



GOOGLE SCHOLAR CITATIONS VISUALISATIONS



Chanta Yang;Lauren Sanby;Ion Todd

VIS MODULE

TABLE OF CONTENTS

Introduction	1
Research.....	2
Alluvial Diagram	2
Citeology	3
Microsoft Academic Search	4
Citation Map	6
Visualisation	9
Visual Queries	9
Initial Design.....	9
Final Design	13
Discussion of Design Strengths and Weaknesses	15
Strengths	15
Weaknesses and Challenges	17
Conclusions	17
Future Work	18
Ideas from Second Presentation.....	18
Team contributions.....	19
References	20

INTRODUCTION

Google Scholar is a search engine used to look up scholarly literature such as journal articles, white papers, theses and books across various disciplines. It allows researchers to see the number of citations for a specific publication as well as a list of related articles. That being said, Google Scholar does not include the number or list of references used in the publication. Thus, users have to directly visit the online database where the paper is stored to view this information. Both aspects of the number of citations and number of references are important measures for determining the quality of a paper.

In addition to searching for papers, Google Scholar allows authors to create public profiles and monitor the number of citations to their publications. Author public profiles display a list of papers published by the author along with the number of citations for each paper and the year in which it was published while an aggregated figure of number of citations for each year is shown in a bar graph.

While researchers are interested in number of citations and number of references of a paper, authors have an interest in the number of citations, topic fields in which their paper is cited and from which institution citing authors originate.

With this in mind, it was thought that a visualisation of the networks between Google Scholar citations (cited by) and references (cited) organised by geographic location and discipline be created for use by both academics and Google Scholar users. This would be done for a single paper of the author.

A visualisation such as this would allow:

- Researchers to easily quantify number of citations in comparison to references for a specific paper
- Academics to identify interested authors for future collaborations.
- Users to identify in which other disciplines the paper has value
- Users to identify where the paper has referenced and been cited from.

Data for this visualisation will be gathered from Google Scholar (number of citations), Google Scholar Metrics (sub-category categorisations for discipline classification) and online journal databases (reference list).

RESEARCH

Before the final design was implemented, research was done into existing citation visualisations. These were found to be in a static and interactive form. Static visualisations included images which simply display the data and interactive visualisations involved interfaces which allow the data to be further explored. The following designs were analysed and their effectiveness evaluated.

ALLUVIAL DIAGRAM

An alluvial diagram is a type of network flow diagram that links different dimensions of information to each other. This example (Figure 1.) applies to the Communication, Rhetoric and Digital Media field and shows which authors published papers in which years. The name of the author appears on the left hand side while the right hand side shows the year in which a paper was published. A line is used to link the two. The thickness of the line indicates how frequently an author has been cited - the thicker the line, the more citations a paper has. Different colours for the lines have been used to represent individual authors.

