Evolving Software Repositories

http://www.netlib.org/utk/projects/esr/

Jack Dongarra
University of Tennessee and
Oak Ridge National Laboratory

Ron Boisvert
National Institute of Standards and Technology

Eric Grosse
AT&T Bell Laboratories

Project Focus Areas

- NHSE Overview
- Resource Cataloging and Distribution System (RCDS)
- Safe execution environments for mobile code
- Application-level and content-oriented tools
- Repository interoperability
- Distributed, semantic-based searching

NHSE

- National HPCC Software Exchange
- NASA (plus other agencies) funded CRPC project
- Center for Research on Parallel Computation (CRPC)
  - Argonne National Laboratory
  - California Institute of Technology
  - Rice University
  - Syracuse University
  - University of Tennessee
- Uniform interface to distributed HPCC software repositories
- Facilitation of cross-agency and interdisciplinary software reuse
- Material from ASTA, HPCS, and IITA components of the HPCC program
- http://www.netlib.org/nhse/

Goals:

- Capture, preserve and make available all software and software-related artifacts produced by the federal HPCC program. (Software related artifacts include algorithms, specifications, designs, documentation, report, ...)

- Promote formation, growth, and interoperation of discipline-oriented repositories that organize, evaluate, and add value to individual contributions.

- Employ and develop where necessary state-of-the-art technologies for assisting users in finding, understanding, and using HPCC software and technologies.

Benefits:

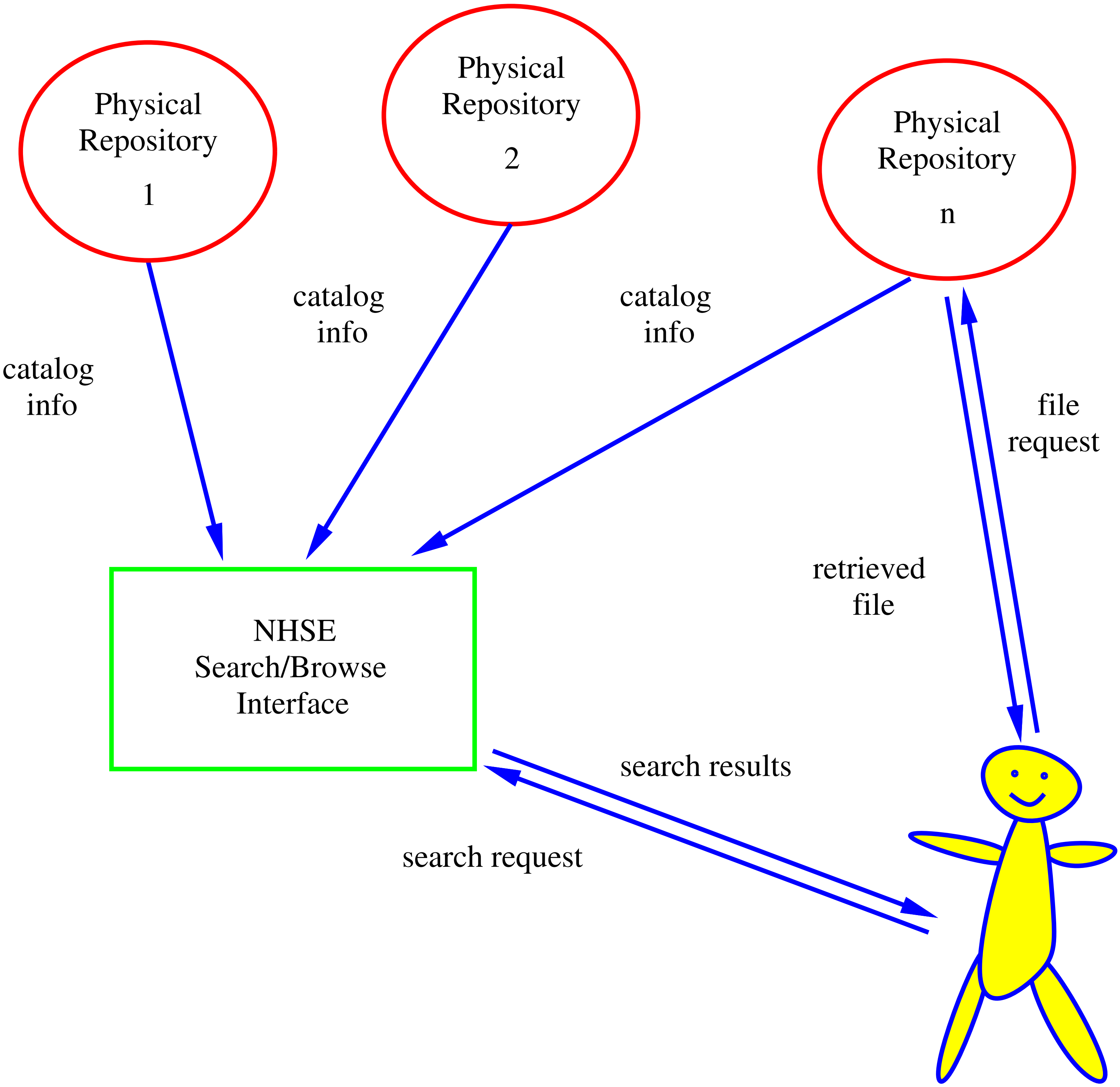
1. Faster development of high-quality software so that scientists can spend less time writing and debugging programs and more time on research problems.
2. Less duplication of software development effort by sharing of software modules.
3. Less time and effort spent in locating relevant software and information through the use of appropriate indexing and search mechanisms and domain-specific expert help systems.
4. Reducing information overload through the use of filters and automatic search mechanisms.

