Geophysical Research Abstracts
Vol. 19, EGU2017-19076-4, 2017
EGU General Assembly 2017

Active Provenance in Data-intensive Research

Alessandro Spinuso (1), Andrej Mihajlovski (1), Rosa Filgueira (2), and Malcolm Atkinson (3)
(1) KNMI, R&D Observation and Data Technology, Utrecht, Netherlands (spinuso@knmi.nl), (2) BGS, British Geological Survey, (3) University of Edinburgh, School of Informatics

Scientific communities are building platforms where the usage of data-intensive workflows is crucial to conduct their research campaigns. However managing and effectively support the understanding of the 'live' processes, fostering computational steering, sharing and re-use of data and methods, present several bottlenecks. These are often caused by the poor level of documentation on the methods and the data and how users interact with it. This work wants to explore how in such systems, flexibility in the management of the provenance and its adaptation to the different users and application contexts can lead to new opportunities for its exploitation, improving productivity.

In particular, this work illustrates a conceptual and technical framework enabling tunable and actionable provenance in data-intensive workflow systems in support of reproducible science. It introduces the concept of Agile data-intensive systems to define the characteristic of our target platform. It shows a novel approach to the integration of provenance mechanisms, offering flexibility in the scale and in the precision of the provenance data collected, ensuring its relevance to the domain of the data-intensive task, fostering its rapid exploitation.

The contributions address aspects of the scale of the provenance records, their usability and active role in the research life-cycle. We will discuss the use of dynamically generated provenance types as the approach for the integration of provenance mechanisms into a data-intensive workflow system. Enabling provenance can be transparent to the workflow user and developer, as well as fully controllable and customisable, depending from their expertise and the application's reproducibility, monitoring and validation requirements. The API that allows the realisation and adoption of a provenance type is presented, especially for what concerns the support of provenance profiling, contextualisation and precision. An actionable approach to provenance management will be also discussed, enabling provenance-driven operations at runtime, regardless of the enactment technologies and connectivity impediments. We proposes a framework based on concepts such as provenance clusters and provenance sensors, envisaging new potential for exploiting large quantities of provenance traces at runtime.
Finally the work will also introduce how the underlying provenance model can be explored with big-data visualization techniques, aiming at producing comprehensive and interactive views on top of large and heterogeneous provenance data. We will demonstrate the adoption of alternative visualisation methods, from detailed and localised interactive graphs to radial-views, serving different purposes and expertise.

Combining provenance types, selective rules, extensible metadata with reactive clustering opens a new and more versatile role of the lineage information in the research life-cycle, thanks to its improved usability. The flexible profiling of the proposed framework offers aid to the human analysis of the process, with the support of advanced and intuitive interactive graphical tools. The Active provenance methods are discussed in the context of a real implementation for a data-intensive library (dispel4py) and its adoption within use cases for computational seismology, climate studies and generic correlation analysis.