PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20380
doi: 10.3897/tdwgproceedings.1.20380

OPEN ACCESS

Conference Abstract

Using YesWorkflow hybrid queries to reveal data lineage from data curation activities

Qian Zhang[‡], Paul J. Morris[§], Timothy McPhillips[‡], James Hanken[|], David B. Lowery[|], Bertram Ludäscher[‡], James A. Macklin[¶], Robert A. Morris[#,|], John Wieczorek[|]

‡ University of Illinois Urbana-Champaign, Champaign, United States of America
§ Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States of America
| Museum of Comparative Zoology, Harvard University, Cambridge, United States of America
¶ Agriculture and Agri-Food Canada, Ottawa, Canada
# University of Massachusetts, Boston, Boston, United States of America

Corresponding author: Qian Zhang (zhangqian06@gmail.com), Paul J. Morris (mole@morris.net)

Received: 16 Aug 2017 | Published: 16 Aug 2017

Citation: Zhang Q, Morris P, McPhillips T, Hanken J, Lowery D, Ludäscher B, Macklin J, Morris R, Wieczorek J (2017) Using YesWorkflow hybrid queries to reveal data lineage from data curation activities. Proceedings of TDWG 1: e20380. https://doi.org/10.3897/tdwgproceedings.1.20380

Abstract

The YesWorkflow McPhillips et al. 2015b, McPhillips et al. 2015a toolkit was designed to annotate data curation workflows in conventional scripts (e.g., Python, R, Java) but it can also be used to annotate YAML-based Kurator workflow configuration files. From just a file that has been annotated by YesWorkflow, YesWorkflow is able to render a top-level graphical view of the workflow structure (prospective provenance), including system inputs and outputs, actors, connections among those actors, and expected data to be passed on those connections.

YesWorkflow also supports dynamic analysis and reporting on the results of the workflow (retrospective provenance) at various levels of granularity (e.g., at the actor level, script level, data level, record level, file level, function level), provided that it has been configured at each. YesWorkflow includes an @Log annotation, which describes the semantic structure of a log message within some actor in the workflow and allows the log message to be linked to the actor within which it was created, and for parts of that log message to be linked to the data passed between actors. YesWorkflow can be used to analyze the log messages after a run of the workflow and construct a store of facts, which can be queried and reasoned upon to make statements about the evolving paths taken by particular data elements through the workflow and assertions made about those data elements within the workflow.

Provenance, like other metadata, appears to be rarely actionable or immediately useful for those who are expected to provide it. However, by refactoring and integrating runtime observables generated from retrospective provenance and context information from prospective provenance analysis into *hybrid queries*, we show how both elements can yield hybrid visualizations that reveal “the plot” of the whole execution. In this way, a comprehensive workflow graph and a customizable data lineage report are made actionable for a workflow run with meaningful provenance artifacts. Queries run on a set of facts extracted from log messages by YesWorkflow after a workflow run, in combination with the facts extracted from the annotated workflow itself, allow for powerful visualizations of the retrospective provenance of a workflow run and of particular data records within a branching workflow.

Keywords

Biodiversity Informatics, Data Quality, Workflows, Provenance

Presenting author

Qian Zhang

Funding program

NSF DBI 1356438 and 1356751

References

- McPhillips T, Bowers S, Belhajjame K, Ludäscher B (2015a) Retrospective Provenance Without a Runtime Provenance Recorder. 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP'15). URL: https://www.usenix.org/conference/tapp15/workshop-program/presentation/mcphillips
- McPhillips T, Song T, Kolisnik T, Aulenbach S, Belhajjame K, Bocinsky RK, Cao Y, Cheney J, Chirigati F, Dey S, Freire J, Jones C, Hanken J, Kintigh K, Kohler T, Koop D, Macklin J, Missier P, Schildhauer M, Schwalm C, Wei Y, Bieda M, Ludäscher B (2015b) YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. International Journal of Digital Curation 10 (1): . https://doi.org/10.2218/ijdc.v10i1.370