STORIES

D-Lib Magazine
May 1998

ISSN 1082-9873

National HPCC Software Exchange (NHSE)[*]

Uniting the High Performance Computing and Communications Community

Shirley Browne, Jack Dongarra, Jeff Horner, Paul McMahan, Scott Wells
University of Tennessee
Knoxville, Tennessee 37996
swells@cs.utk.edu

[*] The NHSE is supported by grant NAG 5-2736 from the National Aeronautics and Space Administration (NASA)

Abstract

The National HPCC Software Exchange (NHSE) is a Web-based resource for promoting software sharing and reuse within and between the High Performance Computing and Communications (HPCC) community including the HPCC agencies. The NHSE also facilitates the development of discipline-oriented, software repositories while promoting contributions to, and use of, such repositories by the HPCC. As part of the sharing process, the NHSE provides information about software and software-related artifacts, which include algorithms, specifications, designs, and software documentation. This paper describes the methods being utilized to achieve such promotion and sharing of software. The NHSE's goal is to decrease the development time of high performance applications while simultaneously reducing software duplication as well as time spent locating relevant software information.

Introduction

Begun in 1994, the National HPCC Software Exchange (NHSE) is being developed through funding by several High Performance Computing and Communications (HPCC) community agencies. The goals of promoting, sharing, and distributing software and related resources are currently being achieved via several projects sponsored by the NHSE. By developing Web-based interoperable, domain-specific repositories, the NHSE provides distributed access to high performance software. Such virtual repositories lessen the burden of locating relevant software in a timely matter. Discipline-oriented repositories maintained by the NHSE include repositories for parallel systems and tools as well as high performance mathematical subroutine libraries.

To facilitate repository creation and maintenance, the NHSE has developed a toolkit for building interoperable, distributed repositories. This toolkit, Repository in a Box (RIB), allows for dynamic sharing of metadata that describes software or any other object suitable for sharing.

Further value is added to the repositories by the NHSE through the use of software reviews, which are performed by domain experts to ensure that software within the repositories adhere to certain standards as well as ensuring relevancy and usefulness to the user. In addition to review information appearing in the repositories, the reviews are published in the WWW-based electronic journal, *NHSE Review*™.

As part of NHSE's ongoing efforts to meet the needs of the HPCC community, additional projects are under development including an access control and on-line licensing system for software distributed from repositories.

1. Repository in a Box (RIB)

Repository In a Box (RIB) is a freely available software package for creating WWW accessible software repositories. These repositories consist of descriptions of software packages and, optimally, the files which constitute those packages. RIB provides an HTML-based administrative interface for creating and maintaining repositories. In many cases, the repository maintainer may choose to only maintain the description of a software package and have the URL for downloading the package point outside of his or her repository. Doing this allows software packages to be included in a repository that cannot or should not necessarily be distributed from the repository. If the repository maintainer chooses to upload the software package itself into the repository when the description is added, then he or she has the option of also setting up a mirror for the software package. What this means is that RIB will remember where the software was downloaded from and will keep its copy of the file consistent with the original copy.

What the end-user sees when visiting a repository created by RIB is a set of HTML pages that describe the software contained in that repository. These HTML pages are generated automatically by a set of administration scripts accessible to the repository maintainer via CGI. The repository maintainer does not need to know HTML to generate or maintain these pages. However, if the maintainer wishes to customize the appearance of his or her repository, then a basic knowledge of HTML is useful. Access to the administration scripts for a repository (and to the contents of the repositories themselves) can be protected by the usual HTTP server access control mechanisms. Each repository under an RIB installation can use its own set of access controls with separately configurable read and write access.

RIB utilizes the Basic Interoperability Data Model (BIDM) [IEEE, 1995] which is an IEEE standard (1420.1) for cataloging software on the Internet. This data model is very flexible and facilitates the reuse of catalog information by using an approach similar to object-oriented programming; information is encapsulated in "class" objects which can be linked by "relationships" to other objects (see figures 1 and 2). For example, two different software packages that were created by the same organization can share the information about that organization rather than recreating it. Any updates to the information about the organization will be reflected automatically in the software package descriptions. Though RIB was specifically designed for cataloging software, the BIDM configuration can be modified to catalog other object types, such as documentation, pictures, etc., allowing RIB to create a variety of on-line libraries.

