Fourth Workshop on Very Large Digital Libraries

On the marriage between Very Large Digital Libraries and Very Large Data Archives

Leonardo Candela
ISTI - Consiglio Nazionale delle Ricerche
Pisa, Italy
candela@isti.cnr.it

Paolo Manghi
ISTI - Consiglio Nazionale delle Ricerche
Pisa, Italy
manghi@isti.cnr.it

Yannis Ioannidis
University of Athens
Athens, Greece
yannis@di.uoa.gr

1. INTRODUCTION

The workshop series on Very Large Digital Libraries (VLDLs) started in 2008 [12] with the aim of fostering and initiating systematic and constructive discussions on the specific and rather novel research area of "very large digital libraries". Before this, building on long experience in the field of digital libraries and data infrastructures, the authors have spent efforts in the definition of the Digital Library Reference Model [4, 3] and in the definition of the Digital Library Technology and Methodology Cookbook [1]. Both initiatives had the common goal of consolidating digital library research as an independent and well established research field with peculiarities characterized by shared foundations. In line with this path, the VLDL workshop series has a twofold target. On the one hand delineating the boundaries of this research area, in an attempt to motivate its existence as an independent avenue of investigation. On the other hand identifying its foundations and grand challenges, so that research results could be classified and compared in a constructive confrontation. The long-term and ambitious goal is to discuss the foundations of VLDLs and establish it as a research field on its own, with well-defined areas, models, trends, open problems, and technology.

The main outcome of the workshop series [12, 8, 9, 5], also published as SIGMOD Record reports [11, 10], was consolidation of its mission. The presentations and following discussions collected in the years clearly promoted VLDLs as a chapter of their own in computer science research. VLDLs cannot be simply regarded as very large databases storing Digital Library (DL) content, as one may be tempted to assess. In fact, as the Reference Model for Digital Libraries well justifies, DL systems cannot be approached from the perspective of content management only; the dimensions of user, functionality, policy, quality, and architecture management are equally important. Accordingly, DLs become Very Large DLs (VLDLs) when any one of these aspects reaches a magnitude that requires specialized technologies or approaches. Actually, the appellation of "very large" is acquired whenever one of the following features apply to one of the dimensions above:

- *volume*, i.e. the dimension in terms of number of entities to be managed or size is huge;
- *velocity*, i.e. the speed requirements for collecting, processing and using entities is demanding; and
- *variety*, i.e. the heterogeneity in terms of entity types to be managed and sources to be merged is high.

However, there is not yet any threshold or indicator with respect to these features agreed by the community in the large that might be used to clearly discriminate very large digital libraries from digital libraries. Regardless of this, very large digital libraries have been developed – e.g. the Library of Congress[1], the National Science Digital Library[2], Europeana[3], DRIVER[4] – and the demand for infrastructures and services promoting collaboration and knowledge sharing on large scale is growing [6, 13].

The fourth VLDL workshop has been organized in conjunction with the 15th edition of the TPDL 2011 conference [7], which is part of the series of *European Conference on Research and Advanced Technology for Digital Libraries (ECDL)* started in 1997. This year workshop called for topics on theory and practice of VLDLs. More specifically, theoretical or foundational topics covered definitional models and measures (content, functionality, users, and policies), architectural models, and design methodologies for VLDLs. Practical or systemic topics covered ideas, experiments, and practical experiences in system design and implementation. Of particular interests were: integration and federation of DLs, user management, security, sustainability, scalability, distribution, interoperability for content, functionality, quality of service, storage, indexes, and preservation.

