Research on the Interoperability Architecture of the Digital Library Grid

Hao Pan
Department of information management , Beijing Institute of Petrochemical Technology, China, 102600
bjpanhao@163.com

Abstract. With the rapid development of the digital libraries, Interoperability is the problem to realize the real information sharing and eliminate the digital information islands. In this paper, the character and requirement of digital library interoperability is discussed. As the limitations of the current technologies in the realization of the large-scale digital library interoperability on the Web, the architecture of the digital library interoperability is put forward based on the Grid technology, and the working process of the digital library Grid is presented. Using the Grid technologies, the high performance OAI federated search service is realized.

1 Introduction

With the quick development of the information technology, people's working, learning and life style have been greatly influenced by the Internet. At the same time, library information enterprise has come to a new phase-DL (digital library), after being through traditional library and automatic library phase. A DL can discover, manage and publish digital information in an effective way in order to provide a simple and convenient digital information querying platform for groups of individuals. Many countries have developed their own digital library one after the other, such as NSDL [1], NCSTRL [2], DLI2 [3]. As different libraries take different technology, platform, protocol and architecture according to its different objective and running manner, it is an impending problem in the field of the DL interoperations that how to integrate these libraries with different infrastructure and technology in order to exchange information and share documents between different DLs, finally to build a big virtual DL, to provide uniform and convenient services for global users.

Present solutions to the DL interoperability, such as distributed search technology [4], metadata collection technology [5], middleware technology, cannot solve the problem of large scale DL interoperability on the Internet in an appropriate

Please use the following format when citing this chapter:

Pan, H., 2007, in IFIP International Federation for Information Processing, Volume 251, Integration and Innovation Orient to E-Society Volume1, Wang, W. (Eds), (Boston: Springer), pp. 147-154.

way [6]. In order to have real DL interoperability, we need to solve the problem from the beginning of the DL architecture design and implementation. This article makes it by offering architecture of the DL interoperability based on Grid technology, which uses the character of resource sharing of the Grid technology to support OAI service.

2 The Outline of the DL Interoperability

2.1 The Character of the DL Interoperability

DL is a network based magnanimous database, which has massive and different types of information such as text, images, video, audio, automation, etc. The characters of a DL are as follows:

1. Heterogeneity. Platform heterogeneity, information source heterogeneity and metadata heterogeneity embody the heterogeneity of a DL. The electronic information resources in a DL are always heterogeneous in the form of resource storage format, service mode offering, access interface and data transport protocol.
2. The physical distribution of the digital resource. The resource of a DL is physically distributed in the network. It is a great challenge to provide service for the user over the tremendous digital resources of the DLs in the WEB.
3. Autonomy. Almost all the digital libraries distributed over the Internet are autonomous information systems. It means they manage and maintain their own resources and services, They also manage their access method and authorization independently.
4. Openness. You can enter or logout an interoperated DL at any time.

2.2 The requirement of the DL interoperability based on metadata harvesting

Many DL projects use distributed searching method. It unites many DLs to provide a single service interface for the users, and thus to implement the DL interoperability. However, this method is only suitable for a DL which has less nodes (e.g. number of nodes <20) because it uses real time searching method. When there are so many nodes (like more than 100), it becomes very difficult to perform a mass distributed searching over the Internet. The good thing is, metadata harvesting method can solve this problem very well. This method is based on OAI/PMH [7] infrastructure. It can offer a uniform querying service for the users by transforming all metadata from different DLs to one format and store them into the metadata depository.

The OAI-PMH offers a technology to support the organizational evolution of digital cultural content creation and collective services to access cross-domain collections. It is an attempt to build a "low-barrier interoperability framework" for archives containing digital content. It allows Service Providers to harvest metadata from Data Providers. The Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) was created to facilitate discovery of distributed resources. The OAI-PMH achieves this by providing a simple, yet powerful framework for metadata harvesting. Harvesters can incrementally gather records contained in OAI-PMH repositories and use them to create services covering the content of several

repositories. The OAI-PMH has been widely accepted, and until recently, it has mainly been applied to make Dublin Core metadata about scholarly objects contained in distributed repositories searchable through a single user interface.Figure gives the metadata interoperability architecture based on OAI-PMH.