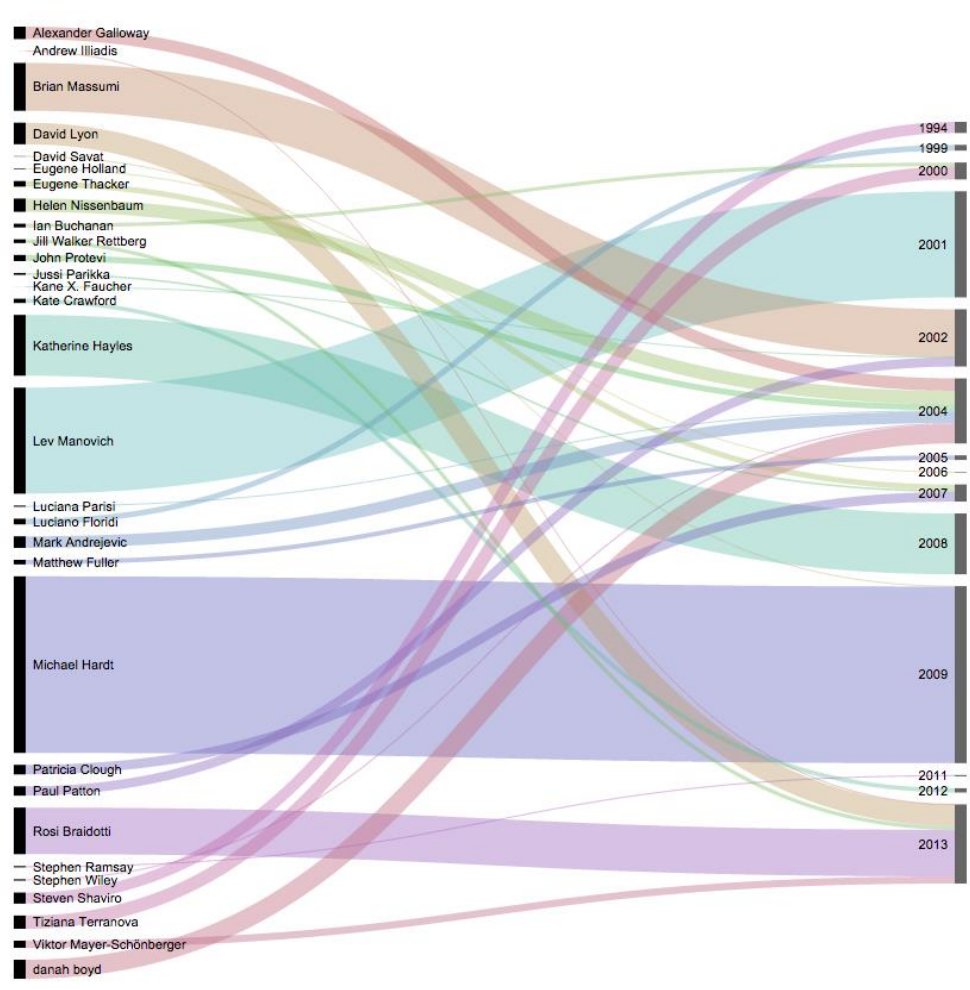


Figure 1 – Alluvial Diagram by J.J Sylvia IV (Sylvia IV, 2015)

From this visualisation, the visual query of "Which author is the most cited and what year was their paper published in?" can quickly be determined by looking for the thickest line on the diagram. One can also easily determine how many papers were published in each year by counting the numbers of lines connected to a certain year.

Drawbacks of this design are that the colours of some of the lines are similar in shade and overlap making it difficult to follow them from the author to the year. Additionally, very thin lines are difficult to see.

CITEOLOGY

Citeology (Autodesk Research, n.d.) is a Java applet that takes papers from the CHI and UIST Human Computer Interaction (HCI) conferences between 1982 and 2010 and visualises citations and references of the most cited paper in each year. The papers for each year are represented as bars of text (the first few sentences of a paper's title) on a timeline starting from 1982 - 2010 where height of the bar indicates how many papers were written in that year. The mid-point of each bar represents the most cited paper from a particular year.

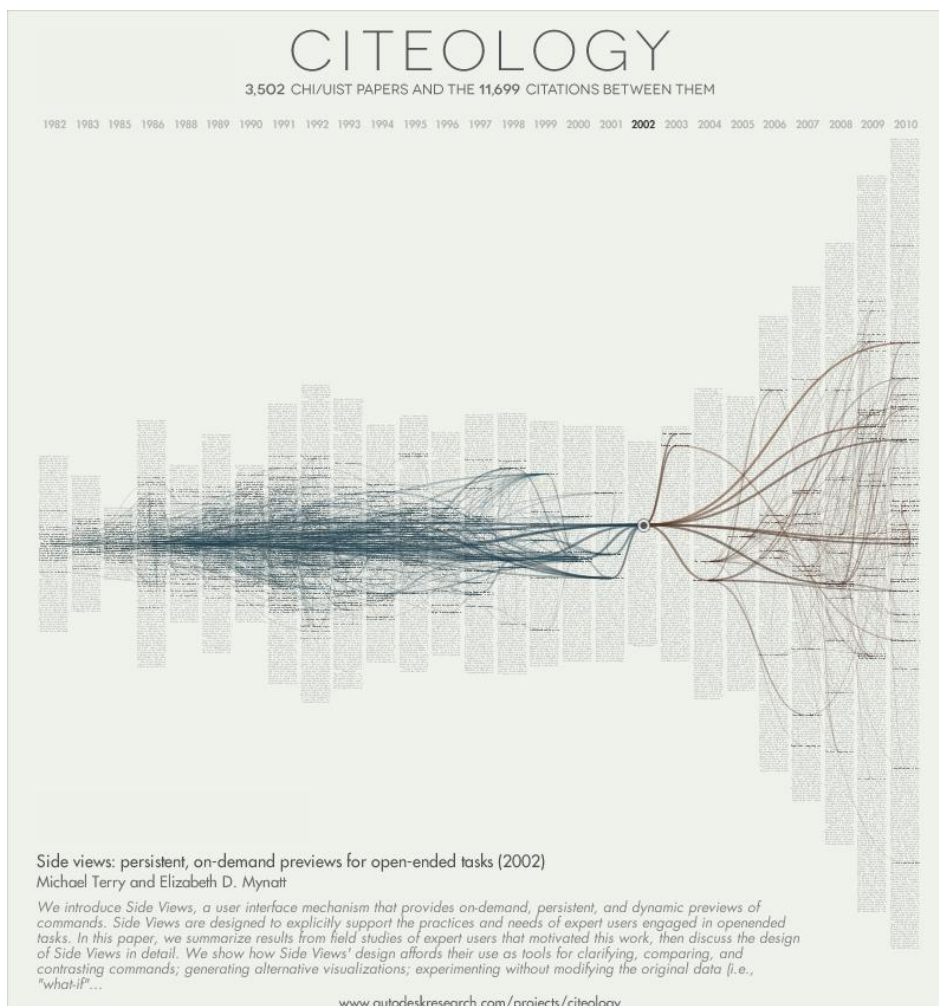


Figure 2 - Citeology Visualisation of the Paper "Side Views: Persistent, on demand previews for open-ended tasks (2002)" (Autodesk Research, n.d.)