Intended Audience:

- HPCC application and computer science community
  - Source of material for NHSE
- Users of NASA, NSF, DOE and other supercomputer centers
  - Good targets for NHSE
  - Natural support organization: supercomputer center staff
- Other users of high performance computers
  - Current and potential industrial users
  - No natural support organization
- Applicable to other domains

NHSE Components:

- Discipline Oriented Repositories
- Submission and Review
- Common Infrastructure
  - Resource cataloging and distributed system
  - Repository tools and resource center
  - Naming and authentication
  - Publishing tools
- HPCC Specific Searching
- Outreach and technology transition
  - To the HPCC user community and industry
- Measurement
- Hypertext Road Map
- Selective Capitalization of Emerging Technologies
- Collection Management



Virtual Repository Architecture

NHSE

- Based on Existing Technologies
  - WWW Browser (Mosaic / Netscape / etc)
    - Distributed / Scalable
    - URL: http://www.netlib.org/nhse/
  - Netlib
    - Repository for math software since 1985
- Repositories Currently Available
  - Netlib, Softlib, CITlib
  - ASSET - (Asset Source for SW Engineering Tech.)
  - CARDS - (Comprehensive Approach to Reusable Defense SW)
  - ELSA - (Electronic Library Services and Appl.)
  - GAMS (Virtual Software Repository)
  - STARS - (SW Technology for Adaptable, Reliable Systems)
  - Many examples related to GC problems
- Currently Available Information
  - NHSE currently points to 350+ modules
    - software catalog, tech reports and papers
    - parallel processing tools
    - numerical libraries
    - Grand Challenge prototype codes
    - data analysis and visulization
    - benchmarks

Discipline Oriented Repositories Interoperation

- The NHSE will catalog and provide access to software and software-related artifacts from all the HPCC software repositories.
- Assets accessible from other existing software repositories, such as ASSET, CARDS, DSRS, and ESLA, may also be of interest to NHSE users.
- The NHSE will be participating in a small-scale interoperability experiment with the above repositories to help define requirements for further interoperation efforts.
- The NHSE will also be working with the Reuse Library Interoperability Group (RIG) on establishing standards for unique naming, asset description and classification, and asset evaluation.
- In the future, the NHSE will interoperate with these other repositories so that software from them may be retrieved directly from the NHSE interface.

