# Getting a grip on the grid: A knowledge base to trace grid experiments

## Overview:

Information management is challenging in Science due to the variety of data produced by the physical instruments and the amount of information generated daily by scientists. In addition to supporting the experiment execution, it is currently crucial for the scientific applications to trail their mechanisms of performing experiments, so that it is possible to trace back the resulting scientific data.

In this work, we describe an approach for building a knowledge base for the scientific experiments performed using the e-infrastructure for bioscience (e-Bioinfra [1]). The e-Bioinfra platform provides grid workflow management and monitoring services for biomedical researchers that use the Dutch Grid. Our approach focuses on gathering meaningful information from these services and populating it into the knowledge base, within its proper context. For this, an agent-based software tool is designed and developed to retrieve, classify and transform existing data into meaningful information.

## Description of the work:

A comprehensive knowledge base gathers relevant information to help scientists clarify their research questions and to validate operational tasks. However, building and populating such a knowledge base, with proper and detailed information resulting from different sources, is a challenge in itself. Although the information is usually accessible, e.g. in logs, it is not trivial to correlate pieces of data. Manual data collection is an error-prone task and requires enormous manpower, due to the amount of log registered by the processes. An automated solution is needed.

Our approach to build a knowledge base is a three-folded mechanism. First, the eBioCrawler[2] is designed to gather automatically already existing logs that contain information generated by different application systems (e.g. MOTEUR[3] and DIANE[4]), including workflow descriptions and execution reports, system outputs, communication accounts and status reports. Secondly, a provenance repository is built around the notion of graphs outlined by the OPM model [5]. The repository is defined using a relational database schema that captures the concepts of the OPM model but extending it to support Events. Event-driven systems are common in scientific environments, but their provenance is not well captured by the OPM alone. The provenance data collected by the eBioCrawler is stored into the repository using the Application Program Interface (API) from the Provenance Layer Infrastructure for E-Science Resources(PLIER[6]), allowing developers to build, store and share (by XML serialization) graphs using the OPM model.

The third mechanism enables the analysis of the provenance data. Using the homogeneous view of the information in the repository based on the OPM ontology[7], the provenance can be now transformed, or serialized, into specific formats (RDF, XML, etc.) or other representations. Examples of viewing mechanisms are a graphical interface to analyze the provenance graphs and a query interface to find events of interest.

## Impact:

The described solution delivers a set of tools that have a potential diverse audience. We foresee three main areas of impact: Scientific, Usability, Operational.

Because of the wide and distributed scale of the resources, the “grid” often becomes a black box to users, due to the fact that: a) jobs may get scheduled arbitrary on distributed resources, b) logs files will be generated on the remote nodes selected by the scheduler, c) the user may lose access to the remote nodes where those files reside, and d) the logs may get cleaned without prior notification. In addition, the increased volume of data, resources, methods and

collaborators in the domain of e-Science, it becomes more and more difficult to perform repeatable, scalable, and traceable experiments.

The Distributed information is now collected in a structured way that can be queried for various goals like experiment results, operational statistics, tracing error, customized reporting, etc. By tracing the whole history of the resources up to the current state, it is possible to confirm or provide evidence of the scientific work done. Analysis and comparative mechanisms, scientific opinions or interpretations, and the results of various kinds of examinations may provide further ways to improve (or debug) the actual scientific experiment (workflow).

Lastly, we think that the insight that our solution provides in all the data and resources, will present a important added value to the operational aspects of running a science gateway. These benefits vary from data retention strategies, assistance for workflow debugging, and producing all kinds of usage statistics.

## Conclusions:

The knowledge base for e-Bioinfra was achieved by implementing OPM and Dublin Core[8] data models into a relational database management system. The rich and complex data available within the e-Bioinfra application were challenging enough to test and validate our approach. These data relates to a few thousands of scientific experiments, which have been performed using Moteur/Diane workflow system.

Provenance data for e-Bioinfra, including workflows description and logs files, is collected automatically from the various e-Bioinfra components and transformed into a knowledge base repository. This knowledge base can be accessed for analysis of the experiment provenance in various forms (e.g. GUI for provenance graphs)

Although establishing the knowledge base is fundamentally for documentation, specific tools are needed to be implemented in order to better explore the information in the knowledge base. Each tool could be tailored to the specific need of the application domain and the type of users.

## Primary Authors:


- Dr. BENABDELKADER, Ammar (PCC UvA) <ammar@pccuva.nl>
- SANTCROOS, Mark (Academic Medical Center Amsterdam) <m.a.santcroos@amc.uva.nl>
- Dr. GUEVARA MASIS, Victor (Nikhef) <vguevara@nikhef.nl>


## Co-authors:


- Dr. MADOUGOU, Souley (PCC UvA) <souley@pccuva.nl>
- Prof. VAN KAMPEN, Antoine (AMC) <a.h.vankampen@amc.uva.nl>
- Dr. D. OLABARRIAGA, Silvia (Academic Medical Center Amsterdam) <s.d.olabarriaga@amc.uva.nl>


## References:


1- ebioInfra: http://www.ebioscience.amc.nl/
2- eBioCrawler: http://www.bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/EBioCrawler
3- MOTEUR: http://www.bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/MOTEUR
4- DIANE: http://www.bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/DIANE
5- Open Provenance Model: http://openprovenance.org/
6- PLIER: http://twiki.ipaw.info/bin/view/OPM/Plier
7- OPM Ontology: http://openprovenance.org/model/opmo
8- Dublin Core: http://www.dublincore.org/