Moreover, unlike previous editions, this year the workshop proposed the traversal topic "...on the marriage between Very Large Digital Libraries and Very Large Data Archives ...". The idea was to call for contributions proposing research issues and solutions regarding VLDLs in relationship with research data. Research data is today an "hot" area in the field of DLs. Scientists are more and more realizing the need of tools capable of dealing with the so-called "tsunami"

[1] http://www.loc.gov
[2] http://nsdl.org
[3] http://www.europeana.eu
[4] http://search.driver.research-infrastructures.eu/

of data in order to make it accessible, searchable, reusable, or linked with the research publications it is related with [2]. The digital nature of research data, its requirements for several metadata descriptions (e.g. geo-reference, provenance), efficient-scalable-secure storage/access, advanced visualization and management tools, interoperability solutions, show many overlaps with the main DLs issues and, due to the multidisciplinary nature and cross-organizational character of data archives, with the very large nature of DLs.

2. WORKSHOP PRESENTATIONS

All submitted contributions were peer reviewed by two of the six members of the Program Committee and six were accepted. The workshop structure comprised an invited speakers session followed by the presentation of the six contributions. Each session is analyzed in a separate subsection below.

2.1 Invited talks

This session featured two invited talks. The first focused on indexing and search challenges in the area of very large visual archives. The second focused on interoperability challenges in the construction of a European data archive infrastructure, arising from the EUDAT project.

Amato in the talk entitled "Dealing with Very Large Visual Document Archives" presented state of the art, issues and open research directions related to content based retrieval in very large datasets of visual documents. Content based retrieval is typically performed searching by similarity on the visual (vectorial) features extracted from images. In the last decades researchers have investigated techniques for executing similarity search efficiently and in a scalable way, mostly based on extraction of global visual features [14]. Several techniques were presented, each resulting as an improvement of the existing ones: tree-based access, approximate similarity search, and permutation-based methods. Finally, approaches based on local visual features were presented. These offer much higher retrieval quality, but introduce efficiency issue which are orders of magnitude more difficult.

Thiemann in the talk entitled "The EUDAT Initiative: Challenges and Opportunities" presented EUDAT, a three-year project starting in October 2011 and funded through FP7 e-Infrastructure Call 9, Data infrastructure for e-Science. Its consortium consists of 23 partners from 13 countries and represents 15 user communities from a wide range of scientific disciplines. Emphasis is on development towards a Collaborative Data Infrastructure across scientific communities. The talk highlighted challenges and opportunities as been seen in the current data infrastructure landscape in Europe and addressed within EUDAT.

2.2 Presentation of contributions

This session included the presentations from the six contributions, which covered very large issues on digital libraries in combination with data archives. In particular, the majority of the accepted papers focused on problems related to large scale content storage and management.

In reference to large scale content storage, Jurik and Zierau in the paper entitled "*Different Mass Processing Services in a Bit Repository*", analyzed the requirements of a general bit repository mass processing service that should capable of abstracting over several programming models and platforms. The service is typically needed in large data archives and libraries, where different ways of doing mass processing is needed for different digital library tasks. The investigation shows that the execution environment has an heavy influence on mass processing requirements, hence a general purpose approach is only possible with respect to a given scenario, where common service parameters and organizational issues can be identified. Thompson, Bainbridge, and Suleman in the paper entitled "*Using TDB in Greenstone to Support Scalable Digital Libraries*", discussed about the issues affecting one of the most diffuse digital library software when dealing with large collections. In particular, they evaluated the behavior of the open source Greenstone digital library software when exploited in parallel tasks, identified a drawback residing in the database component and propose some strategies essentially based on the exploitation of a database supporting parallel access by obtaining significant benefits in terms of import time. Finally, Praczyk, Nogueras-Iso, Kaplun, and Šimko in the paper entitled "*A storage model for supporting figures and other artifacts in scientific libraries, the case study of Invenio*", presented an extension of the data storage model of Invenio, a software platform for building a web-based (document) repository developed at CERN. The extension addresses the requirements arising while extending INSPIRE, the information resource in High Energy Physics, to store figures and preserving data tables on which publications are based. Such requirements are in line with current digital libraries challenges to facilitate discovery and access to digital objects distinct from the traditional full-text documents, e.g. figures, data sets or software related to scientific developments.

