Foregrounding Data Curation to Foster Reproducibility of Workflows and Scientific Data Reuse

Michael R. Gryk
University of Illinois, Urbana-Champaign USA
gryk2@illinois.edu

ABSTRACT

Scientific data reuse requires careful curation and annotation of the data. Late stage curation activities foster FAIR principles which include metadata standards for making data findable, accessible, interoperable and reusable. However, in scientific domains such as biomolecular nuclear magnetic resonance spectroscopy, there is a considerable time lag (usually more than a year) between data creation and data deposition. It is simply not feasible to backfill the required metadata so long after the data has been created (anything not carefully recorded is forgotten) – curation activities must begin closer to (if not at the point of) data creation.
The need for foregrounding data curation activities is well known. However, scientific disciplines which rely on complex experimental design, sophisticated instrumentation, and intricate processing workflows, require extra care. The knowledge gap investigated by this research proposal is to identify classes of important metadata which are hidden within the tacit knowledge of a scientist when constructing an experiment, hidden within the operational specifications of the scientific instrumentation, and hidden within the design / execution of processing workflows.

Once these classes of hidden knowledge have been identified, it will be possible to explore mechanisms for preventing the loss of key metadata, either through automated conversion from existing metadata or through curation activities at the time of data creation.
The first step of the research plan is to survey artifacts of scientific data creation. That is, (i) existing data files with accompanying metadata, (ii) workflows and scripts for data processing, and (iii) documentation for software and scientific instrumentation. The second step is to group, categorize, and classify the types of "hidden" knowledge discovered. For example, one class of hidden knowledge already uncovered is the implicit recording of data as its reciprocal rather than the value itself, as in magnetogyric versus gyromagnetic ratios. The third step is to design/propose classes of solutions for these classes of problems. For instance, reciprocals are often helped by being explicit with units of measurement. Careful design of metadata display and curation widgets can help expose and document tacit knowledge which would otherwise be lost.

TOPICS

data curation; user interfaces; classification; abstracting; metadata; ontologies