Figure 1. Legend for BIDM graphic notation

Figure 2. BIDM relationships and classes

Using the BIDM allows software repositories created by RIB to interoperate, which means that they can share information about the software that they catalog. In order to facilitate the sharing of information over the Internet, RIB binds the BIDM to HTML. This binding enables BIDM information to be placed in the < HEAD > element of an HTML document in < LINK > and < META > tags. Note that the descriptions (i.e., metadata) for the software packages are shared, and not the actual software files themselves.

The interoperation feature of RIB is very important because it allows the creation of high level virtual repositories that import their data from domain specific repositories. Domain specific repositories are repositories that only contain software relevant to a specific discipline. These repositories have a higher level of quality because they are limited to certain domains and are controlled by experts in those domains. An example of this type of interoperation is demonstrated by the NHSE software repository catalog. The NHSE software repository imports a portion of its catalog information from a collection of the domain specific repositories discussed below in section 2.

RIB is currently in use by several organizations, including NASA and the Department of Defense's Major Shared Resource Centers (MSRC). These organizations hope to improve their use of resources by using RIB to archive and share the software that they create.

2. Domain Specific Repositories

The NHSE currently maintains four domain specific, software repositories. Domain experts manage each to ensure that all cataloged software is appropriate to each repository and that the software is placed in the appropriate classification. BenchWeb, Chemistry Software and Information Resources (CSIR), HPC-Netlib, and PTLib are products of NHSE's repository efforts. The catalogs of these repositories were constructed using the RIB toolkit.

BenchWeb is a repository which provides computer system performance benchmark information and software codes. In addition, benchmark results are provided along with links to other benchmark sites. Benchmark data for high performance computers to personal computers are provided along with vendor-supplied benchmark results.

CSIR is a repository housing two chemistry software catalogs. One catalog is for applications that run on machines from those with massively parallel processors down to simple desktops. The other catalog is specifically for high performance, parallel applications.

HPC-Netlib is the high performance branch of the popular Netlib mathematical software repository. In addition to a catalog of high performance mathematical software, both research and commercial, HPC-Netlib provides review information as well as a link to relevant literature.

PTLib (Parallel Tools Library) is a repository of information about high quality, parallel systems software and tools, both research and commercial. Besides the catalog, PTLib contains pointers to review information as well as links to relevant literature about parallel tools and systems.

NHSE provides a centralized "virtual library" for easily locating software applications specific to the high performance computing community. In addition, the maintainers of the NHSE provide help with common questions and forward specific questions to the software authors. By providing such information, the repositories facilitate direct communication between the user community and the software authors.

3. Software Reviews

As an added value to software cataloged within the NHSE repositories, reviews are systematically conducted to ensure that the software meets certain criteria. These reviews are performed by domain experts so that the repository users are provided reliable, unbiased information about the capabilities of software available. Currently, the NHSE provides two levels of reviews.

The first, called the *checked* level, is a partial review which focuses on the application's scope and documentation. The *checked* level is often performed by the repository maintainer and is a filter of sorts before the application is reviewed by a domain expert. Upon satisfactory completion of the first level, the application is moved to the *reviewed* level. Here the documentation is further scrutinized, and the software is checked for correctness, soundness, usability, and efficiency. During the *review* process, the application is checked to make sure it is bug-free; it does what it claims it will do; its methods used to solve the problem are sound; its user interface is understandable; it is easy enough to use by a typical NHSE user; and it is fast enough in that a slow speed does not render it ineffective. It should be noted that software cataloged within the repositories do not require that they have been *checked* or *reviewed*.

