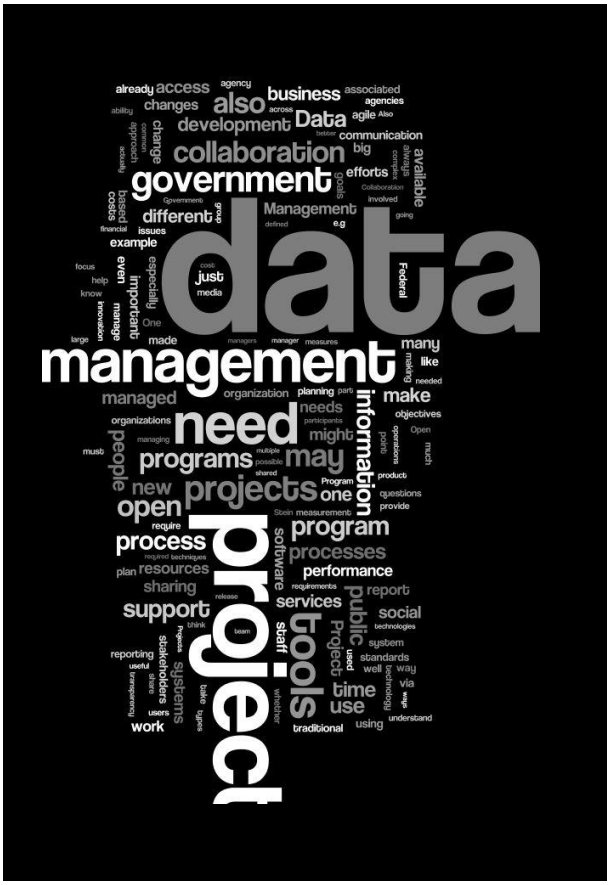

Some Perspectives on Sharing Large Open Research Data Sets

Dennis D. McDonald, Ph.D.

Some Perspectives on Sharing Large Open Research Data Sets

By Dennis D. McDonald[1]

August 3, 2016

Sharing open data

We're probably going to see an increasing number of of reports like Genetic Drivers of Immune Response to Cancer Discovered Through 'Big Data' Analysis where access to and analysis of a large body of previously collected data leads to significant findings. To quote the report:

> *"This work emphasizes the value of open data," Godzik added. "Because we could access genomic data from over 5,000 tumor samples from the Cancer Genome Atlas (TCGA), we could jump straight to analysis without having to set up a big collaborative network to gather and sequence so many samples."*

Getting more eyeballs and brains looking at good data can potentially lead to positive outcomes, including findings that were unanticipated by the original data sources. More analysis and more serendipity means more findings, right? After all, the traditional research model is to repeat experiments in order to prove or disprove findings. "Reanalyzing" data that someone else has already been through would seem to be in line with that.

Questions about large data sets

Yet, re-analyzing the same data is not always what opening up large complex data sets for additional and potentially innovative scrutiny is all about. With today's large and constantly growing data sets it's unlikely -- perhaps even impossible -- that everything useful or interesting will be found the "first time

---

through." Very large data sets, especially when data are regularly added to or replenished (for example, when collaborating researchers add data in standard form to reflect their own experiments, or when remote sensing or satellite data constantly generate new data in real time) it's probably to everyone's advantage to try something new in the way of analysis. Still, there are some cautions that should be taken into account when planning the analysis:

- Is the data set being made available to others really static?
- Are the analysis or modeling approaches more appropriate to static data?
- Is it appropriate to ask the same questions when the underlying data reflects changes or modifications that might make the performance of identical analyses problematic?
- Has documentation about how the data set has evolved been made available and taken into account when additional analyses are being planned?
- If novel or innovative analytical or modeling approaches are being proposed (e.g., using game theory to model tumor behavior) are there special demands on the data that might not have been anticipated by the original data source?

Go "straight to analysis"?

Another consideration is, what (if anything) would be lost when, as Godzik suggests, "… we could jump straight to analysis without having to set up a big collaborative network to gather and sequence so many samples."

My thinking about going "straight to analysis" reflects a combination of economics, philosophical, and community considerations.

Economics

There's no question that analyzing existing data can help the researcher avoid some of the costs associated with data collection. The significance of costs savings will vary across projects given that planning and design costs will still be incurred. The question of "who pays for the data" may also arise in situations where a general infrastructure for collecting and managing standard data and data-related costs does not already exist.

This is one of the reasons, for example, that NOAA has put so much thought into planning and engaging with private sector vendors when developing its own open data program given the costs incurred in making its specialized environmental and satellite data more accessible to the public.

Philosophy

I wonder about what might be lost to the researcher if some of the more complex and even painful -- and messy -- activities associated with data collection are regularly avoided. Working through the details of how each and every data element will be gathered, cleaned, processed, and managed can be both a humbling as well as an educational experience. This is especially the case when complex data requiring human intervention and judgement are involved, as can be the case with clinical data being shared among organizations with different coding standards.

Community

Finally, there is something to be said for what is gained when a community forms around not only data analysis but all the processes associated with a data lifecycle. Much can be gained from collaborating especially when multiple organizations are involved in generating and collecting data, as is certainly the case with many government programs that work through the cooperation of many different state, local, international, private- and public-sector organizations in the delivery of services. Building such relationships can be a significant aid to both data analysis and data impact.

Related reading

- *Challenges of Public-Private Interfaces in Open Data and Big Data Partnerships*
- *Developing a Collaborative Approach to Improving Project Management Practices, Part 1: Culture*
- *Geography and Innovation: Additional Metrics for Assessing the Impact of Collaboration on Federal Acquisition*
- *Interim Report on the Generalizability of the NOAA Big Data Project's Management Model*
- *Needed: Better Integration of Project Management and Data Management*
- *On Managing Health Data Programs: Some Thoughts After the Health Datapalooza Conference*
- *The "Open Movement" Continues: Big Pharma Company Opens Up Access to Clinical Trial Data*
- *Should Clinical Trial Data Sharing Be a Precondition for Refereed Journal Article Acceptance?*
- *Thinking About "Data Program Governance"*