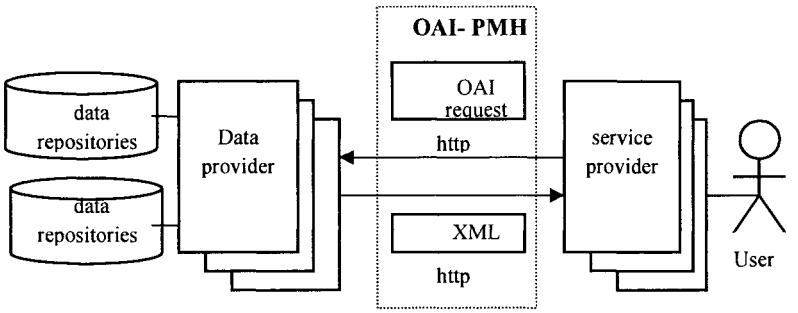

Fig.1. The metadata interoperability architecture based on OAI-PMH

Primarily, there are two kinds of roles in OAI infrastructure: data provider and service provider. Data provider is to provide and distribute the metadata while service provider collects the metadata from multiple data providers and provides querying service to the end users. Service provider communicates with data provider via OAI-PMH protocol. OAI-PMH protocol, whose metadata uses XML format, is based on HTTP protocol to transport data, which makes it widely used in the Web environment and enables information exchange between distributed heterogeneous resources.

In order to effectively implement the metadata discovery, harvesting and indexing service, appropriate architecture to support these operations and high performance servers are required. At the present, all the metadata discovery,, harvesting and indexing jobs are centralized on one or several given servers, which makes the performance and the reliability extremely poor [8]. As the scale of the DLs interoperability enlarges continuously, it will get harder to expand the solution. In this paper, Grid technology is introduced to enhance the reliability and the scalability of the service and the performance of the metadata harvesting and indexing service.

3 The Grid technology

Grid computing is a solution for resource sharing and problem solving in the dynamic and heterogeneous virtual organization, which made the Internet into a tremendous super computer to realize the computing resources, storage resources, data resources, information resources, knowledge resources, equipment resources, etc [9]. The basic character of Grid computing is resource sharing .

3.1 Open Grid service architecture

OGSA is the most widely used Grid architecture at present. The need for integration and interoperability has led to the design of the open Grid service architecture, which offers an extensible set of services that virtual organizations can aggregate in various ways [10]. OGSA defines the Grid service, which aligns Grid technologies with Web Services technologies to take advantage of important Web Services properties such as service description and discovery, automatic generation of client and service code from service description, compatibility with emerging higher level open standards and tools, and broad commercial support.

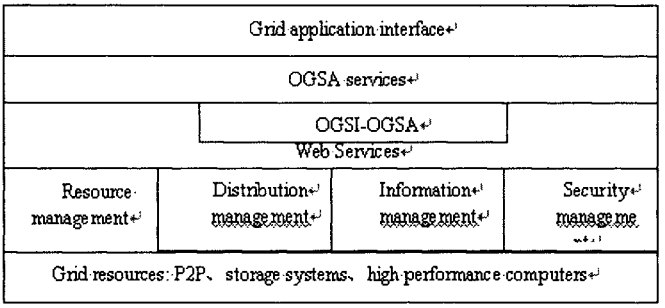

Fig.2. open grid service architecture

In OGSA, every resource is represented by a Grid service, which defines standard mechanisms for creating, naming, and discovering persistent and transient Grid service instances [11]. It provides location transparency and multiple protocol bindings for service instances and supports integration with underlying native platform facilities, which can solve the problem of interoperability among the heterogeneous grid facilities and provides the virtual service. Fig.2 describes the OGSA, the digital libraries interoperability can be realized by using the service of resource discovery, resource distribution, and resource accessory provided by OGSA.