Netlib - Network Access to Mathematical Software and Data

- Began in 1985
  - JD and Eric Grosse, AT&T Bell Labs
- Motivated by the need for cost-effective, timely distribution of high-quality mathematical software to the community.
- Designed to send, by return electronic mail, requested items.
- Automatic mechanism for the disseminate of public domain software.
  - Still in use and growing
  - Mirrored at a number of sites
    - netlib2.cs.utk.edu
    - netlib1.epm.ornl.gov
    - research.att.com
    - netlib.no
    - unix.hensa.ac.uk
    - ftp.zip-berlin.de
    - nchc.edu.tw
- Moderated collection of high-quality math software
- Distributed maintenance
- Model for domain-specific repositories

Netlib – Network access to mathematical software and data

Jack Dongarra *Univ. Tenn. and ORNL*
Eric Grosse *AT&T Bell Labs, Murray Hill NJ*

netlib

Started in 1985.

Motivated by the research community

Uses email for the distribution.

Has grown in popularity and scope.

Funded by the NSF and Bell Labs.

- The development of NETLIB was motivated by the need for cost-effective, timely distribution of high-quality mathematical software to the research community at large.
- The system was designed to send, by return electronic mail, requested routines together with subsidiary routines and any related documents or test programs supplied by the authors.
- Automatic mechanism for the disseminate of public domain software.

Try:

mail netlib@ornl.gov
mail netlib@research.att.com
send index
send rs from eispack
who is Golub

Collection includes:

Linpack	Eispack	Fishpack
Odepack	ACM TOMS	Benchmark
Bihar	Blas	BMP
Conformal	f2c	FMM
Fnlib	Fftpack	Hompack
Lanczos	LP/data	Minpack
Napack	NL2SOL	Odepack
Paranoia	Pltmg	Polyhedra
Port	Pppack	Quadpack
SIAM memship	Sparspak	Typesetting
Vanhuffel	Voronoi	...

Netlib provides the following features:

- There are no administrative channels to go through.
- Since no human processes the request, it is possible to get software at any time, even in the middle of the night.
- The most up-to-date version is always available.
- Individual routines or pieces of a package can be obtained instead of a whole collection.

Over around 15000 requests a day.

Software collection about 1 Gbytes

21K file in 330 libraries/directories

Interdisciplinary resource

- Software
- Parallel processing collection
- Data
- Tools
- Reports
- Documentation
- Benchmarks
- Journal information

Synchronization and Netlib Sites:

- Still in use and growing
- Mirrored at a number of sites
  - netlib2.cs.utk.edu
  - netlib1.epm.ornl.gov
  - research.att.com
  - netlib.no
  - unix.hensa.ac.uk
  - ftp.zip-berlin.de

Offshoots

Other sites running the netlib processor, but to support other databases:

- statlib@temper.stat.cmu.edu *statistics*
- tuglib@science.utah.edu *TEX*
- reduce-netlib@rand.org *Reduce symbolic algebra*
- maple-netlib@can.nl *Maple symbolic algebra*
- nistlib@cmr.ncsl.nist.gov *benchmarks*

