First Workshop on Very Large Digital Libraries – VLDL 2008

Paolo Manghi
ISTI - CNR
Pisa, Italy
paolo.manghi@isti.cnr.it

Pasquale Pagano
ISTI - CNR
Pisa, Italy
pasquale.pagano@isti.cnr.it

Pavel Zezula
Masaryk University
Brno, Czech Republic
zezula@fi.muni.cz

1. MOTIVATIONS

In today's information society the demand for Digital Libraries is changing. The implementation of nowadays Digital Libraries is more demanding than in the past. Information consumers are facing with the need to have access and elaborate over an ever growing and heterogeneous information space while information providers are interested in satisfying such needs by providing rich and organised views over such information deluge. Because of their fundamental role of information production and dissemination vehicle, Digital Libraries are also expected to provide information society with services that must be available 24/7 and guarantee the expected quality of service.

This scenario leads to the development of Large-Scale Digital Library Systems in terms of distribution, integration and provision of services, information objects, end-users and policies of use. Such systems have to confront with new challenges in a context having scalability, interoperability and sustainability as focal points.

The need for concrete solutions can be seen also in the substantial amount of resources invested by the European Union towards the creation of a unifying European Information Space, starting with DELOS [11], BRICKS [10], DILIGENT [6] and DRIVER [4] in the past, and continuing with D4Science [5], DRIVER-II, CLARIN [8], SAPIR [15] and finally with the European Digital Library, a major effort to build a European-scale digital library to make available to everybody the rich cultural assets of the whole Europe.

New approaches and technologies have been devised to appropriately tackle the various matters arising in designing, developing and deploying VLDL systems. The goal of this workshop was to provide researchers, practitioners and application developers with a forum fostering a constructive exchange among all of such key actors.

2. WORKSHOP OUTLINE

The workshop structure comprised an invited speakers session followed by the presentation of the nine accepted contributions, organized into three sessions: *architectures*, *services* and *data management* for VLDLs.

2.1 Invited presentations

The organizers invited two speakers, respectively in the field of content modeling and architecture design for VLDLs. Both presentations had a foundational flavour, with the purpose of encouraging discussions on common patterns, best practices and methodologies in VLDLs. The first presentation by *Carlo Meghini* (ISTI-CNR) focused on representation models for Complex Objects on VLDL scenarios. The second presentation by *Daan Broeder* (CLARIN Project, Max-Planck Institute) illutrated common challenges to be faced in building real, very-large infrastructures for language resources management.

2.2 Architectures for VLDLs

VLDL architectural issues have to do with organizational models, interoperability, integration, federation, sustainability, scalability, quality of service, policies and how these issues may combine and be solved in the context of specific application domains and research communities. The session presented three experiences in architectural design, respectively focusing on service and content management patterns in VLDLs, VLDL cataloging systems, and organizational and policies issues in VLDLs.

The first presentation, by *Andreas Aschenbrenner* (Max-Planck Digital Library labs), introduced a Digital Library System Warehouse framework inspired by successful patterns for large-scale digital libraries definition. The framework combines two common practices: (*i*) integration of external services, e.g. search, which may be part of the core user requirements, but reside outside of the core architecture; and (*ii*) manipulation of distributed digital objects. As examples of the two patterns, two system architectures where cited: the *eSciDoc* project [2] aims at building an integrated information, communication and publishing platform for web-based scientific work, exemplarily demonstrated for multi-disciplinary applications; the *TextGrid* project [3] supports a collaborative research environment for specialist texts. A beta-version containing about 10 terabyte of objects initially (images, XML-based full text, annotations, etc) and an initial, expandable set of functionality went live in September 2008.

The second presentation, by Gianmaria Silvello (Department of Information Engineering, University of Padua), described the design of a Digital Library System able to collect, manage and share very large archival metadata collections in a distributed environment. Archive characteristics were presented, where size, interoperability and heterogeneity were pointed out as the most relevant and peculiar challenges for the architecture design. The work also included an extension to the architecture, so as to include management of special Compound Digital Objects in the archival context.

The last presentation of the session, by Mary Rowlatt (MDR Partners), presented the *EuropeanaLocal* project [7] and its part in the *Europeana* Digital Library architectural framework [14]. In particular, the project plays an important role in ensuring that the enormous amount of digital content (from museums, libraries archives and archives of images, sound, text and movies) provided by Europes cultural institutions at local and regional level is represented in Europeana, alongside that held at national level. The expected results include (*i*) the establishment of a network of regional repositories that are highly interoperable with Europeana, (*ii*) an integrated Europeana-EuropeanaLocal prototype service and (*iii*) the development of thematic areas for Europeana services which integrate content from both the national and the local/regional level.

