Visualization of a Occupational Therapy Research Database

Peter Annable, Roshamiliza Rahman, Shahzad Saleem, Mahesh Suravajjala, Yelena Yezerets

Fig. 1. - Research Keywords Word Cloud

Abstract - The American Occupational Therapy Foundation (AOTF), in partnership with the American Occupational Therapy Association, has started to maintain a database of research in the field. We demonstrate how data visualizations can provide insights for building scientific networks, identifying scientific leaders for specific initiatives, and summarizing capacity to external stakeholder groups. We find that seven institutions have received 80% of total funding dollars. While research connected to Stroke is the largest overall, research related to Autism Spectrum Disorder has grown from none in early years to being the top diagnosis area. Overall there are 29 different research areas funded between 1982 and 2016 with an approximate funding of $153 million. The three most researched areas are Autism, Stroke, and Sensory integration areas. Insights are limited to current data available and do not reflect all research in occupational therapy. We demonstrated the need to expand upon the data collection efforts and use data based techniques for guiding future research.

Index Terms —Occupational Therapy, Information Visualization, Topical Network, Fruchterman-Reingold, Burst Analysis

Introduction

For several years, the American Occupational Therapy Foundation and the American Occupational Therapy Association have built up a researcher database to track research on Occupational Therapy techniques and practices. **426 unique research projects** have been recorded, spanning research **from 1982 through 2016.** The intent of this data analysis is to create a comprehensive understanding of the scientific community for building scientific networks, identifying scientific leaders for specific initiatives, and summarizing capacity to external stakeholder groups.

We have analyzed the data and created five visualization types to accomplish the following tasks:

1. Depict an evolving topic space using text associated with projects and researchers
2. Present insights on where the research is done and possible research linkages between locations.
3. Show the evolution of topics over time and where bursts of activity have occurred.
4. Identify networks of scientific experts and their expertise areas based on diagnosis and agenda categories presented in five cumulative time slices.
5. Dynamically explore the research data by location and topic areas.

1 Method

1.1 Data

Dr. Julie Bass, Director of Research, AOTF, provided a dataset consisting of a single fact table with 437 research project records and dimension tables for the Researcher Profile, Diagnosis Description, Ages of study participants, Practice Settings, and ICF Relation. Key attributes of the research projects included Project Title, Researcher Id, Institution, Funding Start and End Dates, and Funding Amounts.

During data analysis we uncovered several problems that were confirmed by Dr. Bass:

- Duplicate records were removed, reducing the total number of projects from 438 to 426.
- Incorrect names used in the Researcher Profile, invalidating the linkage to Research Projects. As a result we used the Principle Investigator field to connect projects with researchers.

- *Peter Annable*
  Cincinnati, OH . E-mail: pannable@iu.edu
- *Roshamiliza Rahman*
  Denver, CO. E-mail: rosrahma@iu.edu
- *Shahzad Saleem*
  Minneapolis, MI E-mail: saleems@iu.edu
- *Mahesh Suravajjala*
  Salt Lake City, UT E-mail: msuravaj@iu.edu
- *Yelena Yezerets*
  Columbus, IN E-mail: yyezeret@iu.edu

All Visualizations available in high resolution at github.iu.edu [4]

- Also due to lacking Researcher Profiles, 56 research projects could not be associated with an Institution. These data were excluded from Geospatial mapping.
- Institution locations were not specified. We used the primary location address via Google search to determine location for Geospatial analysis.
- Keyword misspellings, which were manually corrected.

To ensure proper data integrity, we imported the dataset into MS Access and extracted lookup tables based on Project ID for Diagnosis, ICF, Agenda, and Age categories. Initial statistical analysis was performed in MS Access by creating crosstab queries for PI and Diagnosis, Diagnosis and Agenda, and Diagnosis and ICF categories. Total funding for each project was calculated by summing up NIH, Federal and Non-Federal Funding amounts. Total cost was calculated by adding up Fed Direct Cost, Non Fed Direct Cost, and NIH Direct Cost. A final de-normalized data extract file was created from MS Access, with multi-valued attributes concatenated into single columns for Age Brackets, ICF Description, Agenda, Diagnosis, and Keywords.

