Geophysical Research Abstracts
Vol. 20, EGU2018-18532-1, 2018
EGU General Assembly 2018

Using data for system-level science: A provenance perspective

Barbara Magagna (1), Malcolm Atkinson (2), and Markus Stocker (3)

(1) Ecosystem Research & Monitoring, Umweltbundesamt GmbH, Vienna, Austria (barbara.magagna@umweltbundesamt.at), (2) School of Informatics, University of Edinburgh, Edinburgh, Great Britain (malcolm.atkinson@ed.ac.uk), (3) German National Library of Science and Technology (TIB), Hannover, Germany (markus.stocker@tib.eu)

The quantities researchers report in scientific literature, say summary statistics such has 5:47 for the mean duration of a studied phenomenon, are generally the result of complex workflows. While not always obvious from reading the reported materials and methods, such values may be derived from numbers generated by an instrument of an observatory; acquired, curated, and published by a research infrastructure; processed using one or more computational models; and interpreted by a postgraduate student supervised by a postdoc who may ultimately derive the reported summary statistics.

At each step in the workflow, the data derived as output in the computation are generally different from the primary data used as input in the computation. Furthermore, workflows involve different kinds of agents, machines and humans, as well as activities, such as data analysis or data interpretation. In using environmental data - specifically observational data but also experimental and computational data - for system-level science we have thus much provenance as a side product.

Unfortunately, provenance is seldom recorded systematically, even though it would arguably be extremely useful. Indeed, provenance helps those who wish to validate results or re-use a method. It also provides evidence of the value of data as well as contributions of individual researchers and institutions. While there exist various conceptual models for provenance representation, e.g. W3C PROV, and workflow engines that integrate them, infrastructural support for provenance recording remains insufficient, especially for human-in-the-loop workflows. Among the challenges, one difficulty is infrastructural discontinuity that results when data published by research infrastructure data portals are downloaded for subsequent local processing. Even if the well-engineered research infrastructure supports recording provenance, the lineage breaks when data published by such infrastructure are downloaded, processed, and analysed outside its realm. When the results of analysis are returned to general use, the researchers who have worked in the downloaded context may have to construct required provenance records from their notes.

Building on a concrete use case in aerosol science, this work highlights the problem and discusses one possible approach whereby a human-in-the-loop workflow using observational data for system-level science is provided by infrastructure “as a service” to research communities. Hence, the approach overcomes the infrastructural discontinuity. As a result, a research infrastructure not only records the provenance of published data derived from the instruments of observatories but also the provenance of abstract statistical data computed for system-level science resulting in analysis of primary data.

The presented work is embedded in a task of the ENVRIplus project dedicated to provenance, where the project more broadly analyses the requirements and develops approaches to address the needs of specific research infrastructures. These will be represented systematically in the ENVRI Reference Model and ENVRI Knowledge Base, feed into the RDA Provenance Patterns WG and possibly into other related activities. Ultimately we aim at comprehensive provenance management for the entire research data life-cycle, and demonstrate the developed approaches on use cases such as the one presented here.