DataNet – GridSpace Data Management Framework

Daniel Harężlak[1], Eryk Ciepiela[1], Marek Kasztelnik[1], Bartosz Wilk[1], Marian Bubak[1,2]

[1] AGH University of Science and Technology, ACC Cyfronet AGH,
Nawojki 11, 30-950, Kraków, Poland
[2] AGH University of Science and Technology, Institute of Computer Science AGH,
Department of Computer Science, al. Mickiewicza 30, 30-059 Kraków, Poland
emails: {d.harezlak,e.ciepiela,m.kasztelnik,b.wilk}@cyfronet.pl,
bubak@agh.edu.pl

Keywords: data management, metadata recording, storage infrastructures

1. Introduction

There are a few ways to write down a scientific experiment. Some use a workflow notation such as [2] or [3] and some use a set of scripts such as [4]. In any case scientific data is produced which often requires further analysis and annotation for future reference or validation. In fact, the amount of produced data is so big that it becomes a problem to manage it and not to forget about it. Another issues related to data management are provenance and ownership. In this paper we propose a lightweight framework for data and metadata management independent of an experiment computing engine.

2. State-of-the-art

Preservation of scientific experiments and data produced by them is a subject of work for different experiment engine teams. For instance, so called Research Objects are investigated in [1]. The work, however, tends to be specific for a given engine. The scientist is therefore forced to use different solutions when switching between engines or cooperating with someone who uses different software. In terms of service maintenance it became a standard to use cloud-based resources for scalable and reliable deployments. PaaS (platform as a service) software is available for setting up private sites (e.g. [5]) to offer robust services for dynamic data management.

3. Description of the solution

The framework (depicted in Fig. 1) supports the scientist in creating the data model, deploying corresponding data repository to the PaaS site and managing the data.

Building the model consists of creating a set of entities of simple data types and defining relations among them. The process is straightforward and allows users to express their data schema without knowing anything about complex data notations (e.g. RDF, XML or OWL). When the model is ready it can be deployed as a repository owned by a user or a group of users. The repository after deployment exposes a REST endpoint through which data management can be performed. The models stored in DataNet can be discovered by other users and if the related data is accessible publicly combined data sets (data coming from different repositories but belonging to the same model) can be acquired. This constitutes a separate research area where data is no longer being produced but only queried to obtain valid scientific results.

DataNet does not replace existing storage facilities but provides a way to annotate the data already stored on dedicated storage sites. This lets scientists keep track of their data and manage it better. Adaptors for different storage sites are also a part of DataNet for easy data access.

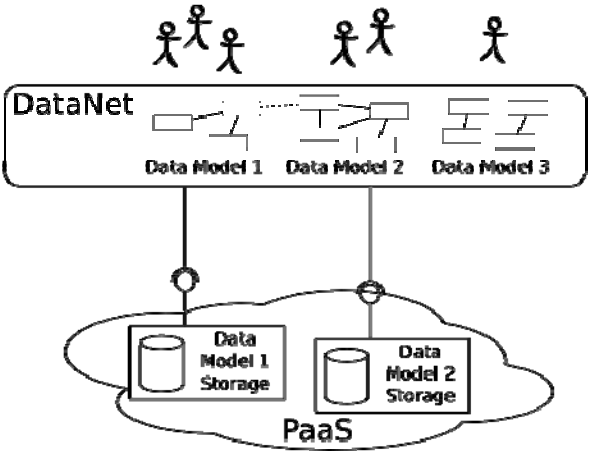

Fig. 1. Concept of DataNet architecture and deployment of data repositories.

4. Results

The first prototype of the framework allows for creation of models with several entities and relations among them. Automatic deployment to the PaaS layer is already in place and a ready to work repository can be setup in a matter of minutes. The deployed repository generates convenient REST templates for a basic CRUD set of operations on entities.

5. Conclusions and future work

The framework prototype for managing scientific data is operational and the tests showed that it can scale well for many repositories. The future work includes handling model modifications and propagating the changes to data repositories. Also, support for different storage technologies would enable wider adoption of the solution in cases of specific requirements such as data size or query rate.

Acknowledgements. The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.02.03.00-00-096/10 as part of the PL-Grid Plus project.

References

1. D. De Roure, K. Belhajjame, P. Missier, J. M. Gómez-Pérez, R. Palma, J. E. Ruiz, K. Hettne, M. Roos, G. Klyne, and C. Goble: Towards the Preservation of Scientific Workflows in 8th International Conference on Preservation of Digital Objects, 2011.
2. P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble: Taverna, reloaded in SSDBM 2010, Heidelberg, Germany, 2010.
3. L. Bertram, A. Ilkay, B. Chad. H. Dan, J. Efrat, J. Matthew, L. Edward, T. Jing and Z. Yang: Scientific workflow management and the Kepler system in Concurrency and Computation: Practice and Experience. Volume 18. Pages: 1039-1065, 2006.
4. E. Ciepiela, P. Nowakowski, J. Kocot, D. Harezlak, T. Gubala, J. Meizner, M. Kasztelnik, T. Bartynski, M. Malawski, M. Bubak: Managing Entire Lifecycles of e-Science Applications in the GridSpace2 Virtual Laboratory - From Motivation through Idea to Operable Web-Accessible Environment Built on Top of PL-Grid e-Infrastructure in M. Bubak, T. Szepieniec, K. Wiatr (Eds) Building a National Distributed e-Infrastructure – PL-Grid – Scientific and Technical Achievements, Springer 2012, ISBN 978-3-642-28266-9, pp. 228-239 (2012).
5. Cloud Foundry Web Site: `http://www.cloudfoundry.com`