The initial data source file, the final MS Access input tool, and resulting visualizations are provided through - github.iu.edu[4].

Table 1. Occupational Therapy dataset statistics

Research Projects	426
Research Profiles	55
Practice Settings	12
ICF Areas	7
Diagnosis Areas	29
Agenda Categories	8
Age Brackets	10
Funding years	1982 -2016

1.2 Temporal Visualization

For Temporal analysis, Sci2[1] was used to normalize, tokenize and remove stopwords from the diagnosis description field. The main objective of the burst analysis was to highlight important periods of time for research for different diagnoses. Our analysis was restricted by using a Gamma of 0.55 and a density scaling of 2.0 in the Sci2 burst detection algorithm. To gain a more complete picture, diagnosis keywords highlighted in burst analysis were used as an input to Tableau dashboard showing research funding by year for each diagnosis. These visuals are combined to show how various diagnoses were researched over the entire 1982-2016 period.

1.3 Topical Visualization

For topical analysis, Sci2 was used to normalize to lowercase, tokenize and remove stopwords from the combined keyword field. A total of 470 keywords resulted for analysis. To understand topic importance, a word cloud was created using the Term Frequency – Inversed Document Frequency (TF-IDF) method. In order to show the number of times each keyword was cited in conjunction with another keyword, Sci2 was used to extract a Word Co-Occurrence Network. The network size was reduced by selecting nodes with at least 6 references and edges with the value greater than 3. Isolates were also removed.

1.4 Geospatial Visualization

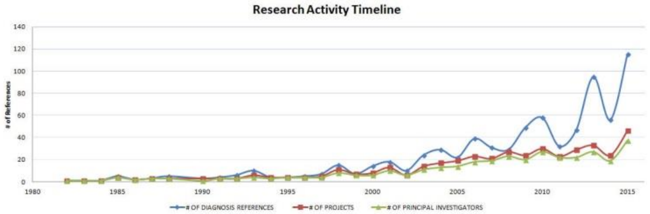
Institutions were mapped based on longitude and latitude markers generated using the Sci2 Generic Geocoder based on zip codes. Two Proportional Symbol Maps were generated with time periods of 1982-1999 and 2000-2016. Nodes size was used to represent the total funding amount, interior color as the number of projects, and the exterior ring colors for the number of principal investigators at each Institution.

1.5 Bipartite Networks

To elicit insights into existing scientific networks and help identify scientific leaders for specific expertise areas, we completed a network analysis and identified networks of experts working on various diagnosis relevant to particular research project. We used Sci2 to extract bipartite networks of researchers united, first, by same diagnosis and, second, by agenda categories that were cumulatively sliced by seven years according to the project funding date for the period of 1982 to 2016. To generate initial network attributes we created a property file to calculate a number of projects associated with individual researchers and diagnosis as follows:
node.countProjectsperResearcher = PROJECT_ID.count edge.count
ProjectsPerDiagnosis=USR_DIAGNOSIS_AREA.count

Preliminary analysis of this network showed that the number of nodes (references between primary investigators and respective diagnoses) increased proportionally during the seven-year periods (Figure 2).

Fig. 2. Occupational Therapy Research Activity Timeline.



Each network time slice was further analyzed to depict the essential linkages between individual nodes using the Blondel Community Detection analysis. Each time slice was imported into Gephi [2] in order to visualize it with the Fruchterman-Reingold layout [5] algorithm. Each modular community was colored by essential diagnoses and nodes were sized by the number of diagnosis references. The networks were further imported into Inkscape [3] to create legends and finalize the label layouts.

To connect diagnosis descriptions with Agenda we tested out several different visualization types including Fruchterman-Reingold with Annotation and Circular Hierarchy; however the Bipartite Network Graph proved to be the most precise and straightforward. The same type of time slicing on the Agenda-Diagnosis data helped to compare them with PI-Diagnosis time slices and to provide more insights into the nature of Occupational Therapy research. Even though these networks were not included in this article, they are accessible through our GitHub repository [4] and would be valuable assets for further discoveries in the Occupational Therapy research community.

