Management of the NHSE – a Virtual Distributed Digital Library *

Shirley Browne†
University of Tennessee

Jack Dongarra
Tom Rowan
University of Tennessee and Oak Ridge National Lab

Ken Kennedy
Rice University

March 27, 1995

Abstract

The National HPCC Software Exchange (NHSE) is a distributed collection of software, documents, and data of interest to the high performance computing community. Our experiences with the design and initial implementation of the NHSE are relevant to a number of general digital library issues, including the publication process, quality control, authentication and integrity, and information retrieval. This paper describes an authenticated submission process that is coupled with a multilevel review process. Browsing and searching tools for aiding with information retrieval are also described.

*The work described in this paper was sponsored by NASA under Grant No. NAG 5-2736

†Author to whom correspondence should be directed. 107 Ayres Hall, Computer Science Department, University of Tennessee, Knoxville, TN 37996-1301, (615) 974-5886, browne@cs.utk.edu

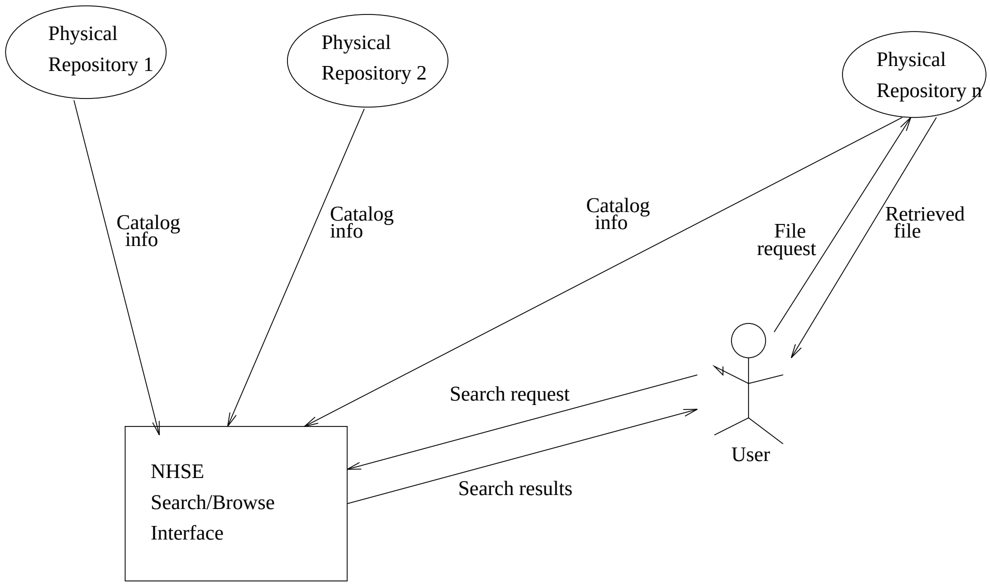

Figure 1: Virtual Repository Architecture

1 Introduction

The National HPCC Software Exchange (NHSE) is an Internet-accessible resource that provides access to software and other information related to High Performance Computing and Communications (HPCC). The NHSE facilitates the development of discipline-oriented software and document repositories. Furthermore, it promotes contributions through and use of such repositories by members of the high performance computing community, via a common World Wide Web interface. The NHSE is also a valuable resource for technology transfer and educational purposes.

The effectiveness of the NHSE depends on discipline-oriented groups having ownership of independently maintained repositories. The information and software residing in these repositories is best maintained and kept up-to-date by its developers, rather than by centralized administration. Developers may also wish to provide specialized services or access methods, depending on the nature of the repository, for example a remote execution facility. Central administration is used instead to handle interoperation and to meet common needs.

Although the different disciplines maintain their own software repositories, users should not need to access each of these repositories separately. Rather, the NHSE provides a uniform interface to a virtual HPCC software repository built on top of a distributed set of discipline-oriented repositories, as shown in Figure 1. The interface assists the user in locating and retrieving relevant resources.

In order for the NHSE to provide an information retrieval interface to the distributed collection of materials, it must have the raw material available from which to build indexes and other searching and browsing aids. Various techniques for collecting and indexing descriptive material are used in the NHSE, including manual construction of catalog records, collection and indexing of unstructured text, and computer-assisted construction of a hypertext roadmap.

Users of the NHSE need to have confidence that the software they obtain is of high quality and well-tested. If the software is experimental or untested, they should be made aware of this. The NHSE has developed a review process that allows authors to submit software for consideration at different levels of review classification, with the rigor of the review process increasing with increasing levels.

