Reviewing Technical Approaches for Sharing and Preservation of Experimental Data

Efraín R. Fonseca C.[*‡] erfonseca@espe.edu.ec
Oscar Dieste[‡] odieste@fi.upm.es
Natalia Juristo[‡♢] natalia@fi.upm.es natalia.juristo@oulu.fi
Estefanía Serral[♯] estefania.serralasensio@kuleuven.be
Stefan Biffl[▲] Stefan.Biffl@tuwien.ac.at

[*]Universidad de las Fuerzas Armadas ESPE, Department of Computer Science, Av. Gral. Rumiñahui s/n, Sangolquí, Ecuador
[‡]Universidad Politécnica de Madrid, Department of Software Engineering, ES-28660 Boadilla del Monte, Spain
[♢]University of Oulu, P.O. BOX 8100, FI-90014 Oulu, Finland
[♯]KU Leuven, Department of Decision Sciences and Information Management, Naamsestraat 69, BE-3000 Leuven, Belgium
[▲]Vienna University of Technology, Favoritenstr. 9/188, A-1040 Vienna, Austria

Abstract

Context: Empirical Software Engineering (ESE) replication researchers need to store and manipulate experimental data for several purposes, in particular analysis and reporting. Current research needs call for sharing and preservation of experimental data as well. In a previous work, we analyzed Replication Data Management (RDM) needs. A novel concept, called Experimental Ecosystem, was proposed to solve current deficiencies in RDM approaches. The empirical ecosystem provides replication researchers with a common framework that integrates transparently local heterogeneous data sources. A typical situation where the Empirical Ecosystem is applicable, is when several members of a research group, or several research groups collaborating together, need to share and access each other experimental results. However, to be able to apply the Empirical Ecosystem concept and deliver all promised benefits, it is necessary to analyze the software architectures and tools that can properly support it.

Goal: Identify the most appropriate technologies for the implementation of the Empirical Ecosystem concept.

Method: For the purpose of technology identification, four features are particularly relevant: Volume of data, architecture, data semantics and manipulation facilities. Those features were surveyed in repositories and data sharing and preservation tools used in the sciences by means of a systematic literature review.

Results: 17 sharing and preservation tools reported in the literature were identified. The fields of Genomics and Proteomics, and secondarily Biology, stand out. Given the importance of those disciplines in today's science and economy, it would not be surprising that many other proprietary tools would have gone unnoticed. Regarding repositories, there are hundreds available (either publicly or restricted access) in the Internet. Typically, they aim at benchmarking, or reanalysis and synthesis of existing empirical studies. Most repositories (both in number and importance) belong to the "hard sciences" (e.g. biology, physics, etc.), but virtually every research area is represented, including ESE.

Most tools and repositories use relational databases for data storage, with very little exceptions. When the amount of stored data is very high (e.g. Genomics), relational databases are being substituted by big data management infrastructures such as Apache™ Hadoop®. Relational databases are also used when data are distributed. Global conceptual models guarantee the interoperability among different data sources. When data are heterogeneous, the situation is more complex. Standard conceptual schemas may not be useful, because the semantics of the local data do not necessarily agree the meaning assigned to the global schema. Likewise, large parts of the conceptual schema may not be applicable to local data sources, and the links among local models may not be easily defined. The current trend is abandoning classical conceptual schemas (e.g. entity-relationship) and standardize the vocabulary of the domain using ontologies.

Manipulation facilities are almost invariably offered using web portals. In some cases, repositories provide web services to give access to data for e-science purposes.

Conclusions: The review of the technologies used for the implementation of repositories and sharing and preservation tools in the sciences shows that common, well-known technologies (particularly, relational databases) can be used for the implementation of the Empirical Ecosystem concept. The only exception is the semantic integration of local models. Instead of comprehensive, global conceptual schemas, ontologies are being increasingly used for semantic integration.