Chapter 21
Integrating Data Management and Collaborative Sharing with Computational Science Research Processes

Kerstin Kleese van Dam
Pacific Northwest National Laboratory, USA

Mark James
University of California San Diego, USA

Andrew M. Walker
University of Bristol, UK

ABSTRACT

Scientific Data Management – the management of storage, access, usage, lifecycle, content, and meaning for scientific data – and the collaborative sharing of it, is not as commonly employed in computational science as it is in other fields. However, where these have been co-developed and in particular tightly integrated with the computational science research process, they have had a transformational influence on scientists' work processes. These efforts enabled not only new and previously impossible research, but also helped to speed up research processes and improve research output.

This chapter describes the key principles and components of a good data management system, provides real world examples of how these can be successfully integrated with scientific research processes and enable successful data sharing, provides an outlook on future developments, and discusses lessons learned. We conclude with a short section on how to get started for those whose interest has been piqued.

DOI: 10.4018/978-1-61350-116-0.ch021

INTRODUCTION

Scientific research can be characterized by its aim to make descriptive, explanatory and predictive inferences on the basis of observed or simulated information about the real world. Ideally it uses explicit, codified and public methods and rules for its data collection and analysis. Repeatability, reproducibility and transparency are seen as the main pillars of good scientific research (King, 1994). In current computational science research these aims are often difficult to achieve because of its inherent complexities and distributed nature.

Scientists today rarely engage directly with their research object, but do so via digitally captured, reduced, calibrated, analyzed, synthesized and visualized data in combination with computer simulations of the processes of interest. Advances in experimental and computational technologies have led to an exponential growth in the volumes, variety and complexity of this data (Southan, 2009; Goble, 2009), and whilst the data deluge is not found everywhere in an absolute sense, it is seen in a relative sense within most research groups. Many lack the methods, tools and infrastructure to deal effectively with the increasing volumes, complexity and geographical distribution of the relevant data. But it is not data alone that challenges the scientific community. Scientists use a much more varied and extensive array of software products to engage with their data, combined in ever more complex workflows that are executed on very different platforms, at times unknown to the user (grids or clouds). This makes it much more difficult to follow the aims of good scientific research practices in terms of repeatability, reproducibility and transparency.

Leaving the aspirational aspects of scientific investigations aside, research practice has become much more collaborative than it was even a few years ago (Jones, 2008; Guimera, 2005), and few research projects do not rely on the sharing of processes and data amongst different group members or groups to accomplish their scientific goals. The increasing complexity of scientific challenges requires more interdisciplinary and multidisciplinary information and knowledge exchange (Committee on Facilitating Interdisciplinary Research, 2004). Whilst multidisciplinary data sharing is still rare, sharing of key data sets within particular research communities has become more mainstream in a range of scientific domains such as environmental sciences or biology (Field, 2009). This is often facilitated through dedicated data centers and expert data collections. In other fields, and specifically computational sciences, working practices around the sharing of research results have, however, not changed much over the past years. Research publications are still the main sources of information exchange in the wider community. Unfortunately publications have certain limitations in conveying comprehensive information on a particular subject; there is the limitation in length and thus detail, its main purpose is to convey the scientists' point of view rather than a comprehensive, objective representation of all facts (Shotton, 2009; de Waard, 2006; Kuhn, 1962; Latour, 1987). Publications thus provide at best a very coarse and high level summary of the research work undertaken by the authors. The associated raw and derived data should be a rich source of supporting information, in particular, if coupled with the appropriate metadata and documented scientific workflows, forming a complete research object (DeRoure, 2009). In recognition of the desire by the research community to have access not only to the summary of a research project, but also the underpinning data, more publishers today require from their authors that they share their raw and derived data by depositing it into publicly accessible archives or by providing it on request. However, recent studies have shown (Savage, 2009; Wicherts, 2006) that few authors comply with the journals data deposition requirement and only the enforced deposition before publication seems to provide the desired result, indicating a continued reluctance to share in-depth research results with the general research community.