A contributor to the NHSE makes a contribution available by placing it on a file server accessible via the FTP or HTTP file access protocols and informing the NHSE of its existence. The NHSE can then provide a pointer in the form of a URL, along with a description of the contribution. For software contributions, it is important for purposes of reviewing and version control and tracking to ensure the property of fixity of publication – i.e., that the software has not been changed since the time of submission unless the NHSE has been informed of the change. Because of copyright, liability, and other legal issues, it is also important that someone not be able to masquerade as someone else or make unauthorized changes to someone else's contributions. For these reasons, the NHSE has developed authenticity and integrity checking mechanisms for software submissions based on file fingerprints and a public-key cryptosystem.

2 Software Submission and Review

Contributors submit software to the NHSE by filling out an HTML form using a forms-capable WWW browser such as Mosaic or Netscape [1]. This form explains the submission and review process, including the authentication procedures, and gives an example of a completed submission form. The form asks the user to fill in values for several attributes, some required and some optional. These attributes form a subset of those specified in the Reuse library Interoperability Group (RIG) Basic Interoperability Data Model (BIDM) [1]. The remaining BIDM fields are generated by the NHSE librarian or from default values. The RIG has been chartered by the IEEE to develop standards for reuse library interoperation. Use of the BIDM standard by the NHSE will facilitate interoperation with other reuse libraries also adopting this standard, including a number of existing government and industry reuse libraries (e.g., ASSET, CARDS, DSRS, ELSA).

[1]The NHSE software submission form is accessible at
`http://www.netlib.org/nse/software_submit/software_submit.html`

Some contributors may have fairly large collections that are already indexed using a different data model. The NHSE will provide assistance to such contributors in converting their indexing information to the form required for submission to the NHSE and in submitting such collections en masse.

2.1 Review Levels

Currently three levels of software are recognized in the NHSE, described as follows:

Unreviewed. The submission has not been reviewed by the NHSE for conformance with software guidelines. This classification is for unreviewed software available on an "as is" basis.

Partially reviewed. The submission has undergone a partial NHSE review to verify conformance with the scope, completeness, documentation, and construction guidelines. These particular guidelines are those that can be verified through a visual inspection of the submission.

Reviewed. The submission has undergone a complete NHSE review to verify conformance with all the software guidelines. This classification requires peer-review testing of the submitted software. This level may be further refined into additional levels in the future.

To receive the Partially reviewed rating, software submitted to the NHSE should conform to the following guidelines:

Scope. Software submitted to the NHSE should provide a new capability in numerical or high-performance computation or in support of those disciplines.

Completeness. Submissions must include all routines and drivers necessary for users to run the software. Source code for widely available software used by the submission, `blas` and `lapack` for example, need not be included as part of the submission.

Documentation. The software contains complete and understandable documentation on its use.

Construction. Submissions must adhere to good mathematical software programming practice and, where feasible, to language standards. Software should be constructed in a modular fashion to facilitate reusability. The use of language checking tools, such as `pfort` or `ftnchek`, is recommended.

To be accorded the reviewed status, the software must first have been accorded the partially reviewed status. This precondition ensures that reviewers will be able to access all the information needed to carry out the review over the National Information Infrastructure.

Software submitted for full review is reviewed according to the following criteria:

Documentation. The software contains complete, understandable, and correct documentation on its use.

Correctness. The software is relatively bug-free and works as advertised on all provided data sets and on data sets constructed by the reviewer according to the documentation.

Soundness. The methods employed by the software are sound for solving the problem it is designed for, as described in the documentation.

Usability. The software has an understandable user interface and is easy to use at the level of a typical NHSE client.

Efficiency. The software runs fast enough, in that slow speed does not make it an ineffective tool.

After software has been submitted for full review, it is assigned to an area editor, who recruits two to six reviewers to peer review the software according the above criteria. To qualify for full review, an author must provide sample data and the output from or a description of results from each sample. Each reviewer is asked to read the software documentation and try the software on some of the data sets provided by the author. In addition, it is recommended that a reviewer test the software on inputs not provided by the author. If source is available, the reviewer examines the source to ensure that the methods and programming methodology are of acceptable quality. Each reviewer prepares all comments in electronic form and returns these, along with a recommendation to the editor in charge of the review. After the peer reviews are returned, the editor makes the final decision as to whether to accept the software and informs the author of the decision. If the software is accepted, the area editor prepares a review abstract for use by the NHSE.

Once the software has been reviewed, one of two things happens. If it is not accepted, the author will be so informed and anonymous copies of the review or reviews will be provided. The author may then choose to address the reviewers' comments and resubmit the revised software. If the software is accepted, the author will be shown a review abstract summarizing the reviewer comments. This abstract will be available to anyone who accesses the software through the NHSE. If the author finds the abstract unacceptable, he or she may withdraw the software and resubmit it for review at a later date.

