Integrated Data Management Processes Expedite Common Data Management Tasks in Autism Research

Frank J. Farach, Naralys Sinanis, Julie Hawthorne, Henry Agnew, Tricia Schantz, Bill Jensen, Leon Rozenblit

Background
Many researchers engage in disposable data management (DDM) data management practices: They clean and organize data after a study has been finished, repeating the process every time a new analysis needs to be done. In contrast, integrated data management (IDM) is a systematic approach to managing data as a reusable resource. It reduces redundancy by centralizing, cleaning, and organizing data once, up front, and making it available via a query interface. Case studies suggest that adopting IDM processes within an IDMS (IDM System) can improve efficiency, reduce costs, increase data quality, and facilitate data reuse. However, no empirical data have been reported evaluating the performance of IDM-like methods in psychological research. We examined whether our organization’s adoption of IDM practices allowed data analysts to more quickly complete investigators’ data-management requests compared to pre-existing DDM practices at client organizations. We predicted that IDM would be associated with faster overall response times and greater efficiency gains when responding to repeated requests.

Methods
We developed the Research Exchange Database (RexDB), an open-source IDMS that can accommodate multiple studies, heterogeneous data types, and multi-site collaboration. Using two years’ worth of available records logged by time- and request-tracking software, we examined the efficiency with which analysts responded to investigators’ data-quality, data-analysis, and study-tracking requests. We computed the average time it took the data managers to complete each request, including time spent communicating with investigators and working with the database system. We asked clients to retrospectively estimate the time it took to manage data with DDM practices, as these data pretrained the data collection period for IDM practices.

Results
Clients’ retrospective ratings suggested that, prior to implementing IDM processes with a compatible system, it took approximately 330 min to complete common data management requests and 240 min to complete subsequent requests of the same kind. However, using IDM practices within an IDMS, analysts more quickly completed the 87 data management requests received during the sampling window (M = 144.0 min, SD = 168.0 min). Time to completion varied by the type of request, F(2, 84) = 0.048, MSE = 26766, p = 0.048, with data quality requests taking the longest time (M = 177.0 min, SD = 157.3 min), followed by data analysis (M = 133.0 min, SD = 212.9 min) and study tracking (M = 62.8 min, SD = 104.4 min). The fourteen (16.1%) requests classified as repeat requests were completed more quickly (M = 56.6 min) than new requests (M = 160.8 min), Welch’s t(71.19) = -4.09, p < .001, 95% CI [-152.0 min, -52.4 min].

Conclusions
Adopting an integrated data management (IDM) process with a compatible system can significantly expedite common data management tasks. The greatest challenge was integrating data sources with varying degrees of cleanliness. Data cleaning and organization are likely to remain challenging, as the number of data sources to manage continues to increase. We expect IDM practices will be most valuable when both the cost of data acquisition and the probability of data reuse are high.

Problem/Challenge
- Disposable data management cleans data from each source multiple times; it’s redundant and inefficient, slowing exploration
- Even small tweaks to a research question may lead to re-cleaning work
- The large amount of re-work required before preliminary data analysis introduces friction and discourages rapid exploration of new research questions

Time to Completion...

...by Request Type

...by Repetition

Conclusions
Adopting an integrated data management (IDM) process with a compatible system can significantly expedite common data management tasks. The greatest challenge was integrating data sources with varying degrees of cleanliness. Data cleaning and organization are likely to remain challenging, as the number of data sources to manage continues to increase. We expect IDM practices will be most valuable when both the cost of data acquisition and the probability of data reuse are high.