1.6 Data Exploration using Tableau Public

To provide further understanding of research data, a demonstration tool was built using Tableau Public. By using visualizations that update in real-time as data selections are made, researchers can gain additional understanding of the data through a “study explorer.” This provides a detailed study listing filtered by one or more selections of Research Institution, Age Bracket, Agenda, Diagnosis, or ICF Description.

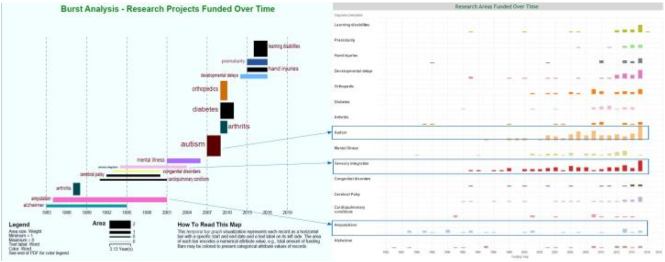
To prepare the dataset for Tableau, an MS Excel file was created with multiple tabs. The first tab used the same data set prepared for the Temporal Visualization, which had multi-valued attributes merged into single columns. OpenRefine [7] was then used to add proper casing to all values for better readability. Additional data tabs were added to the Excel file, corresponding to each of the dimension tables in the Access Database for Agenda, Age Bracket, ICF Description, Diagnosis, and Agenda. At the client’s request, data regarding funding amounts was removed. This data was loaded into Tableau and a dashboard of interlinked visualizations was prepared and published online. [8]

2 RESULTS

2.1 Temporal Visualization

Figure 3 highlights the research topics with important bursts and the persistency of the funding over time. Thirty-one different diagnoses have been funded over the past 30 years, and 14 topics were picked by the burst analysis algorithm and also used as inputs into the Tableau visualization. In total, the most funded areas are Other Diagnoses, Autism, Stroke, Sensory Integration, and Brain Injuries. During the 1980s and 1990s, Amputation-related research projects received the most funding. During the 1990s, Sensory processing disorder related projects were also equally encouraged by AOTF. During the 2000s, Autism related research projects received the most funding, followed by Heart and Stroke related projects. There was a total of 272 different projects during this decade, with $64.7 million of funding granted (almost twice the amount compared to the 1990s). New research areas funded the most during the 2000s were Orthopedics and Neurodegenerative diseases. After 2010, Autism and Heart related research projects have still been the most funded research areas, although the total breadth of diagnosis areas covered has widened. A total of $57.8 million was granted between 2010 and 2016. New research areas funded during this time have been Burns and HIV/AIDS.

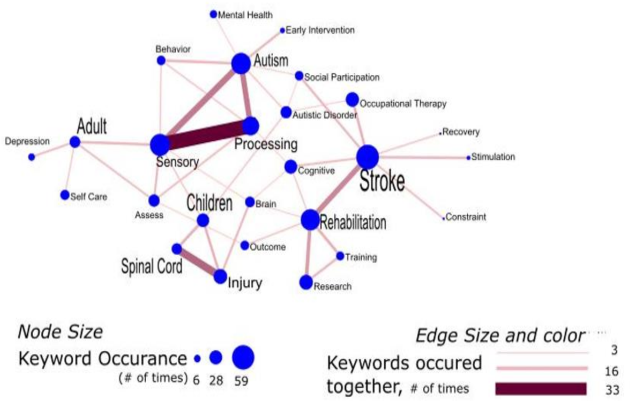
Figure 3. Burst Analysis of Topic Keywords



2.2 Topical Visualization

Using GUESS, a network was laid out with the Kambada Kawai algorithm [9], through which we can see nodes ranging from citation count of six to 59. Node size reflects citation count. Edge color and weight reflects relationship strength, ranging from three to 33 co-occurring keywords. The legend was added using Inkscape. The word cloud can be seen in Figure 1.