Once the reviews are completed, an appropriate icon is placed beside the application within the catalog so that potential users can readily identify those packages which have been examined by a reviewer. Information about the reviewers can also be found within the metadata of an application that has undergone the review process. Periodically, the reviews are made available via the *NHSE Review* ™, a Web-based, electronic journal devoted to providing the high performance computing community with comparative information about the software and related technologies.

4. Access Control and Licensing

Currently under development by the NHSE is a modular mechanism for restricting access to software distributed from virtual repositories. This tool will be compatible with RIB but capable of standing alone for those organizations that already employ repositories not built using RIB. A prototype is presently being tested by the University of Tennessee and Rice University before additional organizational onsite testing is done.

Both authentication and authorization mechanisms will be employed, allowing repository maintainers to protect software applications that have access restrictions such as commercially licensed and export controlled. Technologies such as X.509 certificates and MD5 checksums will be integrated atop an RDBMS architecture to provide a comprehensive, secure tool for licensing and distributing software files. An X.509 certificate contains information about the user which is electronically signed by a trusted third party. This certificate is then used to identify the user. An MD5 checksum is generatd by an algorithm that creates a unique "fingerprint" for a series of bytes contained in a file or string of characters. This "fingerprint" can be published by a repository maintainer and then used to verify the integrity of software files that have been stored or transferred via an untrusted third party.

The use of the database facilitates the ability to log user data so that profiles can be built which offer great detail about the software users. Many organizations will find this useful as a means for streamlining the software licensing process while allowing quicker user-access to the distributed software.

Figure 3 illustrates the prototype access control system currently under development. Once the authentication process has been completed to ensure that the user is who he or she claims to be, software may be downloaded via the following process:

Figure 3. Access control system

1. Send Certificate

The client browser (which must support X.509 certificates) presents the user's certificate to the HTTP server in a request for the restricted file, referenced by the file's catalog number in an extended URL for the CGI.

2. CGI Call

The HTTP server passes the request to the CGI

3-4. Database Queries

The CGI queries the database for the certificate serial number presented from the database list of those authorized to download the requested software package. The CGI also requests the actual filename and path for the software from the database.

If the user is authorized, the file download process begins. If not, the CGI returns an error message to the user explaining that they are not authorized to download the software.

5-7. Begin File Download

If the user is authorized to download, the CGI pulls the file from a directory not accessible to the HTTP server, according to the pathname provided by the database. The file is passed through the HTTP server on to the client browser.

Conclusions and Future Work

The NHSE continues to explore new avenues for providing useful tools and methods that promote software sharing and reuse among members of the HPCC. The RIB toolkit has been accepted as a viable means for creating Web-based, interoperable software repositories. With agencies such as NASA and the Department of Defense using the toolkit to develop in-house repositories, the NHSE hopes that other organizations will adopt RIB as the application of choice for creating and populating repository catalogs that will share metadata with other organizations. RIB's use of the BIDM standard provides a suitable platform for creating catalogs of software divided by specific domains. Currently, NHSE alliances are being established with the National Science Foundation (NSF) funded National Computational Science

Alliance (NCSA) and National Partnership for Advanced Computational Infrastructure (NPACI).

In future versions of RIB, an underlying database will replace the file system currently used by RIB. The employment of a database will allow tighter integration with the access control module. The database benefits to the user will include better concurrency control, faster access to RIB data, more presentation and query options for RIB data, increased portability of RIB data, and a simpler code base. Also, since the creation of the RIB software, a new standard for exchanging structured data on the Internet has emerged called XML (eXtensible Markup Language). Plans are underway to incorporate XML into RIB's strategy for encoding BIDM information. Bringing the RIB data exchange format in line with XML will allow RIB to exchange information with other applications that conform to this standard.

References

IEEE Std 1420.1-1995, *Standard for Information Technology--Software Reuse--Data Model for Reuse Library Interoperability: Basic Interoperability Data Model (BIDM)*.

Top | Magazine
Search | Author Index | Title Index | Monthly Issues
Previous Story | Next Story
Comments | E-mail the Editor

hdl:cnri.dlib/may98-browne