2.2 Authentication Procedures

After a contributor fills out the NHSE software submission form and submits it, a program is invoked at an NHSE server that checks the form for any obvious errors, such as omission of required attributes, incorrectly formed email addresses, or unretrievable URLs. If no errors are found, a plain-text version of the catalog record is returned to the client program, along with instructions to save the plain text version to a file and carry out one of the following authentication procedures:

PGP Authentication [7]. The author uses his public, NHSE-certified PGP key to sign the catalog record and then mails it back to a designated address. The mail server at that address verifies the PGP signature and processes the submission only if the signature is valid.

Notarization. The author prints out the plain text form, signs it, has the signature notarized, and sends the document back via surface mail. When the form is received, a the NHSE librarian PGP-signs the electronic version of the form (using a special proxy key reserved for this purpose) on behalf of the author.

Before using method 1, the author must have PGP installed on his/her system and have obtained a PGP key pair. The author's public key must have been certified by the NHSE librarian. An author may obtain this certification either in person, via a trusted third party who signs the author's key, or by a method similar to 2 above: print out the key fingerprint, sign it, have it notarized, and surface mail it to the NHSE librarian.

We considered other authentication methods, such as email addresses and userid/password based accounts, but rejected such methods as providing insufficient security.

2.3 Identification, Cataloging, and Integrity

Once an author's software submission has been authenticated, it is processed before being placed in the NHSE on-line software catalog. This processing involves retrieval of the files specified by the author as making up the contribution, fingerprinting these files, assigning the contribution a unique identifier, and additional cataloging by the NHSE librarian. If the software has been submitted for partial review, the NHSE librarian also inspects the submission for adherence to the NHSE software guidelines.

After the files making up a contribution have been retrieved, each of these files is fingerprinted using the MD5 secure hash function [6]. The (URL,MD5) pairs for the files are then placed in another file which is itself fingerprinted. This top-level fingerprint is used to construct a unique identifier for the submission which we call a LIFN, or Location Independent File Name. The submission can subsequently be retrieved from the NHSE software catalog by specifying its LIFN.

The LIFN concept is part of a more general naming structure that is being developed to provide for transparent mirroring of files and to address other scalability and reliability problems that will result from the expected growth of the NHSE [4].

As part of the processing, the NHSE librarian categorizes the software submission into one of four main categories: application libraries and programs, data analysis and visualization tools, numerical libraries and routines, and parallel processing tools. Software falling under parallel processing tools is categorized further into one of eight subcategories. The NHSE librarian also assigns keywords drawn from the HPCC thesaurus (currently under development) and, for mathematical software, from the GAMS classification scheme [2].

The NHSE provides a form, called the LIFN verification form, that allows a user to verify the integrity of any submission [2]. A contributor may also use this form to check whether he has changed any of his files since submitting them. To use the form, the user enters the LIFN he wishes to verify and presses the Verify button. This action causes a program to be invoked on an NHSE server that carries out the following steps:

1. retrieves the fingerprint file that was constructed when the LIFN was assigned and which contains the URLs and the stored fingerprints for the files making up the submission

2. retrieves the files using the designated URLs

3. computes the MD5 fingerprint for each of the retrieved files and compares it with the stored fingerprint that was previously computed for the same URL

4. flags any file that has been changed since the LIFN was assigned and asks the user if he would like to retrieve the original file as archived by the NHSE

2.4 Updating a Previous Submission

A contributor may update or withdraw a previous submission by using the NHSE software submission change form [3]. This form asks the contributor to enter the LIFN for the previous submission. If the contributor does not know the LIFN, he can search for his submission in the NHSE software catalog in order to determine it. After the contributor enters the LIFN, he presses a button

[2] The NHSE LIFN verification form is accessible at
`http://www.netlib.org/nse/software_submit/lifn_verify.html`

[3] The NHSE software submission change form is accessible at
`http://www.netlib.org/nse/software_submit/submit_change.html`

that causes the catalog record for the LIFN to be retrieved and displayed in a second form. The contributor may then specify any files that have been changed or added, describe changes made to the files, and/or update cataloging information.

After the contributor fills out the change form and submits it, he is asked to authenticate his change request using one of the two authentication procedures described in section 2.2. Note, however, that if the submission was initially authenticated using PGP, the NHSE will be extremely cautious about accepting updates authenticated using the notarization method.