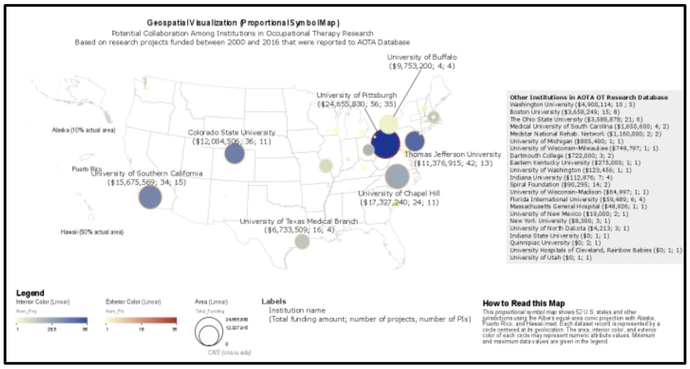
Fig. 4. Kamada Kawai Network Visualization of Research Topic Keywords



2.3 Geospatial Visualization

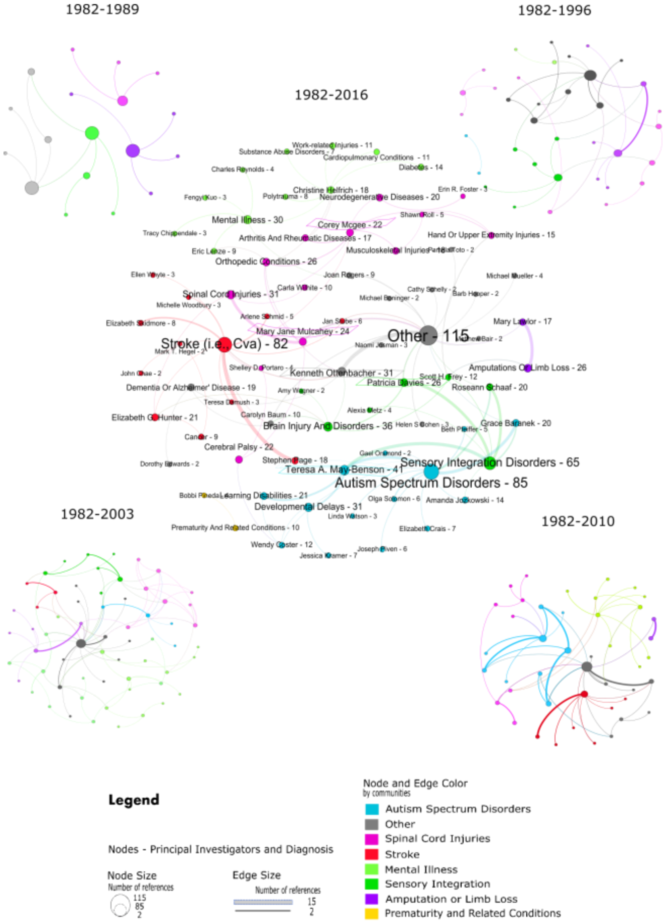
There were 56 records with missing institution names and one institution was located outside of the United States; thus their zip codes could not be identified. From the 380 research projects with geolocation, we identified 29 distinct institutions which were the focus of the geospatial maps.

Fig. 5: Proportional Symbol Map of institutions receiving grant funding in occupational therapy research between 2000 and 2016