This visualisation (Figure 2.) is similar to the previous example of an alluvial diagram in that a network of lines links older referenced papers and newer cited by papers to the chosen paper (represented by a dot on the bar graph). Blue lines link papers referenced by the chosen paper (and the papers they, in turn, referenced) and brown lines link to papers that used the chosen paper as a source (and other papers which cited them, thereafter). Essentially, the visualisation provides insight to the "genealogy" of a paper with "ancestors" (references) and "descendants" (citations).

While this visualisation allows for a quick assessment of the number of references for a paper versus number of citations and comparison of the number of papers from year to year, the diagram quickly becomes crowded - as more references of references are shown (and vice versa: citations of citations), the occlusion of lines makes it more difficult to coherently judge which papers link to which. This, in turn, makes it difficult to extract details such as the names of papers.

MICROSOFT ACADEMIC SEARCH

Similar to Google Scholar, Microsoft Academic Search is an experimental academic search engine created by Microsoft Research to research the ways in which scientists, academics and students find academic content. Although the website explores concepts of data mining, entity linking and visualisation, the site has not been updated since 2013 and will likely be taken down once research goals have been met (Microsoft Academic Research, 2013).

CITATION GRAPH

Citation Graph is an interactive visualisation created by Microsoft Academic Search (2013) that shows citation relationships between authors: nodes in blue show citing authors while the single orange node shows the main author. Distance and size of the node to the main author indicates how often the citing author has cited the main author. The exact number of citations from one author to another can be seen by mousing over the edge between the two nodes.

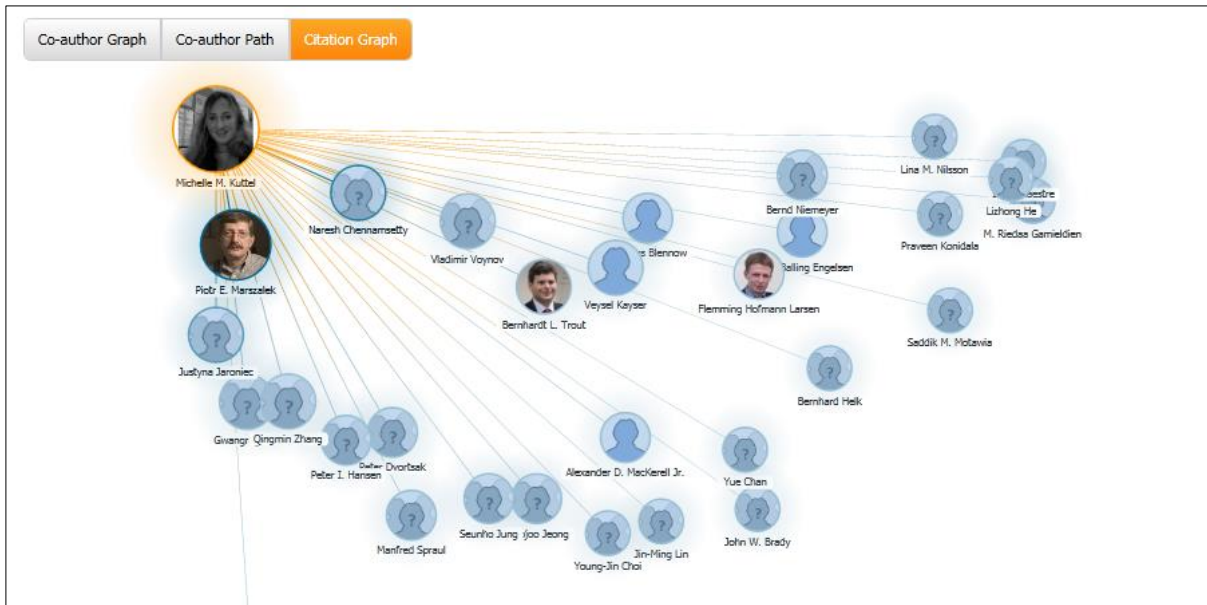


Figure 3 – Citation Graph (Microsoft Academic Research, 2013)

The graph is effective at displaying exactly which author cited the main author and number of citations from author to author but not at showing the exact total number of citations to the author. Additionally, distance and size of node are not adequate means of distinguishing number of citations from the main author as nodes can be moved around via clicking-and-dragging thereby skewing this visual queue. There is also the problem of occlusion to consider as the number of citing authors increases.

ACADEMIC MAP

Created by Microsoft Academic Search, Academic Map (Figure 4.) is an interactive interface that shows the location of various institutions around the world. The size and colour of these points show the amount of authors found at a particular institution. Mousing over points shows the name of the institution, number of authors and publications. Zooming in reveals more institutions in an area. Clicking a particular institution displays nodes of the authors clustered around the name of the institution (Figure 5.).

