Second Workshop on Very Large Digital Libraries

In conjunction with the European Conference on Digital Libraries
Corfu, Greece, 2 October 2009

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

Pasquale Pagano
ISTI - Consiglio Nazionale delle Ricerche
Pisa, Italy
pasquale.pagano@isti.cnr.it

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

1. MOTIVATIONS

The mission of the international workshop series on Very Large Digital Libraries (VLDLs) is to provide researchers, practitioners and application developers with a forum fostering a constructive exchange among all key actors in the field of Very Large Digital Libraries. Its 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 first VLDL workshop [1] was consolidation of its mission. After several presentations on the papers accepted to the workshop and further demonstrations and long debates by all participants, there was eventual agreement that VLDLs deserve a chapter of their own in computer science research. Experience in the field since the historical beginning of DLs has proven that VLDLs cannot be simply regarded as very large databases storing Digital Library content, as one may be tempted to assess. In fact, as the DELOS Reference Model for Digital Libraries [4] well motivates, 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 magnitutde that requires specialized technologies.

VLDLs are clearly in their early stage of development. For example, interdependencies among the aspects above are still to be identified and studied; the same holds for models and measures for evaluating "very-largeness" of given DL systems. In fact, it is not even clear if there are clear enough patterns and best practices that are common in existing DL systems to determine the boundaries of the field or, instead, practical experience is still too much in its infancy.

The second VLDL workshop continued such investigations from where the first one left off. To this aim, its call for papers focused on foundational aspects of VLDLs and real experiences with them:

Foundational topics They covered definitional models and measures (content, functionality, users, and policies), architectural models, and design methodologies for VLDLs.

System topics They covered ideas, experiments, and practical experiences in system design and implementation. Of particular interests were the following: integration and federation of DLs, user management, security, sustainability, scalability, distribution, interoperability for content and functionalities, quality of service, storage & indexes, and preservation.

All contributions submitted were peer reviewed by two of the six members of the Program Committee and nine were accepted. The workshop structure comprised an invited speakers session followed by the presentation of the nine contributions, organized into three sessions: *systems*, *data management* and *functionalities* for VLDLs. Each session is analyzed in a separate subsection below.

2. WORKSHOP PRESENTATIONS

2.1 Invited talks

This session was dedicated to present the experiences of design and development of two major projects, funded by the European Commission, whose challenges are very large in terms of content heterogeneity and size:

Building Europeana v1.0: Towards a Large-Scale Content Ingestion: Julie Verleyen (Europena Office, Koninklijke Bibliotheek, National Library of the Netherlands) presented the data ingestion workflow adopted in Europeana [5], based on the Open Archival Information System (OAIS) [3]. The workflow safely rules the flow of data from content providers to the final receiver, i.e. , Europeana, by addressing the functional needs of updates, traceability, duplication, and conversion.

SAPIR: towards Large Scale Multimedia Content Search: Fausto Rabitti presented the advanced solutions proposed by the SAPIR project [6] (coordinated by Maristella Agosti and himself) to enable content-based searches of audio-visual information in the presence of very large collections of images and audio pieces. The solution exploits a P2P-based architecture for feature extraction, designed to scale and process very large collections of digital objects.

2.2 Session on VLDL systems

This session focused on the problems arising in hardware and software architectures of very large digital library systems.

Utility-based High Performance Digital Library Systems (by Hussein Suleman): Hussein Suleman presented an analysis of current high-performance systems and argued for the adoption of utility computing to tackle scalability issues that arise when dealing with large data collections under high quality-of-service requirements.

MultiMatch: Multiple Access to Cultural Heritage (by Giuseppe Amato, Franca Debole, Carol Peters, and Pasquale Savino: Franca Debole presented the main issues that surfaced while designing the MultiMatch system, enabling users to run searches over cultural heritage material across different media types and languages.

Integrating Multi-Dimensional Information Spaces (by Kostas Saidis and Alex Delis: Kostas Saidis defined very large digital libraries as those managing several large information spaces, not only in terms of data volume but also in terms of diversity of material and heterogeneity of sources they can support. Based on this, he presented a data management specification for digital libraries, enabling the construction of infrastructures for information space integration and interoperability.

2.3 Session on Data management in VLDLs

This session discussed experiences with dealing with different aspects of content management in the context of very large scenarios.

Improving Similarity Search in Face-Images Data (by Pedro Chambel and Fernanda Barbosa): Finding all faces in a large data set that are similar to a given desired face is computationally very expensive. Fernanda Barbosa proposed an approach to face recognition based on metric spaces, in which images are mapped to metric data structures in a metric space and similarity searches are reduced to Euclidean-distance range queries within that space.

Improving Query Results with Automatic Duplicate Detection (by Irina Astrova: A significant challenge in ontology-based data integration is that the process of data integration may lead to duplicate attributes, i.e., attributes that appear different but whose semantics is, in fact, the same. Irina Astrova presented a context-based approach for automatic duplicate detection in very large ontologies.

Enabling Content-Based Image Retrieval in Very Large Digital Libraries (by Paolo Bolettieri, Andrea Esuli, Fabrizio Falchi, Claudio Lucchese, Raffaele Perego, and Fausto Rabitti): The problems arising in the context of efficient Content-Based Image Retrieval (CBIR) are mainly concerned with the design of effective feature extraction algorithms. However, when the number of the images and/or the sizes of the individual images are very large, processing these images and building the corresponding indexes becomes prohibitively expensive. Fausto Rabitti presented the CoPhIR [9] collection (the largest available to the research community) of 106 million images and the GRID-based image crawling process used to extract the features required for CBIR.