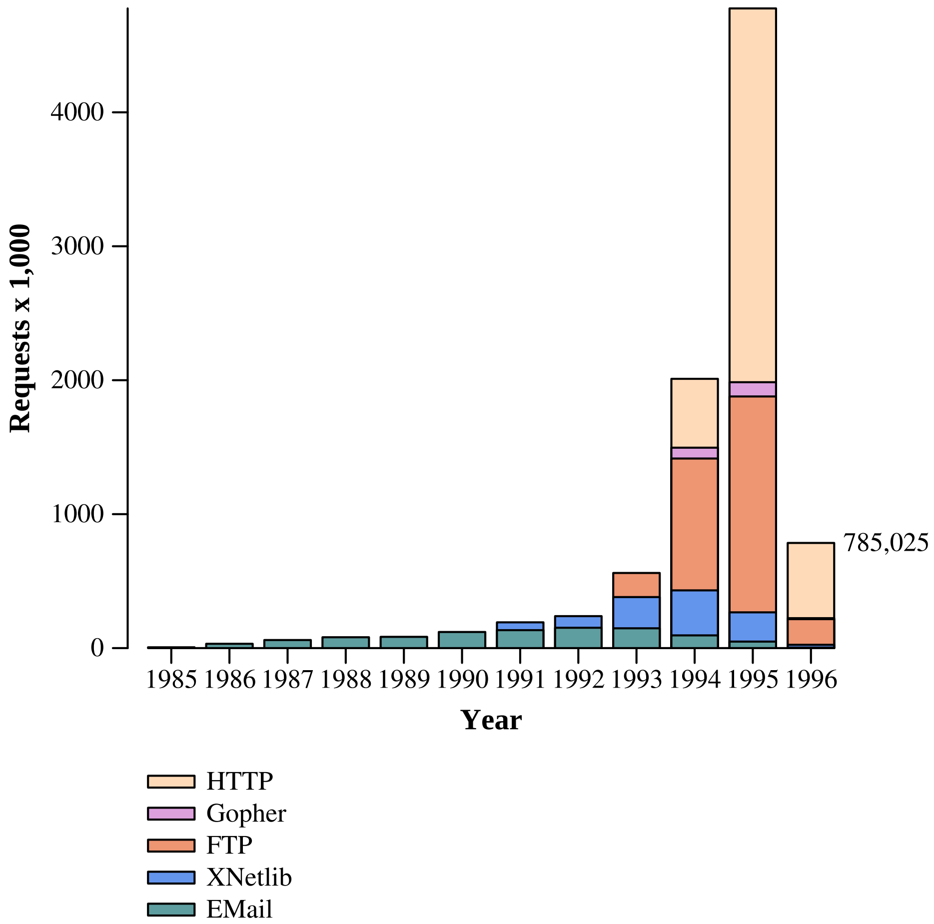
Well over a hundred copies of netlib itself have been shipped.

NETLIB does not offer

- Technical assistance in determining and correcting problems with library software.
- Procedures for testing or validating codes.
- A uniform style for programming and documentation.
- Uniform error handling within the library.

Requests Made to the Netlib Repositories at the Univ. of Tennessee & ORNL

8,946,816 total requests to these repositories as of Feb 15, 1996



Data as of - Feb 15, 1996 at 02:07:27

Breakdown of requests to each Netlib library

(an alphabetical listing is also available.)
Data as of 02/15/96 at 02:09:20

Library Name	Number of accesses
lapack	475,979
pvm3	379,849
linpack	256,403
slatec	248,292
blas	178,728
clapack	129,256
linalg	127,022
eispack	126,116
slatec/src	118,366
toms	117,152
f2c	98,025
c++	96,774
benchmark	85,552
master	69,997
f2c/src	67,415
minpack	60,632
fn	59,781
fftpack	58,805
na-digest	50,970
port	49,496
slatec/lin	46,800
hence	45,229
confdb	43,118
slatec/chk	37,640
c++/answerbook	37,524
napack	36,719

Discipline Oriented Repositories
Different disciplines will maintain their own software repositories

- Users should not need to access each of these repositories separately
- NHSE will provide a uniform interface to a virtual HPCC software repository which will be built on top of the distributed set of discipline-oriented repositories.
- The interface will assist the user in locating relevant resources and in retrieving these resources.
- A combined browse/search interface will allow the user to explore the various HPCC areas and become familiar with the available resources.
- A longer term goal of the NHSE is to provide users with domain-specific expert help in locating and understanding relevant resources.

Discipline Oriented Repositories
HPCC Cataloging

- To enable searching, cataloging information must be made available for NHSE assets.
- Each physical repository will be responsible for maintaining a network-accessible file containing such cataloging information.
- These files will be retrieved and indexed by an NHSE indexer on a regular basis, and the resulting searchable index will be replicated for reliability.
- The NHSE is using the Harvest system from University of Colorado to do the collection, indexing, and index replication.

Discipline Oriented Repositories

Why domain specific?

- Cohesive contributor/user community
- Quality control,peer review
- Searching tuned for subject matter

Why interoperate?

- Reduce redundant software development efforts
- Interdisciplinary problem solving
- Collaboration and technology transfer between industry, government, and academia

Discipline Oriented Repositories

Under investigation...

