STEM: Individual, Local, and Global Flows and Activity Patterns

Prof. Katy Börner, Victor H. Yngve Professor of Information Science and Director of the Information Visualization Lab & Cyberinfrastructure for Network Science Center, School of Library and Information Science Indiana University

Oct 21, 2009 presentations slides are at http://ivl.slis.indiana.edu/km/pres/2009-borner-stem-activity.pdf

PROF. KATY BORNER: I applaud you for arranging this workshop. It is very important to consider not only funding inputs and publication outputs when evaluating science but to also look at data relevant for education and the training of the next generation of scientists.

Slide 2:

In this talk, I would like to make three major arguments. Firstly, just like science and the economy, STEM is global. Hence, we need global data covering all areas of science as people are often trained in one area of the world in one specific discipline of science but later they might work in a completely different area.

The second argument I'd like to make is that STEM is evolving dynamically. It is not static. Therefore, we need to study it using dynamically evolving data streams. Much data that is used to inform STEM decision making today is based on data that is one or more years old. Imagine you would try to regulate the temperature in a room based on one or more year old data.

The third piece I'd leave you with is the importance of having open data and open code, not to use "black box" tools on proprietary datasets as it is oftentimes done today. Reproducibility of results is a hallmark of science and we now have the data, algorithms and computational power to study STEM and inform STEM relevant decision making in a scientific way.

Next, I'd like to show you a number of examples which might inspire you to do similar devices, setups, approaches for the study of STEM education data.

Slide 3:

The setup you see here is part of the Mapping Science exhibit. Sample maps from the exhibit are hung in the back of this room. This particular setup is called "Illuminated Diagram Display." As you can see, there are two large-scale, printed maps. One shows the map of the world and the other a map of all of science. Using a touch panel display, one can click on any place on the world map to see in the map of science what kind of research is being conducted at this geolocation. Analogously, one can select any node in the Map of Science, to highlight on the map of the world who's conducting this kind of research on, let's say, chemistry.

The displays combine 300dpi large scale printed maps and low resolution projectors to illuminate different areas on the maps. Alternatively, the maps can be printed on semi-transparent material and back-illuminated using, e.g., Best Buy TVs.

Slide 4:
This is the interface to the set-up, so you see a map of science and map of the world. There are also buttons for more interdisciplinary areas of science and for people. For example, nanotechnology would need many different fingers to actually pin down all those areas of science which are involved in nanotechnology.   The same applies for sustainability research.
Currently the display renders publication data in a visual form. Wouldn't it be nice to also know who is training the students who are going to work in these areas?

On the right-hand side are the so-called people buttons. Here you could click on Albert Einstein and you'd see where he published his work, located on the map of the world and also on the map of sciences. In the second time slice, you'd see all the works which cite Einstein. And in he third time slice you'd see who is citing the cited works. The combined three time slide animation shows the spreading of his ideas spatially and within scientific space. The display could be used to render the diffusion of educational material, students, post-docs, etc. given appropriate data.

Slide 5:
The interface can be provided in different languages. This is the interface for a setup at the Chinese Academy of Science. DNA is the only word I can read.

Slide 6:
Let's have a closer look at the Map of the World.

Slide 7:
Here you see North America. Every white dot shown is a paper published in the Thompson Reuters database. Some of the areas are circled by red ellipsoids to indicate just how much is actually published there.

Slide 8 & 9:
This is Europe and Asia.

Slide 10:
Similarly, you can also zoom into the Map of Science. Each major area (node) of science represents multiple journals and each has a flowing label of top keywords.

Slide 11:
Here is a zoom into the inorganic chemistry peninsula. This so called base map of science can be (and has been) used to overlay scientific career trajectories, the publication record of a university or country, or to compare funding profiles of different agencies.

Slide 12:

Currently, we are working on setups where you have not only two maps but multiple maps. So you could see a global map of science, together with a zoom into medicine, and further into cancer; a map of the world, and a zoom into Asia and further into Japan. A Wii like interface will be used to let people enter names of scientists or areas of research and matching records will highlight in the Illuminated Diagram Display making a large amount of data accessible for a general audience.

Slide 13:
We also design science maps for children. These maps show the raw data but we added water color paintings to major areas of science.

Slide 14 & 15:
Students, children or caretakers, are asked to overlay different inventors and inventions on the Map of Science and the map of the world.

Slide 16 & 17:
Plus, there is poster of the world map and all puzzle pieces that children can take home.

Slide 18:
Here, you see the physical setup of the two puzzle maps in display at a public library in Bloomington, IN.

Slide 19:
We also took maps into classrooms. In the first session we introduced Google maps and asked children about famous explorers and how they travel over physical space and make discoveries. Christopher Columbus comes to mind.
In the second session, we showed them Maps of Science and discussed how researchers also go from one area of science to another and make discoveries along the way. We asked the children where, e.g., Albert Einstein would go. It wasn't clear because he made contributions to many different areas of science. The idea was then to either put him in all three areas or put him in the middle of those areas, which might be "nowhere land" (in the Science Map) or to cut him apart and put the pieces in different areas. Perhaps the answer is not so important, but it is important for children to think about these issues.
Finally, I asked children to find their home in science. There was one seven-year-old girl, who said: "I want to be a nuclear physicist," and she put her star-shaped sticker in the corresponding place on the map of science.
I think it's important that children don't see science as an obstruction, alien, and abstract, but that they find their place in science and that they discover how the many different subjects they learn all day long are connected with each other.

