Digital Library Grid Scenarios

Michal Kosiedowski, Cezary Mazurek, Marcin Werla

Poznan Supercomputing and Networking Center, ul. Noskowskiego 10, 61-704 Poznan, Poland
{kat,mazurek,mwerla}@man.poznan.pl

Abstract: Grid technologies that have recently emerged around the world enable the integration of distributed computing resources. These technologies allow creating distributed grid services to hide the complexity of the underlying resource network and forming virtual organizations, which is the central concept of grid computing philosophy. Grid researchers and developers gather together under the auspices of the Global Grid Forum to implement common applications of grid services through definition of open standards. We believe that grid research should also cover incorporation of digital libraries concepts and functionalities to work together with other components in complex grid applications. This is why, based on our experience gained from numerous grid and digital library projects, we list some grid opportunities involving digital libraries and draw paths to follow to enable digital libraries to the grid, grid applications and grid users.

Poznan Supercomputing and Networking Center (PSNC) has been involved in digital library and grid research for a few years now. The digital library system dLibra [1], which has been developed in Poznan since 1999 and was presented on several occasions during international conferences. The system has been deployed to serve the regional WBC Digital Library (Wielkopolska Biblioteka Cyfrowa) in 2002 [2] and is planned for new deployments of regional digital libraries in Poland in the near future. As far as grid works are concerned PSNC has been a part of such major initiatives as GridLab [3] and Crossgrid [4] EU funded projects, and the PROGRESS project [5,6], which was co-funded by the Polish State Commitee for Scientific Research. The latter grant aimed to deliver an architecture and a testbed implementation of an access environment to grid resources and services. The PROGRESS project delivers also a Data Management System [7,8], which is supposed to act as a grid storage for scientific data, and is primarily used to assist the computing experiments performed in a computing grid.

While we actively undertake research problems within respective domains of digital libraries and the grid, we also notice the need to take a complex look at enabling the grid to digital libraries and digital libraries to the grid. We believe that this integration should primarily include incorporation of digital libraries into grid-enabled information retrieval systems as digital libraries are systems that are built to deal with information. Exposing digital library functionality as information retrieval grid services can enable multiple complex grid scenarios and applications.

An information retrieval grid can be understood as information retrieval (IR) performed on the grid. IR research deals with identifying documents or sub-documents that meet information needs expressed in the form of queries. The grid, on the other hand, offers new techniques that enable distributed accomplishment of computational tasks on a set of computers connected by a network. Exposing IR systems components as grid services can enable federations of information collections, additionally tuning the participating services for optimal performance. The grid security models can also enable access control of information and document publication.

The task of defining the IR Grid standards has been recently undertaken by a new Working Group [9] formed within the structures of the Global Grid Forum [10]. As described in [11] the GridIR architecture aims to bring the following benefits of the grid to IR systems: the use of divide and conquer approaches to collections, indexing and querying, thus allowing IR on larger collections; the ability to tune collections for better retrieval; event triggered re-indexing via push or pull architectures; security at the collection, query and document level; mechanisms for collecting, weighing and ranking results from different sources: multiple algorithms or IR approaches might be applied to the same data set, and then provide merging and ranking methods to determine the overall ranking from the merged results. On the other hand, with GridIR the grid can benefit from IR systems being provided with: the

embedded and integrated indexing of content which are agnostic about content type, markup, size and location; the publication of local documents or other collections in indexed and searchable formats; a framework for updating collections by push or pull mechanisms; a resource discovery infrastructure for changeable index-able content.

We believe that the information retrieval grid, such as the one proposed by GridIR, must also include digital libraries and their functionality. Such integration can bring various digital library systems within one common framework, thus allowing users to take advantage of multiple document collections while retrieving required information. It can also add new quality to the complex grid applications enriching the opportunities with digital library functionality and contents.

There is a tendency observed that with the size of the data gathered in DL systems growing, there also grows the need for better and simpler access to that data. Nowadays it is not enough to provide users with a good-looking search interface to a DL system. As described in [12], users require a tool that can give access to publications collected in many libraries at once. What is also important, users do not require knowledge on locations or addresses of all these libraries; these should be hidden. Such a multiple-DL search tool should be able to query many DL systems and return search results to users as a single search result list.

Such a distributed DL search system can be created by allowing DL systems to communicate with each other. It is not so difficult in case of communication between multiple instances of the same DL system since they feature the same access interfaces and use the same communication protocols, query languages etc. But the problem becomes much more complicated, when two different digital libraries being instances of two different DL systems are supposed to communicate with each other. Different DL systems use different communication protocols or different query languages. Determining, what DL instances should be examined with a given query and effective presentation of search results, obtained from multiple distributed systems, is also problematic. Adjusting DL systems, so they can be a part of an IR grid, like for example GridIR, is an approach that may solve most of the above listed problems.

But not only the DL can benefit from going grid, also the grid and the users of the grid can benefit from integration with DL systems. These benefits come when DL functionality, exposed as IR grid services, becomes accessible and usable in complex grid applications and scenarios.