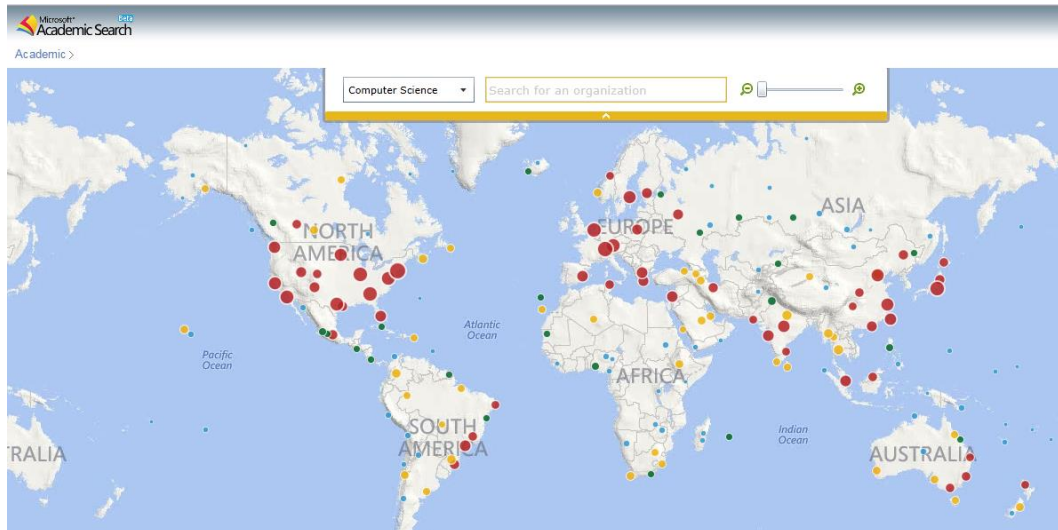


Figure 4 – Academic Map by Microsoft Academic Search (Microsoft Academic Research, 2013).

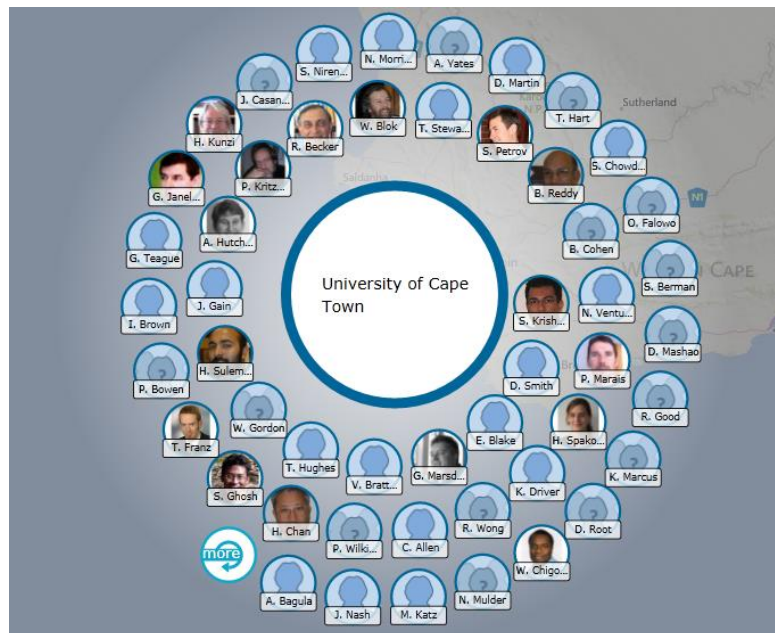


Figure 5 – Clicking on the University of Cape Town on Academic Map brings up a visualisation of authors (Microsoft Academic Research, 2013).

Red dots represent institutions with the most authors; yellow dots, less; green, even fewer and blue the least. Therefore, institutions with the most authors in each region can easily be identified. Difficulty arises in comparing institutions that are categorised as blue and green as circles appear similar in size even though a green institute may have three times more the amount of authors a blue institution may have.

CITATION MAP

Citation Map (Hu, et al., 2013) is an interactive map that shows the geographic distribution of citing authors of a chosen paper and indicates how recent these citations were. Data

shown on the map has been sourced from Microsoft Academic Search. The top ten (most) citations are indicated by different coloured flight paths from the source paper to the cited papers where a key displays the author name and number of citations by that author. Colour has been used to distinguish the source paper (blue) from citing papers (green). Different tints of green are used to indicate the recentness of paper differentiating earlier papers from more recent ones. That being said, it is sometimes difficult to distinguish the newness of a paper as the gradient scale for mid-range colours can be similar and therefore, are difficult to compare. Names of locations on the map are useful in helping to identify specific locations. Black text with white highlights providing sufficient contrast with the grey of the map to allow names to be easily read.

