Int J Digit Libr
DOI 10.1007/s00799-008-0042-0

EDITORIAL

Scholarly digital libraries at scale: introduction to the special issue on very large digital libraries

Min-Yen Kan · Dongwon Lee · Ee-Peng Lim

When people are asked to think of very large information sources, these days the answer most commonly heard is the World Wide Web. Yet the Web does not answer charges to organize, validate and curate its information to yield knowledge useful for an information seeker.

Managers of library systems, who have accepted these charges, have grown to meet this demand, scaling up library management systems to the very large. Very large digital libraries must maintain their quality of service in organizing and curating their data, as well as deal with the myriad technical issues that face large scale systems.

Recognizing the growing importance of these challenges, this issue of the International Journal of Digital Libraries gathers seven articles that showcase fielded solutions to these technical and organizational challenges, comprising of both invited submissions from well-established digital library centers and rigorously peer-reviewed submissions.

Architectural foundations are necessary for building the digital library at a large scale. Standards developed in our digital library community such as Dubline Core, OpenURL, OAI-PMH, DOI and ARC have successfully addressed metadata standardization, appropriate access, object discovery, ubiquitous identifier and serial storage issues, respectively. Putting these component technologies together requires a framework at a higher level. The first part of this issue brings together competing and complementary frameworks of building, growing and maintaining digital libraries at scale using these proven components.

van de Sompel et al. (this issue) lay down the principles of the aDORe three-tier architecture, which were designed with minimality and compatibility in mind. Implementing this protocol is easy, and paves the way for distributed and robust digital object discovery and retrieval. The authors show that the architecture can be fielded and used in diverse situations by detailing two very different use scenarios at Los Alamos National Labs and the University of Ghent. The benefits of a minimalist approach are clear, enabling off-the-shelf commodity hardware to achieve highly-scalable searching and retrieval performance of over 40 million unique items.

Ioannis et al. (this issue) tackle the same problem from a different perspective, in their article on the service-oriented architecture of the DELOS Network of Excellence. In the work on the DELOS digital library management system (DelosDLMS), the unification of existing services is the central focus. By composing different component services together, the DelosDLMS enables quick deployment of new digital library interfaces and collections. A distributed, well-tested, peer-to-peer middleware underlies many DELOS deployments and forms the basis for more advanced services in the next generation digital library: query expansion, content-based search of multimedia items and novel user interfaces.

Which architecture should the practioner pursue? The answer depends on how their audience and maintainers view their digital assets and services. It is important to remember that implementation of one framework does not preclude using or supporting another; and several alternative protocols may be instrumented to provide both object-level or service-level guarantees.

Irrespective of this dichotomy, services are run and objects are stored on machines. In today's distributed environment,

M.-Y. Kan
National University of Singapore, Singapore, Singapore

D. Lee (✉)
The Pennsylvania State University, University Park, USA
e-mail: dongwon@psu.edu

E.-P. Lim
Nanyang Technological University, Singapore, Singapore

load balancing is a critical issue that must be properly handled to ensure data reliablity and service replication. Sulemain et al. (this issue) perform a case study of how this problem can be addressed within context of a service-oriented framework, the Open Digital Library (ODL) framework. They show how hundreds requests can be redistributed and balanced across a digital library network to facilitate optimum service. Such load balancing algorithms are an essential part of the robust digital library and can be integrated within the DelosDLMS or aDORe architectures.

Aside from needing a scalable architectures in a technological sense, libraries also have problems with scale in a very human sense. Such large digital libraries are often collaborative, involving many coordinating parties that have overlapping but distinctly different agendas. Such organizational challenges can bog down and deter progress on technical fronts, especially when intra-project groups have not properly negotiated standards for metadata quality and data exchange.

The European Library (TEL) is a project at such a scale which has been overwhelmingly successful at handling these challenges. A collaboration involving many European national libraries, the Conference of European National Libraries (CENL) established working groups for coordinating TEL and other cross-library initiatives. Cousins et al. (this issue) details how CENL's working groups allow all contributers to have a say in development and puts each team's expertise to use in the appropriate segment of development. Large-scale collaborative digital library projects can adapt such an organizational framework for their development, marketing and management.

Li et al. (this issue) examine the scientific digital library, CiteSeer, which is architected around web-based requests. Like many digital libraries, CiteSeer responds to many different types of requests (in CiteSeer's case over 40 distinct message types), each contributing differently to the overall workload of the system. Li and colleagues describe how they analyzed the system's query logs in terms of temporal and geographic scope of requests. While they conclude most requests to CiteSeer are to download freely-available copies of scientific papers, the authors note a growing proportion of the logs indicate that users also turn to CiteSeer for the purposes of knowledge discovery, including conducting surveys of work on a particular area of computer science.

Supporting knowledge discovery and online usage of papers is one future pathway of digital libraries. The area which is most developed in this regard is the work on scholarly texts in the form of scientific records. Nanba et al. (this issue) examine the context in which a retrieval environment for scientific articles can be used to effectively retrieve related work on patents. They exploit natural language cues to locate, analyze and link and citations across the two genres within a supervised machine learning framework. Richardson et al. (this issue) explore how automatically-generated concept maps can be utilized to provide a high-level summary of theses written in English. While the requisite natural language processing to generate these maps is fragile and computationally intensive, the authors illustrate how such an application is already a useful gisting service for non-native English speakers—the generated concept maps are terse and far easier to translate than the full text. In these works, we are given the glimpse of what is yet to come—in which different document genres or different languages are brought together, by leveraging regular document structures.

In organizing this special issue, we planned to obtain balanced coverage from both the academic and industrial perspectives. While we were successful at finding representation from academia, industry proved to be more challenging. While colleagues in industry were agreeable towards having a balanced perspective in the issue, they were understandably reluctant to expose the technical details of their approaches in solving scalability problems and issues. In this regard, open access to scholarly information may be more easily surmounted than open access to corporate infrastructures in managing large scale content. Despite this dearth of industrial participation, we feel that this issue compiles valuable advice on solutions to distributing, serving, balancing and managing large-scale multimedia collections.

From the perspectives offered by the articles in this issue, one gathers that the future of very large digital libraries lies along two paths: towards a standards-driven, generic and scalable library infrastructure supporting preservation, discovery and distribution of data and knowledge; and towards the specialized niche digital library, serving specialized and storage-intensive contents for specialized purposes.

We hope you enjoy reading about digital libraries at the very large scale, and wish that this issue will inspire the next generation of digital librarians and stewards to think on the pressing issues in the future of large scale digital libraries.