2.4 Session on Functionality for VLDLs

This session attracted contributions related to functionality design and development affected by very large size.

Semantic Journal Mapping for Search Visualization in a Large Scale Article Digital Library (by Glen Newton, Alison Callahan, and Michel Dumontier: Glen Newton analyzed the scalability and utility of semantic mapping of journals in a large science, technology, and medical digital library, numbering 5,7+ Millions articles. The aim was to evaluate the effectiveness of semantic mappings of articles for query result refinement and visual contextualization in very large digital libraries.

Exploiting Individual Users and User Groups Interaction Features: Methodology and Infrastructure Design (by Emanuele Di Buccio and Massimo Melucci: Information gathered by monitoring interactions between users (or group of related users) and a DL system can be used as an indicator of individual user interests. Emanuele Di Buccio proposed a methodology based on a Peer-To-Peer infrastructure for collecting and exploiting behaviors of both individual users and user groups as a source of evidence of their interests.

Maintaining Object Authenticity in Very Large Digital Libraries (by Tobias Blanke, Stephen Grace, Mark Hedges, Gareth Knight, and Shrija Rajbhandari: DLs are increasingly used to manage large volumes of research data. Such data is irreplaceable and their management calls for scalable methodologies for its long-term access and curation. Mark Hedges presented the case study of an iRODS-based architecture [8] for automatic support of rule-based authenticity models in very large digital libraries.

3. WORKSHOP DISCUSSION

The final, brainstorming session of the workshop started by everyone agreeing that "very large" issues in Digital Libraries are concerned with four axes: user management, content management, functionality management, and policy management. For example, a Digital Library may be "very large" in terms of the amount of data it has to manage as well as in terms of the number of communities of end users it has to serve at the same time.

The discussion focused largely on *scalability*. In the database world, a database is very large when it can scale up beyond a given threshold of content (e.g., greater than 1 Terabyte of data). Can we draw a parallel between DLs and databases and claim that a DL is very large when it can scale up beyond a certain size in the four axes mentioned above? In other words, are DLs very large because of well-established *size* thresholds, be them related to users, policies, content, or functionalities? Furthermore, what are these thresholds and how are they to be identified? Analysis of known VLDL scenarios, e.g., those of e-Infrastructures such as Europeana and DRIVER, seemed to answer this question in the negative: DL scalability appears to depend on size as well as on at least two other important interrelated aspects: *sustainability* and *interoperability*.

In the DL world, scalability cannot be measured in an absolute sense as in the database world, where size determines whether or not a database is labeled as very large. In fact,

the adjective "very large" strongly depends on sustainability issues, often ruling the DL world, where funds are generally scarce and hard to guarantee in depth of time. As a consequence, the adjective "very large" may label systems where the problem to be tackled might not look "that large" in other domains. For example, adoption of GRID infrastructures or high performance computing solutions used by the physics community to tackle very large datasets are generally not a reasonable solution in a typical DL application scenario, given the current dominant business models. As a consequence, new VLDL research goals arise, such as developing sustainable and low-cost hardware and software infrastructures or devising alternative business models.

The second key system property for VLDLs is interoperability with respect to content and functionality – whose foundations are the subject of several intense studies recently, e.g., in the context of the DL.org project [7]. For example, Europeana and DRIVER infrastructures aggregate and integrate tens of millions of metadata records from different data sources. This volume of data, however, does not introduce scalability issues, as it can be effectively dealt with standard storage and indexing technology. To the contrary, such systems are not very large in terms of content management, but rather in terms of the arbitrary amount of distributed data sources they have to federate and the heterogeneity of technology, data structure, and data semantics they have to cope with to accomplish this task. Similar reasoning can be applied to other categories of content to be integrated, such as ontologies, policies, user profiles, and others, where diversity of information representation and forms of information exchange are the real "very large" issues.

4. CONCLUSIONS

The main conclusion drawn from all workshop deliberations was that Very Large Digital Libraries research seem to focus on scalability challenges that Digital Library systems manifest in terms of content, user, policy, and functionality management. Furthermore, it is not only size that is a key issue but also levels of sustainability and interoperability. The need for research in these areas is great, and any progress in these directions will be further investigated in the next edition of VLDL.

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, Leonardo Candela, and Costantino Thanos for their research inspiration, and Vittore Casarosa and Francesca Borri for the logistic support.

Special thanks are also due to the members of the program committee, *Stavros Christodoulakis* (Technical University of Crete (MUSIC/TUC), Greece) *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), *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.

Finally, our sincere gratitude goes to the chairs of the sessions, who offered their expertise in coordinatin and harmonizing the sequence of presentations, and of course, to all authors, whose passion and ideas were the real fuel of VLDL 2009.

Workshop proceedings [2] were printed by the DELOS Association for Digital Libraries [4].

6. 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] Yannis Ioannidis, Paolo Manghi, Pasquale Pagano. Proceedings of the Second Workshop on Very Large Digital Libraries, held in conjunction with ECDL 2009, Corfu, Greece, 2009

[3] OAIS: Open Archival Information System. http://public.ccsds.org/publications/archive/650x0b1.pdf

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

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

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

[7] DL.org Project: http://www.dlorg.eu

[8] iRods Project: Integrated Rule-Oriented Data System http://www.irods.org

[9] CoPhIR Image Collection: Content-based Photo Image Retrieval Test-Collection http://cophir.isti.cnr.it