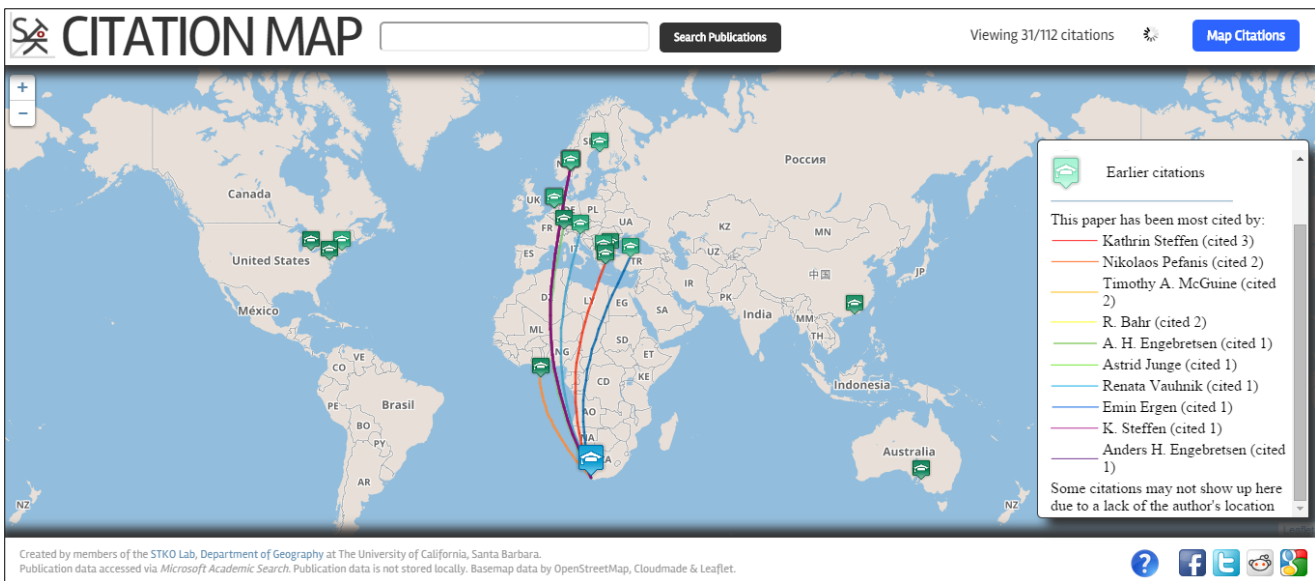


Figure 6 – Citation Map (Hu, et al., 2013).

In terms of interactivity, users can click, pan and zoom in/out of the map. Clicking a point on the map opens up a textbox over the point which shows details of the citing author including their photo, name, institution, paper title and a partial abstract of the paper.

Navigation of the map can be done by clicking and dragging, while scrolling allows the user to zoom in/out of the map to see a more precise location of an institution as occlusion can occur if the map is densely populated in a single region

A problem with the map is that if there are multiple papers from a single institution, map points are occluded preventing the user from clicking on papers located on the marker behind the first one.

After analysing various citation visualisations, we decided that an interactive implementation would be best suited for our visualisation as it would allow various dimensions of data (eg. citations, references, country, institution, discipline, time, etc.) to be shown without the visualisation getting too crowded.

From the aforementioned designs, we found Citation Map to be most similar to what we envisaged our implementation to be like. Thus we will be using some of the design principles in Citation Map (map interface, representation of earlier/later citations) as the basis of our design but add additional dimensions of data and improve on existing aspects of the Citation Map's design.

VISUALISATION

VISUAL QUERIES

In what **area** has the paper been cited **most**?

Where has the paper been **cited recently** (in the last five years)?

In what **discipline** has the paper been cited most frequently?

Which **disciplines** has this paper **referenced**?

INITIAL DESIGN

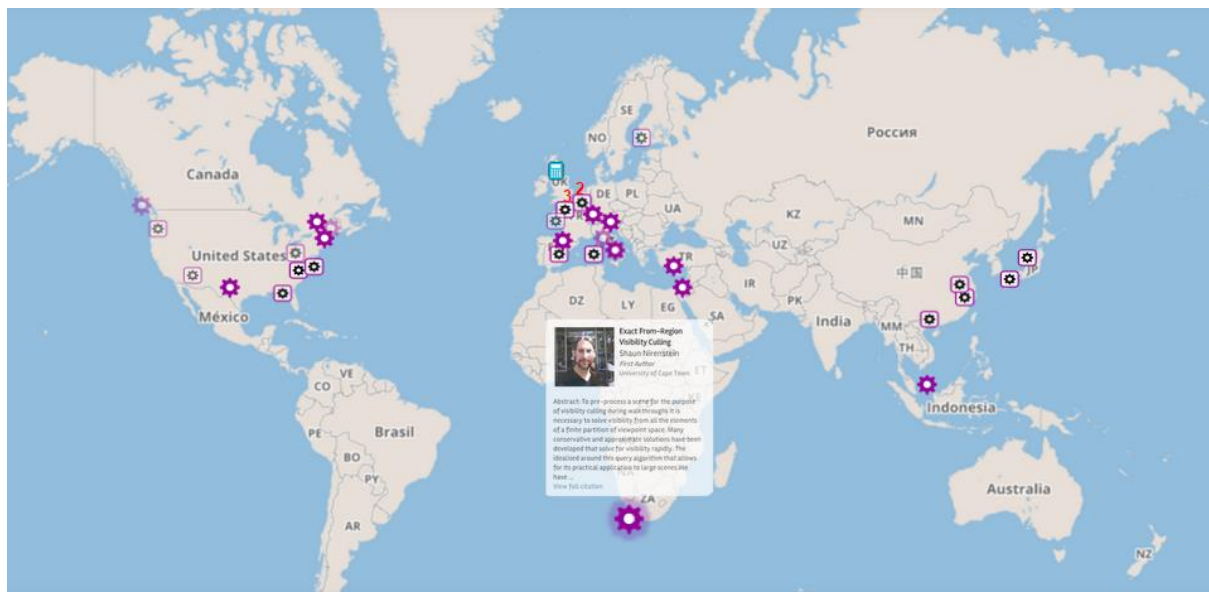


Figure 7 Initial Design

Our design is based on the [Citation Map](#) visualisation. We adapted it so that users could more easily identify:

