Networks in digital libraries, a personal view

By **Thomas Krichel**

Abstract: *This is a personal introduction to the relationship between digital libraries and networks. I recall the way I came to the subject of networks. I describe the way that I have tried to harness networks for digital library building. And I point out some of the difficulties in networking digital objects and their descriptions in digital libraries.*

Keywords: *Digital libraries, Networks, Network effects, Graphs, Co-authorship, RePEc.*

Born in Völklingen, Saarland in 1965, I studied Economics and Social Sciences at the universities of Toulouse, Paris, Exeter and Leicester. Between February 1993 and April 2001 I lectured in the Department of Economics at the University of Surrey. In 1993 I founded NetEc, a consortium of Internet projects for academic economists. In 1997, I founded the RePEc dataset to document Economics. Since January 2001, I teach in the College of Information and Computer Science at Long Island University. Since July 2005, I also supervise a few students at the Faculty of Information Technology of Novosibirsk State University.

Título: Redes en bibliotecas digitales, una visión personal

Resumen: *Introducción personal a la relación entre bibliotecas digitales y redes. El autor recuerda la manera cómo llegó al tema de las redes y describe cómo ha tratado de aprovecharlas para la creación de bibliotecas digitales. Y señala algunas de las dificultades existentes para poner objetos digitales en red y sus descripciones en bibliotecas digitales.*

Palabras clave: *Bibliotecas digitales, Redes, Efectos de las redes, Grafos, Coautoría, RePEc.*

Krichel, Thomas. "Networks in digital libraries, a personal view". En: *El profesional de la información*, 2008, noviembre-diciembre, v. 17, n. 6, pp. 662-663.

DOI: 10.3145/epi.2008.nov.09

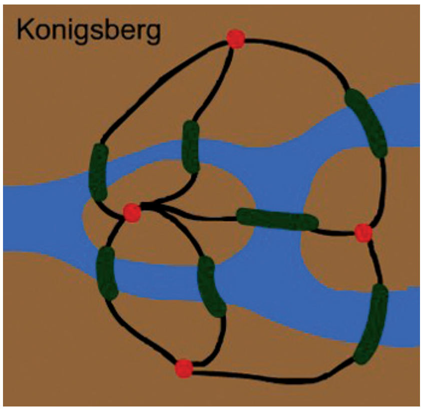
NETWORK RESEARCH has been around for some time now. I was introduced to this from my math books at school. There I saw a little schematic map of the town of Königsberg, now named Kaliningrad.

The town is located in East Prussia and currently forms an enclave of Russia. The town was built on the river Pregel. There were seven bridges going to and from a couple of islands. The task set was to devise a walk that would start and end at the same place but would visit every bridge just once. No solution was given. I vividly remember my frustration in trying to find the answer to this problem but there was no internet to aid my studies. However, 30 years later, with the emergence of the internet, I revisited this problem.

It turns out that there is no mathematical solution to this question. Ever since, I have been cursing the writer of this textbook for wasting my time. But there are some mitigating circumstances. First, the person who showed that there was no solution, in 1738, was **Leonard Euler**, not exactly a mathematician to sneeze at. Second, the problem is considered to be the birthplace of graph theory. You have pieces of territories connected through bridges. This is just like having nodes connected by edges.

Graphs for digital libraries

I became interested in the idea of graphs as a tool to produce value-added services in digital libraries. Surely a distinguishing feature of digital libraries, as opposed just a pile of files on a disk somewhere, is that each element of the digital library has been uniquely identified, and that there are some relationships between them. Thus, every document is a node. The relationship that it has to another node implies that there is an edge between nodes.



Los 7 puentes de Königsberg

The classic example comes if you think of your library as containing scientific papers, and the relationships as citations. One paper cites the other. This is an asymmetric relationship, because in most cases if paper "A" cites paper "B", then paper "B" does not cite paper "A". Citation graphs can be used to find related papers, since related papers will tend to cite the same paper. This is particularly powerful because it cuts across language barriers. For example, two papers, one in Spanish and another one in

English, can be found to be related even though they may not share the same terms at all. Free citation data is scarce at this time. But anybody who is interested can have a look at **José-Manuel Barrueco-Cruz**'s *CitEc* dataset at

http://citec.repec.org

which does have a good free citation dataset of the *RePEc* digital library.

Co-autorship graph

One feature that I have worked on is the structure of co-authorship in the *RePEc* digital library. The library has an author registration system. Authors contact that system to state which papers, as catalogued in the library, they have written. If two authors claim to have written the same paper then they are co-authors.

Co-authorship then sets up a graph between people. This is something that we can call a network. The graph is symmetric, because if I am somebody's co-author then he is my co-author. Nowadays co-authorship is very common. Thus, as more and more authors register, we can build a network of co-authorship across all authors in the discipline. Authors are linked through co-authorship links. These links can be short. I have a co-author "Joe", so the path between me and Joe is of length one. Joe has a co-author "Jane". I have neither had the pleasure nor the ability to write a paper with Jane, so my path to Jane is of length 2. All authors who can be reached by paths between them form a co-authorship community. Typically, the largest such community covers more than 50 percent of all authors, often 80 or 90 percent. For each author in the community, we can then observe how they are linked to other.

As the simplest measure, we can calculate the average length between the author and all others. If that average length is small, then we can say that the author is central to the community. If that average length is large, we can say that the author is marginal. This gives us a way to rank authors in the community.

Why do this? If an author wants to improve their ranking, then he can write more papers with more co-authors. This is not something that can be accomplished immediately as it takes time and effort to write papers. A cheaper and easier alternative is for the author to make sure that all of their co-authors are registered and that their list of papers is up to date. Thus the centrality ranking can be a tool to get others involved.

If we want to set up such a system in practice, we find that it is not trivial to do so once the number of authors reaches a few thousand. For each author we need to find the shortest number of paths between themselves and their remaining co-authors. So, if there are 10.000 authors we need to calculate 49.995.000 shortest paths. We cannot simply do this only once because new authors may register at any time and will appear in the community of connected authors, necessitating path recalculation. I have designed software to be able to accomplish this task, but I do not have a dedicated machine for running this service.

Digital libraries development

So what is holding back network usage in digital libraries, be it networks with documents or networks with people? Well first it is a fundamental matter. We really need a stable identification of items in the digital library. This is something that is still quite hard to achieve, because the evaluation of sameness cannot be left to a computer and it has to be lead by humans. We need an organized collection description, hopefully something that is freely available. These are issues that I have written about in other papers, most recently in "From open source to open libraries", available at

http://openlib.org/home/krichel/papers/kuyus.html

Second, as I have pointed out here, we need better computational resources. While computers are improving, it is still hard labor to go through the items in a digital library. Since *RePEc* now has about 600.000 items, we are talking about a huge amount of calculations that would be required to produce centrality calculations when full citation information for all is available.

Finally, we have the problem of visualization. Just a centrality table is not something that will get us a lot of attention. We really need to have pictures that communicate stories, show clusters of nodes that belong together and show peaks and troughs. We are still quite far from intensive usage on networks in digital libraries. A lot remains to be done.

***Thomas Krichel**, Palmer School of Library and Information Science, College of Information, and Computer Science, CW Post Campus, Long Island University 720 Northern Boulevard, Brookville 11548-1300, USA*

Faculty of Information Technology, Novosibirsk State University, 2, Pirogov Street, 630090 Novosibirsk, Russia

http://openlib.org/home/krichel
krichel@openlib.org