Slide 20:
Eight out of 50 maps from the Mapping Science exhibit are on display at the back of this room. This is a 10 iterations in 10-year effort—10 new maps are added every year.
The first iteration was devoted to communicating the power of maps to a general audience. The second iteration discussed the importance of reference systems. Many scientists have created reference systems to locate and access data. Astronomers for

instance can point to any segment of the sky and retrieve all data, all simulation results, all imagery, all measurements ever done there. Couldn't we have a similar system of reference for science? The fourth iteration introduced the power of forecasts, i.e., how we can learn from epidemic models, from economic models, oil depletion models, etc. for the forecasting of science itself.

Then there are six iterations which address the needs of specific user groups such as economic decision-makers, science policy makers, scholars (in 2010), maps as visual interfaces to digital libraries, science maps for children, and also science forecasts for the general public.

The 10th iteration will be on the topic of how to tell lies with maps. Maps have always been used for the interests of those who had the money to pay for them.

Slide 21:

If you would like to see the maps in their native size and archival quality—they are on display at Stanford University. It is our hope to bring the exhibit to Washington DC and your suggestions for possible venues are most welcome.

Slide 22:

My second point today is that: science and technology but also STEM are evolving dynamically. I don't think we need to do Meltdown modeling yet. However, we can learn from economic models and in particular this latest article by Mark Buchanan.

Slide 23:

It does not suffice to just count. We need a way to monitor, analyze and visualize STEM in real time so that anybody can see research results, policy decisions, teaching materials, job advertisements, together with bursts of activity and evolving communities of research and teaching practice, positive and negative feedback cycles. We need validated techniques, which are documented in a way that they can be understood and replicated across disciplinary boundaries.

Slide 24:

I would like to inspire this by showing you maps developed for other datasets and different questions. However, these maps might inspire you to do similar analyses using STEM data.

This is a map of all funding awarded by the National Institutes of Health (NIH) in 2007. A simple Google map browser is used to interactively navigate the topic space of all awarded extramural NIH projects. Each project is represented by a dot that is color coded by the respective NIH institute. You might interactively explore it via the link given below the map. Specifically, the online version let's you zoom and pan, select projects to see what institutes fund them and to access information on their titles, PIs, and abstracts.

Slide 25:

This map shows a map of science generated based on download logs by Johan Bollen and his colleagues. Whenever a user accesses paper A and then paper B, papers A and B are assumed to be related. Given more than 1 billion download counts a map can be created that reflect the access to (not citations) of scientific papers. As you might realize many doctors actively read Medline papers, but they might never write or papers or cite other papers. More details on the used approach are given in the text that accompanies the map and the associated paper. I would like to encourage you to actually read the text.

Slide 26:
Another map I found interesting is an interactive map of financial funding. It has a lot of very easy to read yet insightful ways to analyze and visualize philanthropy data.

Slide 27:
This is a map entitled “Chemical R&D Powers the U.S. Innovation Engine” created by the Council for Chemical Research in 2009. It shows how $1 billion of federal funding supports basic research, the chemical industry spends another 5 billion to fund invention development and technology commercialization, and how the revenue generated by the chemical industry results in $40 billion in Taxes and 600,000 jobs ultimately providing the resources for the next round of federal funding of chemical research.
This map is in display here. Please do read the accompanying text for more information.

Slide 28-31:
Last but not least, I would like to demo a map of job market data that resulted from a 6 week class project as part of the Information Visualization class I teach at IU. Here, Angela Zoss and Michael Conover acquired information on available jobs from job market RSS feeds. The data is then overlaid on a map of the world and a map of science (similar to the Illuminated Diagram Display). Circles represent sets of similar jobs; clicking on a circle results in a listing of all relevant jobs. Both map types support zoom and pan, search for keywords, e.g., nanotechnology, and access of detailed job descriptions.
As for the UCSD map of science shown here, 7.2 million papers published over a five-year timeframe went into the making of this map by my colleagues Kevin Boyack and Dick Klavans.
Science news, curriculum material, or other STEM relevant data could be rendered in a similar way. One could even add salary data to see what STEM teaching positions offer what kind of salaries.
If two students can create such a service in a six-week class project, then government could do much more insightful yet easy to use interfaces to STEM relevant data.

Slide 32:
The third argument I would like to make, relates to open data and open code for studying individual, local, and global STEM flows and activity patterns. Data needs to be made available in digital, fielded format (not as scanned pdf files) to be useful for data analysis and visualization. Ideally, data is shared as a database dump or MS Excel file with a data dictionary right next to it.
Plus, open code is key to replicable studies that can be trusted. If you work with contractors that use patented, “back box” software and you never get to understand what algorithms or exact parameter values are used then the results cannot be replicated, compared, or interpreted in a scientific manner. Ideally, there would be open data and open code for different levels and types of analysis as listed here.