3.2 The advantage of the Grid digital libraries

The Grid technology which supports the large scale resource sharing has the advantage in the respect of heterogeneous platform compatibility and system harvesting. The DL interoperability and the Grid computing have many similar characters, such as distribution, heterogeneity, autonomy. Therefore, the Grid technology can be used to solve the problem in the DL interoperability.

1. The heterogeneity of the resource. Two aspects of the resource heterogeneity are solved in the Grid technology, firstly, the heterogeneity of different resources, such as the computing resources, the storage resources; secondly, the heterogeneity of the structure of the same resources. The DLs interoperability can solve the resource heterogeneity using the Grid technology.
2. The sharing of the resources. The aim of the Grid computing is to realize the

resource sharing, and eliminate the information Isolated island and the knowledge Isolated island [12]. Using the Grid technology, the DL interoperability can be enhanced, and the resource sharing can be realized.

3. The autonomy of the resource. The resources in the Grid belong to different organizations. The resource owners can manage and control their resources absolutely. The Grid systems cannot control these resources. The DLs interoperability also requires solving the autonomy and sharing of the resources.

4. The individuality of the service. The Grid provides the intellectual and individual services for the users. The DL also provides the individual services for the users, and realizes the information formulation and the initiative pushing of the information.

4 The architecture of Grid digital library

4.1 The architecture of digital library based on the Grid technology

To solve the interoperability of the digital library based on the metadata level and the service level, the architecture of DLs interoperability based on the Grid technology is proposed, Fig.3 gives the architecture.

The architecture of Grid digital library contains 3 layers. The base layer is the data layer, which is made up of the DLs distributed on the different regions, as the data provider of the DLs interoperability, the data layer provides the metadata that data format is decided by OAI-PMH protocol to form the Metadata Repository based on the OAI-PMH protocol; the top layer is the application layer of the virtual digital libraries, which provides the united service interface, and provides the data retrieval service and data query service; the middle layer between the data layer and the application layer is the Grid layer, which shields the distribution and the heterogeneity of the base layer, and provide the united service interface for the application layer by the way of metadata discovery, harvesting and indexing.

4.2 The principle of the digital library Grid

The digital library Grid realizes the shielding of the distribution and heterogeneity of DL and the metadata discovery, harvesting and indexing of the digital libraries distributed on the different places, and provides the generous information service for the users. There are three layers in the architecture of DL grid, the application layer, the grid layer and the data layer. The application layer is to provide the generally query interface to the readers, the readers start a query not knowing how and where to get the data. The data layer contains the heterogeneous DL, every DL in the data layer realize the function of the DP of the metadata. In the grid layer, the metadata is collected by the metadata harvesting point to the metadata generally harvesting point, which provides the unified metadata to the application layer.

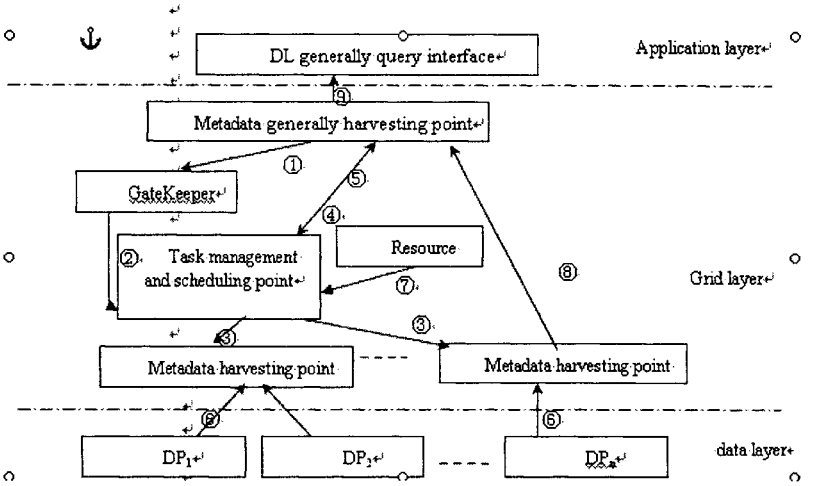

Fig.3. the interoperability architecture of digital library Grid

The working steps of the digital library Grid in the Fig.3 is as follows.