Geospatial Visualization (Proportional Symbol Map)
Potential Collaboration Among Institutions in Occupational Therapy Research
Based on research projects funded between 2000 and 2016 that were reported to AOTA Database
University of Buffalo ($9,753,200; 4; 4)
University of Pittsburgh ($24,655,830; 56; 35)
Alaska (10% actual area)
Colorado State University ($12,084,506; 36; 11)
Thomas Jefferson University ($11,376,915; 42; 13)
Puerto Rico
University of Southern California ($15,675,569; 34; 15)
University of Chapel Hill ($17,327,240; 24; 11)
University of Texas Medical Branch ($6,733,509; 16; 4)
Hawaii (50% actual area)
Other Institutions in AOTA OT Research Database
Washington University ($4,900,114; 10 ; 5)
Boston University ($3,658,249; 15; 8)
The Ohio State University ($3,588,878; 21; 6)
Medical University of South Carolina ($1,650,600; 4; 2)
Medstar National Rehab. Network ($1,160,000; 2; 2)
University of Michigan ($805,400; 1; 1)
University of Wisconsin-Milwaukee ($749,797; 1; 1)
Dartmouth College ($722,000; 3; 2)
Eastern Kentucky University ($275,000; 1; 1)
University of Washington ($123,456; 1; 1)
Indiana University ($112,876; 7; 4)
Spiral Foundation ($90,295; 14; 2)
University of Wisconsin-Madison ($64,997; 1; 1)
Florida International University ($59,489; 6; 4)
Massachusetts General Hospital ($48,926; 1; 1)
University of New Mexico ($19,000; 2; 1)
New York University ($8,300; 3; 1)
University of North Dakota ($4,213; 3; 1)
Indiana State University ($0; 1; 1)
Quinnipiac University ($0; 2; 1)
University Hospitals of Cleveland, Rainbow Babies ($0; 1; 1)
University of Utah ($0; 1; 1)
Legend
Interior Color (Linear)
Num_Proj
1
28.5
56
Exterior Color (Linear)
Num_PIs
1
18
35
Area (Linear)
Total_Funding
24,655,830
12,327,915
0
CNS (cns.iu.edu)
Labels
Institution name
(Total funding amount; number of projects, number of PIs)
How to Read this Map
This proportional symbol map shows 52 U.S. states and other jurisdictions using the Albers equal-area conic projection with Alaska, Puerto Rico, and Hawaii inset. Each dataset record is represented by a circle centered at its geolocation. The area, interior color, and exterior color of each circle may represent numeric attribute values. Minimum and maximum data values are given in the legend.

2.4 Bipartite Networks

Fig. 6. Network map of 51 Principal Investigators working on 29 diagnoses for a period of 1982 through 2016.

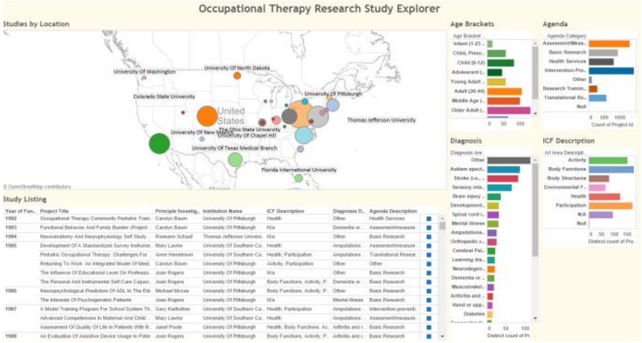


The dataset was cumulatively sliced by seven years. The nodes were assigned to eight communities of diagnoses based on the Blondel Community Detection algorithm. The nodes represented primary investigators and diagnoses. Nodes, respective labels and linkages between the nodes were sized by the number of references to diagnoses. The maximum number of references (115) was attributed to the 'Other' diagnosis. The maximum number of connections (15) between nodes was associated with the 'Stroke' diagnosis and Principal Investigator Stephen Page. The top five researchers with the most diagnoses referenced on the final cumulative time slice were outlined with the colors of their respective communities as follows: Theresa May-Benson – Autism Spectrum Disorders, Kenneth Ottenbacher – Other diagnoses, Patricia Davies - Sensory integration/processing disorders, and Mary Jane Mulcahey and Corey McGee - Spinal cord injuries.

2.5 Data Exploration using Tableau Public

Figure 7 shows the main screen in the Tableau tool [8]. Additional visuals were added for value-added analysis, such as studies by Institution and Diagnosis.