Slide 33:
The table lists examples of individual, local, and global level analyzes that use statistical analysis and profiling, temporal analysis, topical analysis, or network analysis. Temporal analyses answer "When" something happens. Spatial analyses answer "Where" questions. Topical analysis answers "What" questions. Network analyses answers "Who" is involved, who is teaching whom, or who had an impact on what questions.

Micro and individual level studies deal with up to 100 records. Here interactive online interfaces can be designed where people can interact and manipulate each single record in real time. A the local scale, e.g., the job visualization, all data can be shown overlaid n a static reference system or base map. At the global scale, in the million or 10 million records range, displaying all records via online interfaces is not feasible. Instead, data might be rendered using desktop tools or simply printed into files. Plus, most analysis algorithms don't scale beyond one million records.

Slide 34:
These are some of the maps which you might be familiar with. If you are interested to learn more about these studies, please contact me. As one can see, temporal analysis looks at bursts of activity or plots data over time, e.g., in 113 years of physics research. Spatial analyses commonly use geospatial maps. Topical analyses, might use science maps or topic network layouts to display topic spaces. Social networks are commonly shown as networks—sometimes as non-legible "spaghetti balls" but more and more commonly in a more legible manner using clustering and backbone identification techniques to give you more understanding about the main structure of the network, the groupings that exist.

Slide 35:
We have managed, thanks to NSF, NIH, and James S. McDonnell Foundation funding, to create a suite of different tools and databases which are useful for the study of science.
The Scholarly Database supports to cross-search 23 million scholarly records—Medline papers, U.S. patents, and NSF and NIH funding. You simply type in the name of the creator, title, years, and select a database to retrieve all relevant records, e.g., all papers, patent or grants that have "RNAi" in the title. Search results can be examined and downloaded in bulk in well-documented data formats.
The Network Workbench Tool has 160 different algorithms for the preprocessing, modeling, analysis, and visualization of networks.
The Science of Science Tool will be demo'd in the remainder of the talk. The Epidemiology Cyberinfrastructure supports the study of diffusion processes, e.g., the spreading of H1N1 but also the diffusion of ideas or the workforce.
Flyers with more information on these infrastructures can be found on the table over there.

Slide 36:

The Science of Science Tool was funded by NSF"s SciSIP program. The tool can be used to extract scholarly networks and to render them in meaningful ways, but also to do science map overlays or so-called horizontal bar graphs. I have examples of the latter. The bar graphs basically show you how much funding from the starting date of the funding to the ending date for a specific project which is titled here, and then the area is size coded by the amount awarded or the number of jobs created.

Slide 37:
The tool also supports very simple geomaps such as world map and U.S. map overlays. Plus, it can do circular hierarchy renderings. Again all this code is open and well documented. It is licensed using Apache 2.0 so that anybody can take it and create opportunities with it.

Slide 38:
These are some of the supported data formats. You can read in your own bibliography data if you would like to map your personal academic profile. You can also read in Scopus or Web of Science publication data or funding data by the National Science Foundation. Thanks to Jim Onken (NIH), who is in the audience today, we now also support reading in RePORTER data about NIH funding and associated Medline papers and Edison patents.

Slide 39-41:
The Sci2 Tool offers easy access to more than 100 algorithms. Fortunately you have printouts of those slides, so I'm going to rush over those because I want to leave you with an understanding that you can use this tool to make sense of your very own data.

Slide 42-44:
According to the "Federal K-12 STEM Education Program Funding in 2006" figure in the 2007 Report of the Academic Competitiveness Council by the Department of Education, STEM is heavily funded by the National Science Foundation.
To understand what projects are funded by NSF, I downloaded all currently active NSF awards that have "stem" and "education" in title and abstract from the NSF Awards Search site at http://www.nsf.gov/awardsearch. On Oct 18, 2009, the total number of awards was 1,340 with a total awarded amount to date of $1,347,802,833.

Slide 44:
The top 10 projects with the highest award amount are shown here.

Slide 45:
Interested to gain an overview of all funding durations and amounts, I used the Sci2 Tool to render the 1,340 awards as a so called "horizontal bar graph". Time runs from left to right and each project is represented by a horizontal bar labeled by its title on left. The beginning and ending of each bar corresponds to the project start and end dates respectively while its area size corresponds to the total award amount. Major projects are easily identifiable. Hardware funding (high award amount over a short duration) look like vertical bars. Color coding can be used to distinguish different award types.

Slide 46-55:
The Sci2 Tool can also be used to geolocate all awards, to see who is collaborating (Co-investigating) with whom, to understand what projects fund which PIs or what NSF Programs are Co-Funding STEM, etc.

Slide 55:
It was interesting to see that the "Hist Black" and "Tribal" programs funds institutions that are only funded through these respective programs and by no other program. According to this dataset, S-STEM:SCHLR SCI TECH actively funds 179 institutions.

Note that these analyses can be replicated by anybody.

Slide 56:
If you are interested to learn more, please consult the Science of Science Cyberinfrastructure Portal at http://sci.slis.indiana.edu.

This is the end of my talk. Thank you for your interest.
Your questions and comments are welcome.