figshare • Computer Science

Reusing data and building upon it

a case study of Computer Scientist Mayank Kejriwal

Key Points

- Open, reusable data is important for computer scientists like Mayank to be able to pull together and develop upon.
- Mayank suggests splitting your data up into manageable chunks and sharing them separately, but linking them together. This makes it easier for people to decide what data to use and makes it less daunting to approach.
- When talking to others about making their data open, Mayank suggests sharing some examples of open data with researchers so they can understand what the end product looks like.

About Mayank

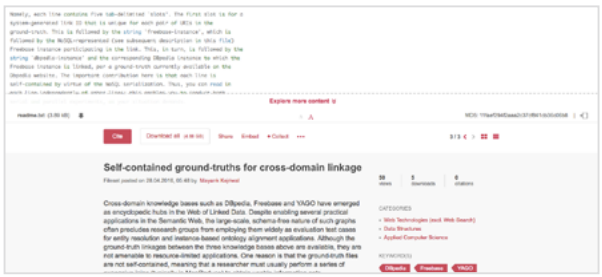
Mayank Kejriwal is a computer scientist in the Information Sciences Institute at the University of Southern California. Prior to starting this position, Mayank completed a PhD at the University of Texas at Austin. This case study is a look at his PhD dissertation and the results of making his outputs openly available.

In 2016, Mayank was working on his PhD in computer science in Austin, focussing specifically on the semantic web, an area of computer science that specializes in intelligent web design that can be read directly by machines. Mayank developed techniques within this area to help identify connections between things that had previously been unconnected by computers. For example, two reviews for the same item on two different store websites could now be identified and brought together within the semantic web.

These techniques had to be evaluated in order to ensure they were accurate. Mayank used Figshare to openly share a portion of the outputs from his PhD as a fileset, titled Self-contained ground-truths for cross-domain linkage.

To obtain this data, Mayank used openly-available data, ran some processing algorithms on them, and developed a gold-standard to be able to evaluate new entity resolution algorithms. An entity resolution algorithm is a computer program that can tell when two things are one and the same.

In the case of products, for example, the program would be able to make out that both products are one and the same, even though they are on different websites and look slightly different. Because machines still don't do as well on this task as humans, such programs have to be evaluated on gold-standard data that have been manually labeled by humans.



Mayank's reserarch on Figshare

However, not all data is as openly available as the datasets Mayank used to generate his own data.

"There's a real disconnect between what's happening in academia and what's happening in industry," said Mayank. "In academia, people develop very fancy algorithms but they don't know if they're going to work or not on real datasets because they don't have access to these datasets. Industry are keeping these real datasets behind closed doors."

Mayank's dataset is big, noisy, and captures what computer scientists would expect to see from an industry-held dataset. He is keen to continue to develop these datasets using data that is available for reuse and doesn't contain restrictive licenses. Computer scientists spend a great deal of time getting datasets in shape for reuse time that could be better spent elsewhere.

The major drivers for Mayank sharing his work openly were to expose the data, to have others to build upon the work, and to make it as easy as possible to reuse the data. For anyone looking to similarly share their data, Mayank recommends splitting the data up into manageable chunks and sharing them separately, but linking them together. This makes it easier to sift through the data and find what's relevant.

When talking to others about making their data open, as well, Mayank suggests sharing some examples of open data with researchers so they can understand what the end product looks like.
"When I go to conferences and present my data, I don't just present my data: I present the process for sharing the data," said Mayank. "I also walk them through the process of how to download the data and how to reuse it."

Mayank continues to try to convert other computer scientists one at a time to openly share their own data.

Read more about Mayank's research!