- Disciplines
- Cited by vs References
- Old citations vs Recent citations
- Additional information on the paper
- Multiple papers in one location

DISCIPLINES DIFFER ON SYMBOLS AND COLOURS

















Discipline	Cited by	Reference
Business, Economics and Management		
Chemical and Material Sciences		
Engineering and Computer Science		
Health and Medical Sciences		
Humanities, Literature and Arts		
Life Sciences and Earth Sciences		
Physics and Mathematics		
Social Sciences		

Figure 8 Discipline Table

In order to ensure fast queries, the pins differ on two channels (both shape and colour). This makes it very easy to filter out the unnecessary data when performing a query. Our data is based on Google Metrics Categories. We chose a different symbol and colour for each category. Symbols and colours are linked to the "what" channel in the brain, so users should be able to easily identify which disciplines are which.

CITED BY VS REFERENCES DIFFER ON BORDER AND COLOUR



Figure 9 Cited by vs Reference

Papers which the source paper has been *cited by* are indicated by a black image with a white background and a border in the colour of the discipline.

Papers which the source paper *references* are indicated by the symbol in the colour of the discipline.

We chose to do this because we feel that it is more important to quickly identify *cited by* papers and having more colours will make *cited by* papers stand out more. In this way, users can distinguish *cited by* from *references* on the colour channel and still be able to identify the disciplines.

NEW CITES VS OLD CITES DIFFER ON SHADE OR TRANSPARENCY



Figure 10 Recent Citations versus Earlier Citations

Recent papers are those which have been written in the five years before the current year or the year in which the paper was written.

We had two choices/versions for papers which were written earlier. The first option was to make the symbol slightly transparent. The second option was to make the symbol a shade lighter. Our goal was to enable the user to tell newer papers apart from older papers, but still be able to identify the paper to be in the same discipline.

SHOWING EXTRA INFORMATION

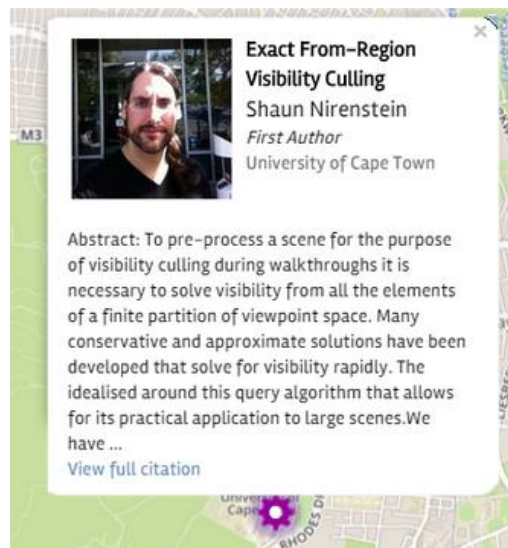


Figure 11 Showing extra information

When the user clicks on a symbol, they are presented with extra information on that paper. We based the popout box on [Citation Map](#)'s design, but edited it to make it easier to search.

The Title, Authors name and institution are shown at the top. A picture is also shown, if available (this is pulled from the Google Scholar profile). The picture isn't essential in terms of information, but it adds value for the user because they can put faces to names. An excerpt of the abstract is also included to aid the user in gaining a greater understanding of the paper and there is link to the paper if the user would like to research further.

MULTIPLE PAPERS IN ONE LOCATION



Figure 12 Showing multiple papers in on location

One of the challenges of the design was to show multiple papers in one location. We attempted to solve this by adding a number above locations with more than one paper indicating how many papers were written at that location. Additionally, when a user clicks on the university it shows all of the different citations from that university. These pop up around the given institution with the relevant information being displayed, as shown in the example. This is necessary to prevent occlusion from becoming an issue and ensuring that all papers can be represented.

FINAL DESIGN

In interactive Proof of Concept of our final design is available [here](#). A static image of the final design can be seen in Figure 15.

The following is a discussion of the implementation of feedback and redesign.

DISCIPLINES

It was suggested that we should categorise disciplines on the *Sub-category* level of Google Metrics because papers are often only cited from within their own category, but there could more likely be differences in sub-category. It was suggested that we keep the symbols to show in the "More Information" box, and use a recurring list of random geometric shapes to indicate to which sub-category a paper belongs.

CITED BY VS REFERENCES

In the final design, cited by is indicated by the symbol in the colour of the category and references are indicated by the same symbol, but white and with a border in the colour of the category (example shown in interactive map above). This is the opposite of what was decided for the initial design. We made this decision after trying both options on the map and concluding that the symbol in the colour of the category stood out more and should therefore be used to refer to 'cited by' papers.