Fig. 7. Study Explorer Screen from Tableau Public



Study Listing

Year of Fun..	Project Title	Principle Investig..	Institution Name	ICF Description	Diagnosis D..	Agenda Description
1982	Occupational Therapy Community Pediatric Train..	Carolyn Baum	University Of Pittsburgh	Health	Other	Health Services
1983	Functional Behavior And Family Burden (Project ..	Carolyn Baum	University Of Pittsburgh	N/a	Dementia or ..	Assessment/measure..
1984	Neuroanatomy And Neurophysiology Self Study ..	Roseann Schaaf	Thomas Jefferson Univers..	N/a	Other	Basic Research
1985	Development Of A Standardized Survey Instrume..	Mary Lawlor	University Of Southern Ca..	Health	Amputations ..	Assessment/measure..
	Pediatric Occupational Therapy: Challenges For ..	Anne Henderson	University Of Southern Ca..	Health, Participation	Amputations ..	Translational Resear..
	Returning To Work: An Integrated Model Of Medi..	Carolyn Baum	University Of Pittsburgh	Activity, Participation	Other	Other
	The Influence Of Educational Level On Professio..	Joan Rogers	University Of Pittsburgh	N/a	Other	Basic Research
	The Personal And Instrumental Self-Care Capaci..	Joan Rogers	University Of Pittsburgh	Body Functions, Activity, P..	Dementia or ..	Basic Research
1986	Neuropsychological Predictors Of ADL In The Eld..	Michael Mccue	University Of Pittsburgh	Body Functions, Activity, P..	Other	Basic Research
	The Interests Of Psychogeriatric Patients	Joan Rogers	University Of Pittsburgh	N/a	Mental illness	Basic Research
1987	A Model Training Program For School System Th..	Gary Kielhofner	University Of Southern Ca..	Health, Participation	Amputations ..	Intervention-preventi..
	Advanced Competencies In Maternal And Child ..	Mary Lawlor	University Of Southern Ca..	Health	Amputations ..	Assessment/measure..
	Assessment Of Quality Of Life In Patients With R..	Janet Poole	University Of Pittsburgh	Health, Body Functions, Ac..	Arthritis and r..	Basic Research
1988	An Evaluation Of Assistive Device Usage In Patie..	Joan Rogers	University Of Pittsburgh	Body Functions, Activity, P..	Arthritis and r..	Basic Research

3 Discussion

3.1 Temporal Analysis

Despite the increasing number of research areas receiving funding only a few, such as Autism and Stroke, were consistently funded. Most of the research projects covered multiple diagnosis areas. The results of temporal analysis could be overlaid with disease history information to understand how research topics are affected by current disease prevalence. Comparing Occupational Therapy research areas and funding amounts to other research areas such as Physical or Speech Therapy could help understand how funding is allocated between related areas. Such analysis would also highlight if the research areas were complementary or supplementary in nature.

3.2 Topical Analysis

The Word Cloud (Figure 1) and the Word Co-Occurrence Network (Figure 4) suggested that major topics across all research data were Stroke, Rehabilitation, Autism, Sensory Processing, and Spinal Cord Injury. Common topics related to Children were Spinal Cord Injury, Autism, and Autistic Disorder. On the other hand, topics related to Adults were Sensory Processing, Self-Care, and Depression. The word cloud of keywords seemed like a good proxy to understand the topic space. However, client feedback indicated that it didn't properly show the breadth of research and likely indicated some deficiencies in data collection. The network diagram was found to be more useful as it provided more context.

3.3 Geospatial Analysis

Based on available data, prior to year 2000, there were only nine institutions conducting occupational therapy research. There were 31 principal investigators conducting 33 research projects with a total of $31 million in research funding.

In the more recent funding period since 2000, total research has increased dramatically, with seven major institutions - **University of Pittsburgh, University of North Carolina at Chapel Hill, University of Southern California, Colorado State University, Thomas Jefferson University, University of Buffalo, and University of Texas Medical Branch** - receiving about 80% of total funding, as shown in Figure 5. In total, there were 29 institutions involved in research, 141 principal investigators conducting 313 research projects, and more than $115 million granted in funding.