3 Information Retrieval Aids

Depending on the size, rate of change, and nature of the underlying software or document database, the NHSE uses different techniques for assisting the user in searching and browsing the information. Smaller and/or fairly stable collections may permit a labor-intensive indexing and abstracting process, with resulting benefits of improved recall and precision for searches. Larger or rapidly changing collections require the use of less precise automatic indexing techniques.

The current NHSE software catalog [4] is fairly small, with fewer than 300 entries. Thus, it has been possible to manually abstract and index this collection. The cataloging process has been carried out jointly by the authors and the NHSE librarian, with the authors providing the title and abstract fields, and the NHSE librarian categorizing each entry and assigning thesaurus keywords. The NHSE software catalog is available in the following formats:

1. An HTML version that may be browsed by category.
2. A searchable version that allows the user to search separately by different attributes or to do a free-text search on the catalog records. A link to an on-line copy of the HPCC thesaurus is provided so that users may select controlled vocabulary terms for searching. The current interface requires users to cut and paste thesaurus terms into the search form. We plan to develop a hypertext version of the thesaurus that will statically link thesaurus terms to scope and definition notes and to related terms (also broader terms and narrower terms) as well as dynamically link thesaurus terms to indexed catalog entries.
3. A PostScript version that may be downloaded and printed.

A number of sites involved with the NHSE maintain collections of technical reports on numerical and/or high performance computing. These collections are frequently already indexed and abstracted, although they may use different

[4] Accessible from `http://www.netlib.org/nse/home.html`

indexing formats. One such collection is maintained at the University of Tennessee Computer Science Department (UTKCS). UTKCS is joining the CSTR project, and other NHSE sites will be encouraged to do likewise. The CSTR project is developing standards and technologies for digital document repositories [5]. The Dienst server software available from Cornell University facilitates searching for and retrieving documents from a repository and linking together different repositories so that all may be searched from any site. Dienst also provides utilities that assist sites with installing the document database and converting from other indexing formats [6].

In addition to the software catalog, the NHSE has a distributed hypertext structure that contains a variety of information on high performance computing. Most of this information is in the form of HTML pages, but there are also links to documents in other formats, such as plain text and PostScript. Links are provided to various HPCC programs and activities, to descriptions of grand challenge applications, and to other software repositories. Because the collection of information has grown very large, a search interface has been provided. This search interface currently uses the Harvest system [3] to collect information from remote sites, index that information using WAIS, and process queries from users. The Harvest system worked satisfactorily at first, but the underlying database has now grown so large and diverse that 1) the gathering takes on the order of several days to a few weeks, and the search interface becomes out-of-date in the meantime, 2) extremely large result sets are returned by many searches. Work is underway both by the Harvest development group and by NHSE researchers at Argonne National Laboratory to address these scalability problems.

Hypertext roadmaps are being developed at Syracuse University to provide guided tours to HPCC software and technologies [7]. The roadmap consists of encyclopedia-like articles written by experts in the field, with links to relevant software and technologies. Because construction of such a guide is labor-intensive and because the resulting structure is static, the roadmap can encompass only a portion of the available information. However, we hope to use semantic indexing techniques such as LSI [5] to leverage the work on the roadmap by automatically inferring relationships to new material.

References

[1] Standard reuse library Basic Data Interoperability Model (BIDM). Technical Report RPS-0001, Reuse Library Interoperability Group, 1993.

[5] More information about the CSTR project is available at `http://www.cnri.reston.va.us/`

[6] More information about Dienst is available at `http://cs-tr.cs.cornell.edu/Info/server.html`

[7] A prototype roadmap is accessible from the NHSE home page at `http://www.netlib.org/nse/home.html`

[2] R. F. Boisvert, S. E. Howe, and D. K. Kahaner. The Guide to Available Mathematical Software problem classification system. *Comm. Stat. – Simul. Comp.*, 20(4):811–842, 1991.

[3] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz. Harvest: A scalable, customizable discovery and access system. Technical Report CU-CS-732-94, Department of Computer Science, University of Colorado - Boulder, Aug. 1994.

[4] S. Browne, J. Dongarra, S. Green, K. Moore, T. Pepin, T. Rowan, and R. Wade. Location-independent naming for virtual distributed software repositories. In *ACM-SIGSOFT 1995 Symposium on Software Reusability*, Seattle, Washington, Apr. 1995.

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshamn. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, Sept. 1990.

[6] R. Rivest. The MD5 message-digest algorithm. *Internet Request for Comments*, 1321, Apr. 1992.

[7] P. Zimmerman. Pgp user's guide. PGP Version 2.6.2, Oct. 1994.