The most known type of the grid is a computing grid. Such a grid features multiple distributed computing resources managed by a grid management system, for example Globus Toolkit, sometimes with resource allocation managers applied upon it. Such a grid can be assisted by several higher-level grid services, like for example those grouped within the Grid Service Provider (GSP) module designed by the PROGRESS project [13]. Additionally, some grid data storage services are usually required. These are provided by special systems, like for example the Data Management System (DMS) initially proposed by the PROGRESS project, that are capable of providing distributed storage to store large amounts of scientific data, allowing seamless access to the data and enabling a high-level management of the grid data pieces. Such computing and data grid services work together in various grid scenarios that can also include assigning job tasks into workflows. We believe that these scenarios can expand to cover more complex applications when given an opportunity to utilize information retrieval, and digital library in particular, grid services.

Let us imagine a medical team that faces a strange illness striking one of their patients. It seems like this is a virus, but that virus is unknown. Having the probes of the virus the doctors can turn this into a digital form of parts of a DNA sequence. Now a proper piece of software comes into place to help and find the structure of the virus through assembling the whole DNA sequence of the virus. This result sequence can now be compared to known virus sequence, helping to find similar or maybe even

exactly the same sequences and thus viruses. Knowing the name of the virus, or at least the type of the virus, the doctors can look over a library to find some literature on how to beat such a virus. Then, they can hopefully cure the patient.

The whole above mentioned scenario can be performed with the use of traditional tools, including digital libraries to search for the required literature. Yet, such a process takes time and/or people resources as the data must be transferred from one tool to another manually, and to have no time loss between each stage the team must commit one person to incessant monitoring of the analysis process. However, the whole scenario can be closed into a grid workflow that can be executed without any assistance from the submitting user. The grid can simply notify the doctors that the results of their search are already there.

The computing experiments and research using the grid-enabled applications and grid computing resources is a problem that was resolved to a much extent by the grid community. This includes also comparison services to compare various pieces of scientific data, be it a specially designed grid application or a service featured by the grid data storage system itself. A new addition to this infrastructure can be the IR grid services featuring the digital library functionality. Integration of the IR grid services with existing DL systems can provide an opportunity to seamlessly add DL functionality to enrich grid scenarios and opportunities. A scenario discussed in this subsection is an example one that we have in mind while continuing our research in the field of grid and digital libraries. There are, probably, many more similar uses of the DL grid services within complex grid applications and in other configurations than the one described above.

While discussing the grid opportunities for digital libraries we also notice that a new type of publications comes to play with the grid on hand. If a DL system is a part of the IR grid, the publication metadata can contain also associations to other grid resources and services. These can be associations to other IR grid services (and thus other digital libraries), but they can also be associations to computational services. Such an approach gives a possibility to build a quasi-intelligent user interface for DL systems connected to IR grids. That interface should be able to read and interpret the special metadata connected with a given part of a publication. After interpreting the metadata, the interface should give users a possibility to utilize the grid services associated with the browsed grid publication.

The management of grid publications raises interesting research problems, such as designing a grid publication description language, designing patterns and best practices concerning building a flexible and intuitive user interface to create, manage and access grid publications. And, as we emphasized in this paper, the IR grid researchers need to come out with standards to allow the integration of digital libraries and other sources of information and publication content. We certainly want to commit ourselves to finding solutions to enable grid publications to the wide community of users.

References

1. dLibra, http://dlibra.psnc.pl/
2. WBC Digital Library, http://www.wbc.poznan.pl/
3. GridLab project, http://www.gridlab.org/
4. Crossgrid project, http://www.crossgrid.org/
5. PROGRESS project, http://progress.psnc.pl/
6. Kosiedowski, M., Mazurek, C., Stroinski, M.: PROGRESS – Access Environment to Computational Services Performed by Cluster of Sun Systems. Proceedings of the 2nd Cracow Grid Workshop, Cracow Poland (2002) 45-56
7. Grzybowski, P., Kosiedowski, M., and Mazurek, C.: Web Services Communication within the PROGRESS Grid-Portal Environment. Proceedings of the International Conference on Web Services ICWS 2003, Las Vegas, USA (2003) 340-345

8. Grzybowski, P., Mazurek, C., Spychala, P., Wolski, M.: Data Management System for grid and portal services. Accessed from http://progress.psnc.pl/
9. Grid Information Retrieval Working Group, http://www.gridir.org/
10. Global Grid Forum, http://www.ggf.org/
11. Dovey, M. J., Gamiel, K.: GRID IR — GRID Information Retrieval. Poster at EuroWeb 2002. Accessed from http://www.gridir.org/
12. Soergel D.: A Framework for Digital Library Research. D-Lib Magazine, Volume 8, Number 12 (2002)
13. Bogdański, M., Kosiedowski, M., Mazurek, C. and Wolniewicz, M.: GRID SERVICE PROVIDER: How to improve flexibility of grid user interfaces?. Lecture Notes in Computer Science 2657: Proceedings of the International Conference on Computing Science ICCS 2003, Springer-Verlag, St. Petersburg, Russia (2003) 255-263