2.3 Data management in VLDLs

VLDL data management issues regard content-related aspects, such as services for manipulation, integration, storage, access, search and federation of data in VLDLs. General-purpose solutions are of interest, as well as others specific to given application domains. The session presented three different experiences in data integration, data storage and access, and data ranking in VLDLs.

The first presentation, by Daan Broeder (Max-Planck Institute), focused on the research activities carried out in past projects at the Max-Planck Institute, regarding management and integration of very large heterogeneous multimedia archives. The activities should in synergy serve the purpose of the CLARIN European project: the main issues regarded data models and interoperability (DOBES project), data archiving (LAT project) and synchronization (DOBES project), single sign-on access (DAM-LR EU project) and persistent identifiers.

The second presentation illustrated the Greenstone system [12] experience with the *Papers Past newspaper collection* [13]. This collection, containing 670,000 newspaper digitalized pages (7.5 million articles), growing to approximately 1.2 million pages over time, counting 20Gb of raw searchable text, 2 billion words, 60 million unique terms and 52Gb of metadata is (almost) certainly the largest Greenstone collection ever built. In this scenario, Greenstone developers had to cope with the large quantities of images to be analyzed and with the peculiarity of large number of unique terms to be indexed, which degraded the performance of the standard Greenstone system.

The last presentation, by Mikalai Krapivin (University of Trento), showed the results and benefits of a new ranking algorithm applicable to very large pools of scientific papers. The algorithm, named Focused Page Rank, proposes a trade-off between traditional citation count and basic Page Rank (PR) algorithms. The author believes this solution to be closer to the expectations of real users because, in accordance with the one of the most significant principles of Scientometrics, highly cited papers tend to be more visible in the results and thus attract more citations in future. The rank evaluation technique is scalable and may be applied to very large libraries.

2.4 Services for VLDLs

VLDL service issues include the design and development problems arising in the realization of functionalities for VLDLs. The session presented results in designing and developing three service typologies: a loan service, a store service and user interface service.

The first presentation, by *Ciro D'Urso* (Italian Senate), presented the design of an event-based loan service that provides a single access point over distributed and autonomous digital libraries of textual or electronic or microform books, music, sound recordings, visual materials. The service, currently under experimental use to integrate catalogs from Italian libraries, is designed to scale with the number of data sources and registered users and to cope with the interoperability issues introduced by the different catalog standard and technologies.

The second presentation, by *Stephen Green* (British Library), illustrated the inter-related challenges in building long-term store services for very large document collections in the British Library. The important features of the service, developed by the Library labs, are: (*i*) privileging uninterrupted access to stored objects to continuity of ingest, by supporting disaster tolerance and recovery functionalities; (*ii*) automatic self-monitoring, with the ability to recover damaged files within a store; (*iii*) design independent of any hardware manufacturer, in order to allow storage units to be swapped in and out as required; (*iv*) digital signing techniques to deliver continuous assurance of the authenticity of stored objects from the time of ingest to any future time; and (*v*) metadata-driven management of versioning and successor objects.

The third presentation, by *Massimiliano Assante* (ISTI-CNR, Pisa) presented portal services capable of dynamically integrating and adapting user interface capabilities from functionality services within an e-Infrastructure. e-Infrastructures are very large dynamic service-based environments where user communities can build applications exploiting a set of functionality services that can join or leave the system any time. In this scenario, communities may require centralized web-usage and access to a tailored subset of such functionalities. Due to the dynamic environment, building from scratch centralized portals can be very expensive, due to inevitable maintenance cost. Portal services offer a way to automatically configure a centralized portal that responds to specific user interface needs, based on the functionality services currently available.

3. WORKSHOP CONCLUSIONS

"What exactly are Very Large Digital Libraries?" Some, to answer this question, blur the separation between Very Large Databases (VLDBs) and Very Large Digital Libraries (VLDLs) and regard the latter as VLDBs storing Digital Library content. Since in databases the adjective "Very large" strictly refers to "size of content" (nowadays about 1 Terabyte of space), the implication is that, similarly, VLDLs ought to be DLs storing digital content beyond a given threshold.