With regard to large scale content management, Lemire and Vellino in the paper entitled "*Extracting, Transforming and Archiving Scientific Data*", proposed a scalable strategy for automatically addressing research-data problems, ranging from the extraction of legacy data to its long-term storage. The automation of these tasks faces three major challenges: (*i*) research data and data sources are highly heterogeneous, (*ii*) future research needs are difficult to anticipate, (*iii*) data is hard to index. To address these problems, the authors reviewed existing solutions in the business world and proposed the Extract, Transform and Archive (ETA) model for managing and mechanizing the curation of research data. Freitas and Ramalho in the paper entitled "*Relational Databases Conceptual Preservation*" addressed the digital preservation of relational databases by focusing on the conceptual model of the database, hence considering database semantics as an important aspect of preservation "property". This technique enhances previous approaches, which were based on raw format preservation of relational database data and structure. The method is based on Web Ontology Language (OWL) ontologies used to express database semantics as inferred by special algorithms devised by the authors into a prototype. Finally, Elbers and Broeder in the paper entitled "*Federating Live Archives*" described The Language Archive (TLA) infrastructure and its transition towards open federated archive environment by means of openness to novel metadata formats. In this federated archive environment both data and

services are synchronized to multiple sites in order to provide long-term persistency on both levels. The change towards a new metadata format opens the possibilities for other domains to define their own metadata components and structure and thus join the infrastructure. However, this increase in flexibility requires a major update of the archiving and exploitation tools.

3. WORKSHOP DISCUSSION

As in the previous edition, the concluding brainstorming session confirmed the general agreement that a "very-large" Digital Library can be considered as such if any of the axes user management, content management, functionality management, and policy management becomes "very large" with respect to *volume*, *velocity*, or *variety*. This statement gives a particular flavour to this research field, which distinguished it from digital libraries and very large databases. In fact, a digital library with a small-size content base may be considered very large because of its large users base or because of its challenging evolving and unpredictable functional requirements. From this claim, the discussion moved towards the question "what does very-large mean w.r.t. the four axes or to any permutation of them?". Again, the discussion converged on believing that very-largeness is a matter of thresholds and hard-challenges which today shape the limits of our solutions and tomorrow will be hopefully tackled to evolve into newer and harder problems. Very large digital library foundations are still in an early stage and do not help in giving a formal specification to these challenges. As a matter of fact, VLDL limits, hence today's very-large issues, manifests themselves only in real-case scenarios, which researchers are still unable to classify w.r.t. a general theory of VLDL. Consequently, the same holds for the solutions proposed by researchers, which find it hard to confront their work with that of others. More generally, researchers cannot decide to tackle a problem of VLDL research starting from a broader perspective, given clearly stated and agreed on VLDL problematics.

In order to start this re-organization, the audience suggested to pursue a pragmatic approach by first trying to identify common "grand challenges" in the field and subsequently narrow the scope of research to a list of "focused challenges", over which researchers can measure their competences and compare ideas and solutions. This discussion will continue during the next year, through collaborative web tools and based on the volunteering work of researchers. The intention is for the Fifth Workshop on Very Large Digital Libraries to bear these grand and focused challenges as list of topics for article submission.

4. CONCLUSIONS

The main conclusion drawn from all workshop deliberations was that VLDL research has all the attributes to candidate as an independent research field. Not only, its DL flavour rotating around the four dimensions of users, content, functionality and policies, only overlaps in some sense with very large databases (on the content issues) and makes its particularly interesting and innovative in terms of challenges. It was agreed that next year's workshop questions should move one step forward. On the one hand into identifying, across the axes of investigation, a categorization of very large problems for DLs, and on the other hand to continue the quest on solutions to these issues. The LinkedIn group "Open Forum on Very Large Digital Libraries"[5] is today open for researchers willing to cooperate on the first task so as to pave the way to a more constructive confrontation on common research avenues next year, in the next edition.

5. ACKNOWLEDGMENTS

We would like to thank all those who contributed directly or indirectly to the event, especially our colleagues at ISTI-CNR Donatella Castelli and Pasquale Pagano for their advices.