1. The metadata generally harvesting point initiates the harvesting duties, the DL Grid initiates the Gatekeeper procedure of providing working assignment.

2. The Gatekeeper procedure builds the duties according to the duty requiring, the task and resource distribution is carried on by the task management and scheduling point.

3. The task management and scheduling point carries on the metadata harvesting jobs.

4. The task management and scheduling point provides the condition information of the task for the metadata generally harvesting point.

5. The metadata generally harvesting point submits the request of duty canceling.

6. The DL distributed on different regions provides their own metadata for the metadata harvesting point.

7. In the procedure of the metadata harvesting, if there is breakdown on one harvesting point, the duty management and scheduling point will transmit its harvesting code to other points using Grid FTP to guarantee the advancement of the metadata harvesting.

8. After the metadata harvesting works, the metadata harvesting point sends the metadata to the metadata generally harvesting point, and the corresponding metadata is store on the metadata repository.

9. After the metadata generally harvesting works, the metadata is stored on the metadata repository, the generally query interface can be provided for the user, which send the request to the corresponding metadata repository, and collect the query results, and submit the results to the user.

5 Conclusions

The aim of the digital library interoperability is to unite the isolating DLs over the Internet, to provide the united and virtual information provider to the user, and to realize the information resource sharing of the global digital libraries. According to the requirement of the large scale digital library interoperability and the advantage of the Grid technology on the aspect of resource sharing, the architecture of the digital libraries interoperability is proposed, which unites the Grid technology and the OAI/PMH protocol, realizes the high performance metadata discovery, harvesting and indexing. As the Grid can make fully use of the idle resources over the Internet, the cost of the digital library interoperability can be greatly reduced. The digital library interoperability can also realize the metadata redundancy and enhance the reliability of the service using the idle resource over the Internet.

References

1. C. Lagcne and W. Hoehn, Core services in the architecture of the national digital library for science education (NSDL), *Proceedings of the Second ACM/IEEE Joint Conference on Digital Libraries* 201~209(2002).
2. B.M. Leiner, The NCSTRL approach to open architecture for the confederated digital library,
http://www.dlib.org/ dlib/december98/ leiner/121einer.html.[2003-02-05].
3. Digital Libraries Initiative Phase2, http://www.dli2.nsf.gov/. [2005-12-12].
4. B. Matthias, M. Sebastian, W. Gerhard and Z. Christian, Challenges of Distributed Search Across Digital Libraries.
http://www.mpi-inf.mpg.de/~czimmer/papers/delos05.pdf. [2005-04]
5. M. Zhang, D.Q. Yang and S. A. Wang, An Architecture Supporting Metadata and Service Interoperability in Digital Libraries, *Computer Science*, 22(4), 482-487(2004).
6. C. Leonardo, C. Donatella and P. Pasquale, A service for supporting virtual views of large heterogeneous digital libraries, *7th European Digital Library Conference*, 362-373(2003).
7. S. Hussein, Introduction to the open archives initiative protocol for metadata harvesting, *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital libraries* (2002).
8. D. David and F. Muriel, Open Archives Initiative – Protocol for Metadata Harvesting Practices of cultural heritage actors.
http://www.oaforum.org/ otherfiles/oaf_d48_cser3_foullonneau.pdf.[2003-9].
9. F. Ian, K. Carl and T. Steven, "The anatomy of the grid: enabling scalable virtual organizations", *International Journal of Supercomputer Applications*, 15(3), 200-222(2001).
10. F. Ian and K. Carl, et al, The physiology of the grid: an open grid services architecture for distributed systems integration, open grid service infrastructure WG.
http: //www. Gridforum.org/ogsi-wg/drafts/ogsa-draft.pdf.htm. [2004-12-9].

11. F. Ian, "Grid computing: concepts, applications, and technologies". http://www.mcs.anl.gov/~foster.htm. [2004-12-18].
12.B. Fran,F. Geoffrey and H. Tony, *Grid Computing: Making the Global Infrastructure a Reality* (Wiley & Sons Ltd, 2003).