Despite being intuitively correct, this answer only partly satisfies DLs practitioners. Indeed, DLs design paradigms cannot be conceptually separated by the relative applications as it happens for DBs. DLs are of use to peculiar user communities whose functionality needs, best practices and behavior are well-accepted DL systems requirements. As captured by the DELOS reference model for DLs, user management, content management, functionality management, and policies are equally important in the definition of a DL. Accordingly, as demonstrated by real DL system experiences, VLDLs should be described as DLs featuring "very large features" in one or more of such aspects.

Models and measures for evaluating the "very-largeness" of the DL-axes content, functionality, users, and policies, are still an open issue. In this respect, the following observations and open questions naturally surface:

- What are the very large criteria for the DL-axes? How can such criteria be described and measured? What is their interrelationship?
- Should the definition of VLDLs be absolute or relative? If relative, for example, being "very large in content" should not depend on a given number, but on a number whose calculation depends on the limits of current technology (e.g. 20 times the average memory limit).
- Should the definition of VLDLs depend on challenge and complexity of design? If so, for example, DLs that can be built with existing solutions could not be in principle very large. The intuition is that "very large" equates to "unsolved because of inherent complexity" in one of the DL-axes.
- Should the definition of VLDL depend on federative aspects of DLs? In that sense, DLs would be very large whenever the user communities involved require the integration of a set of DLs at the level of one or more of the DL-axes.

The lesson learned from the workshop presentations and final discussion, is that VLDL research candidates to be an independent DL topic but still is in its early stage. It is to be investigated whether the foundations required for the consolidation of a research field per-se can be found in the common patterns and best practices of extant DL technologies or instead we still have to wait for more practical experience to come. These matters reveal a number of novel and interesting research avenues, from foundational to applicative, which will certainly be among the call topics of the Second Workshop on Very Large Digital Libraries next year.

4. PROGRAM COMMITTEE

Finally, our special gratitude goes to the members of the Program Committee: *Stefan Gradman* (Institut fr Bibliotheks und Informationswissenschaft, Humboldt-Universitt zu Berlin, Germany), *Kat Hagedorn* (OAIster System, University of Michigan Digital Library Production Service, USA), *Dean B. Krafft* (National Science Digital Library Project, Cornell Information Science, USA), Yosi Mass (IBM Research Division, Haifa Research Laboratory, University Campus, Haifa, Israel), *Yannis Ioannidis* (Department of Informatics, National and Kapodistrian University of Athens, Greece), *Peter Wittenburg* (Max-Planck-Institute for Psycholinguistics, The Netherlands), whose long research experience contributed in making this workshop an attractive and fruitful experience for all authors and participants.

Workshop proceedings [1] were printed by the DELOS Association for Digital Libraries [11].

5. REFERENCES

[1] Paolo Manghi, Pasquale Pagano and Pavel Zezula. Proceedings of the First Workshop on Very Large Digital Libraries, held in conjunction with ECDL 2008, Aarhus, Denmark, 2008

[2] Max-Planck Institute and FIZ Karlsruhe eSciDoc, funded by the German Federal Ministry of Education and Research (BMBF). http://www.escidoc-project.de

[3] Andreas Aschenbrenner Editing, analyzing, annotating, publishing: TextGrid takes, the a, b, c to D-Grid. In: iSGTW 30 January 2008, Jg. 54

[4] DRIVER: Digital Repository Infrastructure Vision for European Research. http://www.driver-community.eu

[5] D4Science: D4Science Project: DIstributed colLaboratories Infrastructure on Grid ENabled Technology 4 Science. http://www.d4science.eu

[6] DILIGENT: DILIGENT; a DIgital Library Infrastructure on GRID ENabled Technology. http://www.diligentproject.org

[7] EuropeanaLocal: Best Practice Network project, funded under the eContentplus programme. http://www.europeanalocal.eu

[8] CLARIN: Common Language Resources and Technology Infrastructure. http://www.clarin.eu

[9] EFG: European Film Gateway Project. http://www.europeanfilmgateway.eu

[10] BRICKS: Building Resources for Integrated Cultural Knowledge Services. http://www.brickscommunity.org

[11] DELOS: Digital library rEference modeL and interOperability Standards. http://www.delos.info

[12] Greenstone: Digital library Repository Software. http://www.greenstone.org

[13] Paper Past newspaper collection: http://paperspast.natlib.govt.nz

[14] Europeana: Connecting cultural heritage. http://www.europeana.eu

[15] SAPIR: Search In Audio Visual Content Using Peer-to-peer IR . http://www.sapir.eu