- CFD - NASA Langley, ICASE
- Computational Chemistry - Sam Trickey, University of Florida
- Material Science - Kevin Kremeyer, University of Arizona
- Climate and Groundwater Modeling - Ken Kliewer, ORNL

Common repository tools and infrastructure provided by NHSE developes

Overall Strategy for the NHSE:

- Effectiveness of the NHSE will depend on discipline-oriented groups and Grand Challenge teams having ownership of the discipline-oriented software repositories.
- The information and software residing in these repositories will be best maintained and kept up-to-date by the individual disciplines, rather than by centralized administration.
- Central administration will be used instead to handle interoperation and meet common needs.
- Although the various disciplines will have ownership of the repositories, they should not be expected to develop the software and tools for building, managing, and interfacing to their repositories.
- Much useful information retrieval (IR) software is currently available, and this software should be incorporated into the NHSE.

NHSE Software Submission

Goals

- Exercise quality control
  (review classification)
- Ensure *fixity of publication*
  (file fingerprints, unique name)
- Prevent impersonation and unauthorized changes
  (digital signatures)
- Promote interoperability
  (RIG Basic Interoperability Data Model)

Review Procedure

- Submissions are “subject” to (ongoing) review.
- Review status abstract for each submission.
  - Based on author comments
  - Package documentation
  - Our independent reviewer testing
  - Comments from users.

NHSE Software Catalog

- Benchmark and example programs (4)
- Data analysis and visualization (22)
- Numerical libraries and routines (57)
  - Computational geometry (7)
  - Linear algebra (18)
  - Optimization (4)
  - Partial differential equations (3)
  - Other (25)
- Parallel processing tools
  - Communication libraries (25)
  - Execution and performance analyzers (31)
  - Parallel I/O systems (5)
  - Parallel programming environments (12)
  - Parallel programming languages and compilers (26)
  - Parallel runtime systems (10)
  - Source code analyzers and restructurers (7)
  - Miscellaneous (16)
- Scientific and engineering applications (66)

The following types of software are being made available:

- Systems software and software tools.
  - compilers
  - message-passing communication subsystems
  - parallel monitors and debuggers.
- Basic building blocks for accomplishing common computational and communication tasks.
  - Building blocks are meant to be used by Grand Challenge teams
- Research codes that have been developed to solve difficult computational problems.
  - Many have been developed to solve specific problems
  - Serve as proofs of concept
  - Models for developing general-purpose reusable software

Technical and Political Issues

- How will the naive user find the right software?
  - Answer: Via the NHSE search/browser interface and the Road Map
- How will authentication, integrity, and version control be implemented?
  - Answer: By a publishing system that includes unique naming & digital signatures.
- Will the NHSE support distribution of software that is not free or cannot be freely distributed?
  - Answer: Yes, but...
    * only provide classification, review, and access
    * use of encryption and separate key distribution
    * NHSE will not have an accounting department
- Will the NHSE be responsible for support of software?
  - Answer: NO!
    * Any support will be by author (or appropriate agent)

Searching
User Profiles

- **Currently processed manually**
  - **Click on Submit: User Profile on NHSE home page**
  - **Fill in email address, software needs, information needs**
  - **NHSE Librarian sends you search results and keeps you posted on future items of interest enditemize**
  - **Future - automate processing**
    - **Handle larger volume more quickly**
    - **Intelligent information agents**
    - **Semantic filtering**

Affiliations, Collaborations, and Intimacies

- **W3 Consortium affiliate member (W^3C)**
- **Reuse Library Interoperablity Group (RIG) Member**
- **Active within IETF**
- **Interaction with Corporation for National Research Initiatives (CNRI)**

