

Having more complete information from the client seemed to help. I would be curious to see the diagrams drawn in the same way, however, but with the updated information.

5. **Overall quality of content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas using text and visualizations, and the appropriateness of references to/attribution(s) for the work of others**

A lot of good and detailed work has gone into this project, and it sounds like there's a full, accurate data set to work with now. Attribution seemed detailed and appropriate. What's really missing is a discussion of the results that brings together statistics and visualizations and provides insights into what they mean and to what extent the client's questions can be answered

at this point. I see that you said you're not convinced you can answer the client's questions with what you have, but what insights (beyond showing the visualizations and listing the statistics) can you make?

Reviewer 12

1. ***Quality of Data Selection:*** There was a clear discussion of why the authors chose to focus on the individual interactions of authors within the data set. However, I believe more could have been done--or clearly shown in the report--regarding the basic features and descriptive patterns in the data. For instance, identification of author occurrences and related records can provide insights into what needs to be cleaned and removed in the data set as not useful. In the final version, I would suggest either taking steps to do this or, if it was done, clearly communicate this within the final report.
2. ***Appropriate Selection of Tools:*** The use of network analytic graphics was justified in this data set. I would have liked to see more connections, though, between the author interactions and the content or topical analyses through word clouds.
3. ***Insights:*** While these are initial result and largely descriptive, I am left wondering what these descriptive aspects answer in terms of the clients needs and requests. How does this information assist the client in making decisions regarding the geospatial and temporal patterns that emerge in the work of digital humanities? I believe more could have been said here and suggest much of the focus be on these issues for the final report.
4. ***Validation:*** The validation was very thorough, but I was left with questions regarding the clients' input in the validation process. Consultation with the client in the validation process is often helpful, and I would recommend discussing the findings and validation process with them.
5. ***Presentation and Communication:*** Over all, what was contained in the report was clear, concise, and somewhat justified. I would have liked to see more discussing and defense of the algorithm and data selection process. This assists in making your inferences stronger and more refined. I suggest adding such considerations to the final report. I would also try to remove or reduce the Sci2 output. This output appeared unnecessary to me and, depending on the client and related shareholders, may easily get in the way of communicating the relevance of the result. If it is left in, I think a clear purpose of this output needs to be communicated in the report. In general, though, a strong effort and look forward to seeing the final results.