CASE STUDY: USING DISSERTATIONS DATA FOR RESEARCH

School of Informatics and Computing at Indiana University

ProQuest®

Cassidy Sugimoto, Assistant Professor in the School of Informatics and Computing at Indiana University, is one of three advisors that head up the MPACT Project. MPACT (www.ibiblio.org/mpact) is an academic genealogy project devoted to defining and assessing mentoring as a scholarly activity, through examining the emergence and interaction of disciplines and identifying patterns of knowledge diffusion. It's a joint project between Indiana University Bloomington and the University of North Carolina at Chapel Hill; and the focus is on collecting dissertation data from a wide variety of disciplines and institutions.

As you can imagine, sourcing such data was never going to be a simple and straightforward task. "I started working on this project when I was still technically a master's student," said Cassidy. "The aim was to explore relationships between authors, advisors and committee members to understand how this has influenced Library and Information Science (LIS) as a discipline. We also wanted to know how these relationships may have shaped the development of new disciplines over time."

Using ProQuest® Dissertations & Electronic Theses database, Cassidy looked for information on authors that had published dissertations in LIS. At the time, the database included metadata such as title, school, author name, advisors (although this was only available for the past few decades)—but it didn't include a field for committee members. Cassidy had to read through the dissertations to find out names of advisors, and then look these up to determine the subjects in which they'd published their own dissertations. In some cases, the dissertation was not available electronically meaning time spent in front of a microfiche reader too. "Yeah, it was quite a tedious task!" agreed Cassidy. "So it was great when a few years ago ProQuest started to include metadata for advisors too. Finding the information became much easier."

To speed the research process, Cassidy decided to reach out to ProQuest on the off-chance they might be able to help. "I was delighted when ProQuest said that not only would they be happy to help, but they were also willing to provide all the information I needed in XML format. Suddenly, I'd saved myself weeks and weeks of work. I thought they might be happy to provide some sample data—as other companies most probably would have done—but to receive the entire data file was amazing!"

The dissertations data provided the framework which the team worked from to find out who had graduated from which school and in which subject area. "ProQuest is the most comprehensive source of information out there that tells you who has graduated from where and in which subjects in the USA. We started by looking at programs that offered Library and Information Science degrees initially but soon realized as we delved deeper in the backgrounds of advisors, that in many cases they'd got their own degree from another discipline, such as English. Had we continued to focus the research just on those LIS programs, we would be missing very important data."

Effectively, the dissertations data could be used to identify interdisciplinary works, determine which disciplines these included, and how they may have contributed to the development of new disciplines. "What has also been fantastic about this project is that it's helped us to generate metrics for measuring other work that researchers do in their role as mentors and advisors. For example, researchers get credit for their work through citations, but never get credit for all the mentoring they do."

Because ProQuest constructs abbreviated abstracts for their dissertations alongside the title and other metadata, it provides a rich source of information that supports topic modeling too. "If you wanted to know what is trending in academia, the ProQuest data can help you," continued Cassidy. "I think ProQuest is putting the finger on the pulse of science, and gives a more honest representation of what's really happening. For example, in the hard sciences there is a lot more funding which means more research gets published, so people have better knowledge of what is "hot" right now. But in other disciplines, such as the humanities, there is much less funding so it's hard to gauge the same level of detail.

"Using ProQuest means it's possible to find what is being published and how that research is impacting other disciplines, which in turn helps to identify where to put the money to progress key areas."

While the purpose of the data was to help populate the MPACT database, during the research process other avenues presented themselves which Cassidy was keen to explore. These included exploration of productivity and accountability; and one project Cassidy is working on right now is the productivity of males vs. females. This looks at, among other themes, how the mentoring of females by males has impacted future career direction. "As ProQuest provides the name of the author, the date of graduation, the institution etc., it's been possible to run an algorithm to overcome author disambiguation and ask more complex questions about mentoring, such as: How successful was the relationship? Was gender a factor? Was institution a factor? What did the female do after her degree? Did she remain in academia? And to what extent did gender of mentor impact this decision, if at all."

Another avenue is exploring the use of words in dissertations and the extent to which this usage is impacted by the discipline or the gender of the person who has written it. "The rich abstracts included in the data, and the mono-disciplinary nature of ProQuest, means that we can really hone in on the use of language, and look at ways of speaking to work out if it's the gender or discipline that controls an author's use of language. We can also to try and answer questions such as whether having a male or female advisor influences the use of words in any way."

Without doubt, access to ProQuest's dissertations data has meant that Cassidy and the MPACT team can forward their research much more quickly than would have been possible otherwise. "I've been so grateful to have this data. ProQuest has been forthcoming with the information they've provided and if we've had any gaps, they've been really happy to help us plug them. I think that ProQuest really cares about how its dissertations data is being used, and they want to provide service that has good information and is really useful. There are so many things I wouldn't be able to do if I didn't have this data and I want to thank them for their help and support."

D3710