3.4 Bipartite Networks

According to the Bipartite Primary Investigator - Diagnosis network, the number of nodes tripled for every time slice. Starting with only eight Diagnoses, the field significantly expanded to 29 Diagnoses by

2016. From 1982 to 1996 the research was concentrated on Amputations or limb loss, Other diagnoses, and Dementia or Alzheimer's disease diagnoses with M. Lowler and J. Rogers consistently leading the research in their respective areas. After 1996 the field of research expanded significantly and started moving to the areas of Autism spectrum disorder, led by Teresa A. May-Benson, which is strongly connected to the Sensory integration/processing disorders research community that was headed by Patricia Davies. The third largest, the Stroke research community, was connected to the Other diagnoses by Kenneth Ottenbacher and was engaged in collaborations with Brain injury and disorders and Cerebral Palsy. Two more dispersed communities were identified. The first consisted of the area of diagnosis associated with various body injuries such as Spinal Cord, Upper Extremities and Musculoskeletal injuries with Corey McGee and Mary Jane Mulcahey leading the field. The second emerging community consisted of the Mental illness, Diabetes and Cardiopulmonary conditions field led by Christine Helfrich.

3.5 Data Exploration using Tableau Public

While this was not envisioned in the original set of deliverables, we received positive feedback from the client about the value of this tool for understanding research data on an ongoing basis. The most useful visuals for the client were the Study Explorer Dashboard, the Study Diagnosis by Institution Stacked Bar Chart, and the Study Agenda Bar Chart. In the future, this tool could be connected directly to the database to provide a living visualization and data exploration tool.

4 Conclusion

As a method for understanding Occupational Therapy research and guiding future allocation of resources, we found that each of the visualizations provided valuable perspective. Temporal and Topical analysis showed what research has been the most important over time, and when important periods of research occurred. Geospatial analysis showed how a small number of Institutions dominated total research. Network analysis showed who the important researchers were and connections between them. Unfortunately, we cannot make many absolute conclusions due to the incompleteness of the data. It was revealed to us during client feedback that data collection had been voluntary and not deployed consistently to all researchers.

Given our findings on the value of the visualizations in this report, we have the following recommendations to improve the data for future efforts:

1. Increase efforts to gather more research project data, and set a goal to collect at least 80% of research since 2000.
2. Review the data collection process to ensure the input is guided to avoid mistakes.
3. Review and correct Researcher Profile connections.
4. Investigate using data overlays such as public health records to look for correlations.
5. Include a master data table to capture Institution addresses.

Acknowledgements

The authors wish to thank Julie Bass, Director of Research, AOTF, Professor Katy Börner, IVMOOC teaching assistants, and the participants of Indiana University’s Information Visualization MOOC for providing us valuable feedback during the process of the project work.

References

[1] Sci2 Team. Science of Science (Sci2) Tool. (2009). Indiana University and SciTech Strategies, https://sci2.cns.iu.edu.

[2] Bastian M., Heymann S., Jacomy MGephi: an open source software for exploring and manipulating networks. . (2009). International AAAI Conference on Weblogs and Social Media.

[3] Inkscape is a professional vector graphics editor for Windows, Mac OS X and Linux. It's free and open source. https://inkscape.org/en/learn/tutorials/ [accessed 3/3/2016]

[4] Visualization of a Research Database in Occupational Therapy Foundation and American Occupational Therapy Association; 2016; https://github.iu.edu/yyezeret/Viz_OTResearch/wiki. (Github.iu.edu account is required).

[5] Fruchterman, T. M. J., & Reingold, E. M. Graph Drawing by Force-Directed Placement. (1991). Software: Practice and Experience, 21(11).

[6] GUESS Visualization Tool is maintained by Eytan Adar. http://graphexploration.cond.org/

[7] OpenRefine is data cleansing tool maintained by the team at http://openrefine.org/community.html

[8] Tableau Study Explorer (2016) https://public.tableau.com/views/AOTAResearchDatabase/StudyExplorer?:embed=y&:display_count=yes&:showTabs=y

[9] Kamada,T., and Kawai, S.. An algorithm for drawing general undirected graphs. Inform. Process. (1989). Lett., 31:7–15.