D-LIB MAGAZINE
doi:10.1045/dlib.magazine | ISSN:1082-9873
NOVEMBER/DECEMBER 2011

IN BRIEF

November/December 2011

Report on the Fourth Workshop on Very Large Digital Libraries

Contributed by:
Leonardo Candela & Paolo Manghi
{leonardo.candela, paolo.manghi}@isti.cnr.it
Italian National Research Council
Pisa, Italy
Yannis Ioannidis
yannis@di.uoa.gr
University of Athens
Athens, Greece

The 4th Workshop on Very Large Digital Libraries was held in Berlin, Germany on September 29, 2011, in conjunction with the International Conference on Theory and Practice of Digital Libraries (TPDL 2011) – formerly known as the European Conference on Research and Advanced Technology for Digital Libraries (ECDL). The workshop series started in 2008 with the aim of promoting discussions on the specific research area going under the umbrella of "very large digital libraries" – with a long term goal of establishing it as a research field on its own. This year the workshop called for contributions focusing on "research data" in the context of very large digital libraries and archives.

The workshop program comprised two invited talks and six presentations. The invited talks elaborated on: (*i*) state of the art, issues and open research directions related to content based retrieval in very large datasets of visual documents; and (*ii*) challenges affecting the development of Collaborative Data Infrastructures across scientific disciplines. The six presentations elaborated on: (*i*) functional and architectural requirements of a general bit repository mass processing service, capable of abstracting over several programming models and platforms; (*ii*) digital preservation of relational databases by focusing on the conceptual model, hence considering database semantics as an important aspect of the preservation strategy; (*iii*) a scalable strategy based on the "extract, transform, archive" workflow for automatically addressing research-data problems, ranging from the extraction of legacy data to its long-term storage; (*iv*) The Language Archive (LAT) infrastructure and its transition towards open federated archive environment by means of openness to novel metadata formats; (*v*) performance improvements in the open source Greenstone digital library software that resulted from a more detailed understanding of the demands made of its database component when building large collections; and (*vi*) an extension of the data storage model of Invenio – a software platform for building a web-based (document) repository developed at CERN – to efficiently deal with figures and data.

A brainstorming session concluded the workshop. This final session confirmed the agreement that a digital library can be considered "very large" if any of the aspects on user management, content management, functionality management, and policy management becomes "very large" with respect to (*a*) *volume* – the number of entities is huge; (*b*) *velocity* – the speed requirement for collecting, processing and using entities is demanding; or (*c*) *variety* – the heterogeneity in terms of entity types to be managed and sources to be merged is high. Moreover, it was agreed that very-largeness is a matter of thresholds and challenges capturing the limitations of current solutions.

To pursue the workshop discussion and involve the community in the large in the identification of "challenges" characterising Very Large Digital Libraries, the LinkedIn group "Open Forum on Very Large Digital Libraries" has been created.

A detailed report of the workshop has been submitted to the *SIGMOD Record*. More information about the workshop, the accepted papers and the presentations are available from the dedicated website.