RECENT CITES VS EARLIER CITES

The final design used different shades of colour instead of transparency to show older citations. This was an intentional design decision because it was discovered that transparent symbols tended to fade and blend into the background. Thus, two different shades of colours were used to distinguish earlier/recent references and citations from each other. The chosen shades are different enough to be easily distinguishable from each other while still allowing the eye to group them together as they are from the same colour range.

SHOWING EXTRA INFORMATION

It was pointed out that the Extra Information pop out occludes too much of the map. It was suggested that only very basic information is shown on the pop out and that extra information is included in a sidebar.

MULTIPLE PAPERS IN ONE LOCATION

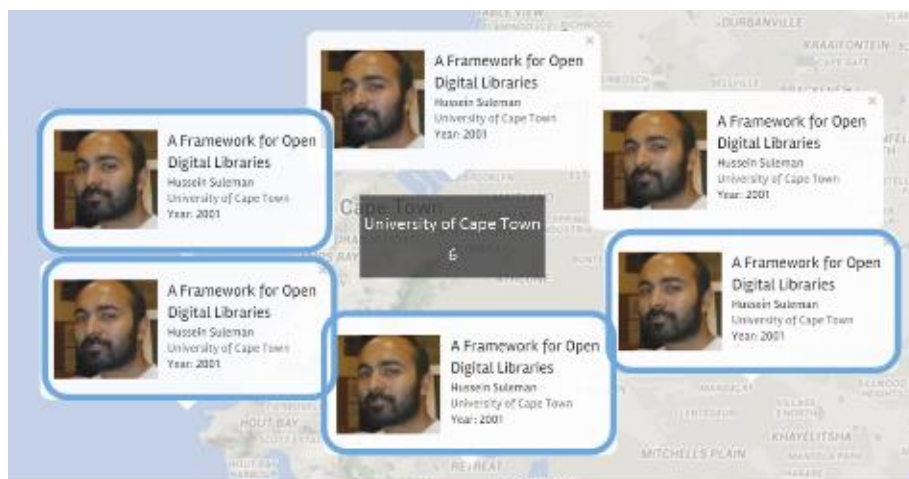


Figure 13 Final design of multiple papers in one location

While collecting the data, it was discovered that a single location could contain both papers which the source paper referenced and papers which the source paper was cited by. Therefore, we chose to indicate "cited by" papers with a blue border so as to distinguish "references" from "cited by" papers.

*Please note that Figure 13 is just an example of how multiple papers would be shown, and that it is not completely based on real data.

SIDEBAR

One of the major suggestions from the first presentation was to put in a sidebar for extra information, this was incorporated into our final design in order to provide necessary information which couldn't fit cleanly onto the map. The sidebar includes information such as the key, information on the selected paper, and aggregate information such as number of papers per discipline, per university, or per author.

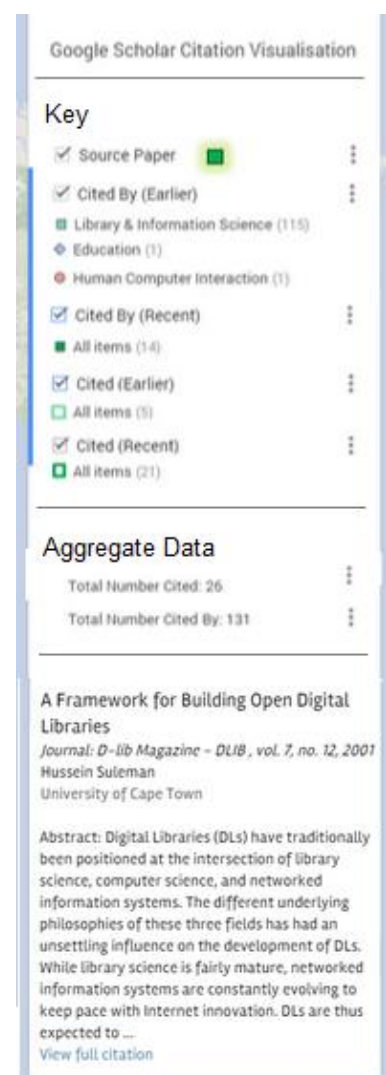


Figure 14 Sidebar



Figure 15 – A static image of the final visualisation.

DISCUSSION OF DESIGN STRENGTHS AND WEAKNESSES

Once the final design has been implemented, we evaluate the strengths and weaknesses and challenges of our visualisation based on whether we were

1. Able to accurately visualise the given data
2. Able to effectively convey the dimensions of the given data that were chosen (cited by, cited, earlier/recent, category).
3. Interactivity and usability of the Interface

STRENGTHS

The strengths of our visualisation fall into 4 main categories:

- Overview
- Zoom
- Filter
- Details on Demand

These categories are related to the interactivity offered by our visualisation and are strengths as they form a part of Ben Shneiderman’s Visual-Information Seeking Mantra which explains principles related to interface design which aid in addressing visual queries.

Other strengths are related to design decisions implemented based on visual thinking principles which help people make more effective visual queries.

OVERVIEW

The global map overview allows simple visual queries to be answered such as “Which region cited the paper the most/least?” In the particular example we implemented in our visualisation, it is clearly visible that most of the citations come from the same subcategory and that the article has less references than citations.

