



5th BELIEF Brainstorming Workshop
Co-organised with the CASPAR project
6 & 7 April 2009 – Athens, Greece

Sustainable e-Infrastructures: Challenges in Data Provenance and Authenticity

Today's e-Infrastructures provide online scientific communities with a viable means for large-scale, cross-domain collaboration, leveraged by their inherent ability to store and efficiently share huge volumes of distributed resources and data. In this context of global computing and integration, where new data sets are derived out of existing information at high rates, either manually or automatically, significant challenges are posed to the management, semantics assessment, and quality assertion of the shared data objects.

Data provenance provides a way to address the aforementioned challenges by enabling stakeholders to track down the origins of data, specify their purposes and help to explain their semantics. Moreover, when used together with interoperability, provenance facilitates data re-use and sharing, permits data to be discovered *in context* (remote representation of distributed data objects), and allows their aggregation (contextualization). Indeed, data provenance is of considerable value for all users, especially for scientists, and, to date, a number of provenance systems are already applied to various scientific domains, supporting data management in many ways. However, given the importance of standardization, the tentativeness of interoperability across e-Infrastructures, and the need for sustainability, the accommodation and provision of sophisticated data provenance services are specifically challenged by the lack of the following:

- Widely accepted, domain-independent models for stakeholder responsibility assignment and rights management
- Standardized schemes and vocabularies for the semantic annotation of data with provenance/lineage information
- Reliable methods to trace and evaluate the authenticity and audit the source of born digital data
- Large-scale architectures facilitating collaborative management and sharing of data provenance
- Algorithmic support to tackle the complexity of the problem, and reduce the performance and resource overheads induced by the need to propagate provenance information



- Unambiguous techniques for provenance storage, representation and dissemination

The joint BELIEF & CASPAR Brainstorming workshop leads as the follow up of the discussions held in Lyon, France, on November 24, 2008, during the 6th e-Infrastructure Concertation Meeting and will try to respond to the outcomes of this meeting. It will be held in Athens, Greece and participation will be free of charge and by invitation only. It will consist of a mixture of presentations and discussions addressing several aspects around Provenance of Scientific Data: Authentication, Authenticity, Annotation, Archiving, etc. The workshop aims to tackle the issues above by bringing together participants from several disciplines and with different background and roles – scientists, end-users, librarians. The output of the workshop will be a report which intends to contribute to the drafting of the research agenda on the issues related to data provenance and standardization in e-Infrastructures.