Special thanks are also due to the members of the program committee: *Daan Broeder* (Max Planck Institute for Psycholinguistics, The Netherlands), *Kat Hagedorn* (OAIster System, University of Michigan Digital Library Production Service, USA), *Norbert Fuhr* (Department of Computer Science, University of Duisburg-Essen, Germany), *Leonid Andreevich Kalinichenko* (Institute of Informatics Problems of the Russian Academy of Science, Moscow), *Fabrizio Silvestri* (Institute of Information Science and Technologies, National Research Council, Italy), *Hussein Suleman* (Department of Computer Science, University of Cape Town, South Africa). Their active participation, enthusiasm and research experience largely contributed in making this workshop an attractive venue and in the end a constructive experience for all authors, participants and the community in the large. Finally, our sincere gratitude goes to all participants, invited speakers and authors, whose enthusiasm and vision constitute the soul of this series of workshops.

Workshop proceedings [5] were funded by ISTI-CNR.

6. REFERENCES

[1] G. Athanasopoulos, L. Candela, D. Castelli, K. E. Raheb, P. Innocenti, Y. Ioannidis, A. Katifori, A. Nika, S. Ross, A. Tani, E. Toli, C. Thanos, and G. Vullo. Digital Library Technology and Methodology Cookbook. Deliverable D3.4, DL.org, April 2011.

[2] C. L. Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, pages 1–40, 2011.

[3] L. Candela, G. Athanasopoulos, D. Castelli, K. E. Raheb, P. Innocenti, Y. Ioannidis, A. Katifori, A. Nika, G. Vullo, and S. Ross. The Digital Library Reference Model. Deliverable D3.2b, DL.org, April 2011.

[4] L. Candela, D. Castelli, Y. Ioannidis, G. Koutrika, P. Pagano, S. Ross, H.-J. Schek, H. Schuldt, and C. Thanos. Setting the Foundations of Digital Libraries – The DELOS Manifesto. *D-Lib Magazine*, 13(3/4), March/April 2007.

[5] L. Candela, Y. Ioannidis, and P. Manghi, editors. *Proceedings of the Fourth Workshop on Very Large Digital Libraries (VLDL 2011)*, Berlin, Germany, 2011. ISBN 978 88 95534 11 4.

[5] `http://www.linkedin.com/groups/Open-Forum-on-Very-Large-4118623`

[6] E. Commission. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions a digital agenda for europe. Technical report, European Commission, August 2010.

[7] S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, editors. *Research and Advanced Technology for Digital Libraries - International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26-28, 2011. Proceedings*, volume 6966 of *Lecture Notes in Computer Science*. Springer, 2011.

[8] Y. Ioannidis, P. Manghi, and P. Pagano, editors. *Proceedings of the Second Workshop on Very Large Digital Libraries (VLDL 2009)*, Corfu, Greece, 2009.

[9] Y. Ioannidis, P. Manghi, and P. Pagano, editors. *Proceedings of the Third Workshop on Very Large Digital Libraries (VLDL 2010)*, Glasgow, Scotland, UK, 2010. ISSN 1818-8044.

[10] P. Manghi, P. Pagano, and Y. E. Ioannidis. Second workshop on very large digital libraries: in conjunction with the european conference on digital libraries corfu, greece, 2 october 2009. *SIGMOD Record*, 38(4):46–48, 2009.

[11] P. Manghi, P. Pagano, and P. Zezula. First workshop on very large digital libraries – vldl 2008. *SIGMOD Record*, 37(4):115–117, 2008.

[12] P. Manghi, P. Pagano, and P. Zezula, editors. *Proceedings of the First Workshop on Very Large Digital Libraries (VLDL 2008)*, Aarhus, Denmark, 2008.

[13] C. Thanos. Global Research Data Infrastructures: The GRDI2020 Vision. Technical report, GRDI2010 `www.grdi2020.eu`, 2011.

[14] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search - The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer, 2006.