ZOOM

By allowing the user to zoom into a specific area of the map, users can see exactly where citations have come from (which institution). Additionally, this helps solve the problem of occlusion that may occur when looking at the overview of a region as points disperse to their exact location allowing individual markers to be identified.

FILTER

Using the Google Maps to implement our visualisation allows filtering to be done on earlier/recent citations/references. This allows the user to easily find the information they are looking for (eg. see only recent citations) by removing points on the map and aids in answering the visual queries a user may have.

DETAILS ON DEMAND

Detailed information about authors are shown when clicking a marker on the map. This allows users to only see information related to authors when needed allowing the map to remain clutter-free and more general visual queries (eg. comparison of cited papers vs cited by papers) to be made.

DESIGN DECISIONS

In order to ensure fast visual queries, map markers differ on two channels (shape and colour). This makes it very easy to filter out unnecessary information when performing a visual query.

The final design used different shades of colour instead of transparency to show older citations. This was an intentional design decision because it was discovered that transparent symbols tended to fade and blend into the background. Thus, two different shades of colours were used to distinguish earlier/recent references and citations from each other. The chosen shades are different enough to be easily distinguishable from each other while still allowing the eye to group them together as they are from the same colour range.

WEAKNESSES AND CHALLENGES

One of the major weaknesses is occlusion, both in busy areas (east coast of America for example) and at universities. We don't allow a user to quickly see how many times the paper was cited at Harvard for example. In order to do this they would need to click on the relevant (glowing) pin to see the data.

Another weakness has to do with how the data was collected. Due to the fact that Google Scholar doesn't have an API and automatically scraping results is against Google's Terms of Service, the data collection process was far more manual than the average computer scientist would deem acceptable. One of the issues with the manual collection of data in this way is that it is sometimes hard to pinpoint the location of a paper, especially for foreign language papers and unknown authors. As a result it is probable that some of the location data points to the current institution of an academic, instead of where the academic was at the time of writing.

CONCLUSIONS

The visualisation benefited a lot from the initial presentation, as many of the suggestions provided were incorporated into our final design. One of the suggestions which we struggled to incorporate was to show the numbers of citations at each university. The decision was made not to include this as it would simply add noise to the busy areas.

The final design was able to show the previously mentioned visual queries, it displays the necessary information in a clear and concise ways and allows users to interact with it in useful ways.

FUTURE WORK

The visualisation is still a prototype and functionality hasn't fully been implemented yet. The implementation and coding of more advanced features (eg. dynamic sidebar with author abstract information, linking author's to their Google scholar profile, showing multiple papers in a single institution as nodes clustered around a point) and pulling the data in automatically instead of having to scrape it manually would be a possible line of future work.

Automatic integration would rely on a change of Google's terms of service or the creation of an API. Sometime could also be spent exploring what summary data would be useful, possibly by continent, country or state/province.

This application would be interesting for both students and academics. Students could use it to get more in depth and contextualised information about their research and academics could see where and who is citing their work. The generalisation of this visualisation could have broad applicability, a fairly uncreative example of this could be showing participation at international conferences.

IDEAS FROM SECOND PRESENTATION

Future work could include being able to compare two papers at the same time by having side-by-side visualisations. This would be useful for the user to see how closely one paper relates to another or to see how much one paper may have been influenced by the paper it is citing.

The pictures of the authors help the user to put a face to the name. The value of this could be expanded - this helps in easier recognition of recurring authors.

TEAM CONTRIBUTIONS

The group worked collaboratively on Google Drive. Everyone in the group helped with finishing touches in completing the report. With regards to the design of the visualisation, the group worked together in critiquing each other's ideas and deciding on a final design.

Areas where individual members made a significant contribution are listed below:

CHANTAL YANG

- Introduction/overview
- Research section
- Interactive Google Map PoC Visualisation

ION TODD

- Data mining
- Discussion section

LAUREN SANBY

- Editing HTML code/webpage layout
- Visualisation section
- About section

REFERENCES

Autodesk Research, n.d. *Citeology - Projects - AutoDesk Research*. [Online]
Available at: <http://www.autodeskresearch.com/projects/citeology>
[Accessed 14 March 2015].

Hu, Y., Mckenzie, G. & Gao, S., 2013. *Citation Map: Visualizing The Spread Of Scientific Ideas Through Space And Time*. [Online]
Available at: <http://stko.geog.ucsb.edu/node/6>
[Accessed 14 March 2015].

Microsoft Academic Research, 2013. *Help Center - Microsoft Academic Research*. [Online]
Available at: <http://academic.research.microsoft.com/About/Help.htm#3>
[Accessed 2014 March 2015].

Shneiderman, B., "The eyes have it: a task by data type taxonomy for information visualizations," *Visual Languages, 1996. Proceedings., IEEE Symposium on* , vol., no., pp.336,343, 3-6 Sep 1996

Sylvia IV, J., 2015. *Visualizing My Interdisciplinary Field (Part 2)*. [Online]
Available at: <http://www.hastac.org/blogs/jsylvia/2015/01/19/visualizing-my-interdisciplinary-field-part-2>
[Accessed 15 March 2015].