Roadmap to HPCC Enabling Technologies and Software

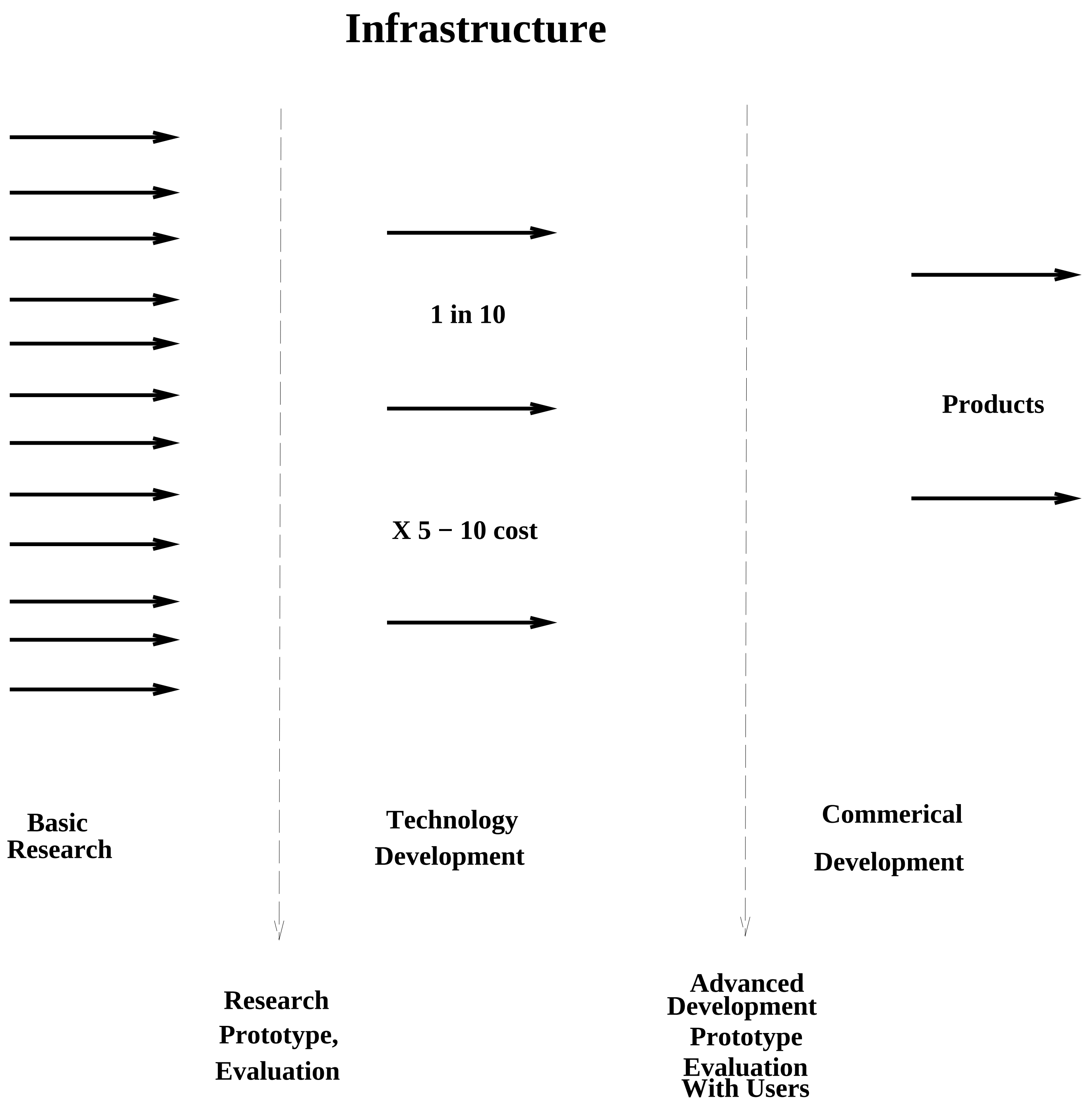
An integrated hierarchy of information

- **Similar to a hypertext encyclopedia**
- **Assembled with the help of panels of experts**
- **A glossary of HPCC terms, algorithms, applications, and enabling technologies.**
- **Further information on each topic**
- **Pointers to relevant software in the NHSE.**
- **More comprehensive overviews of some enabling HPCC technologies, applications and algorithms, in the form of short review articles.**
- **A collection of material relevant to HPF has been prepared and current efforts are focussed on improving quality and presentation of material.**
- **A description of HPCC systems was prepared**

Proposed
Capitalization of Emerging Technologies

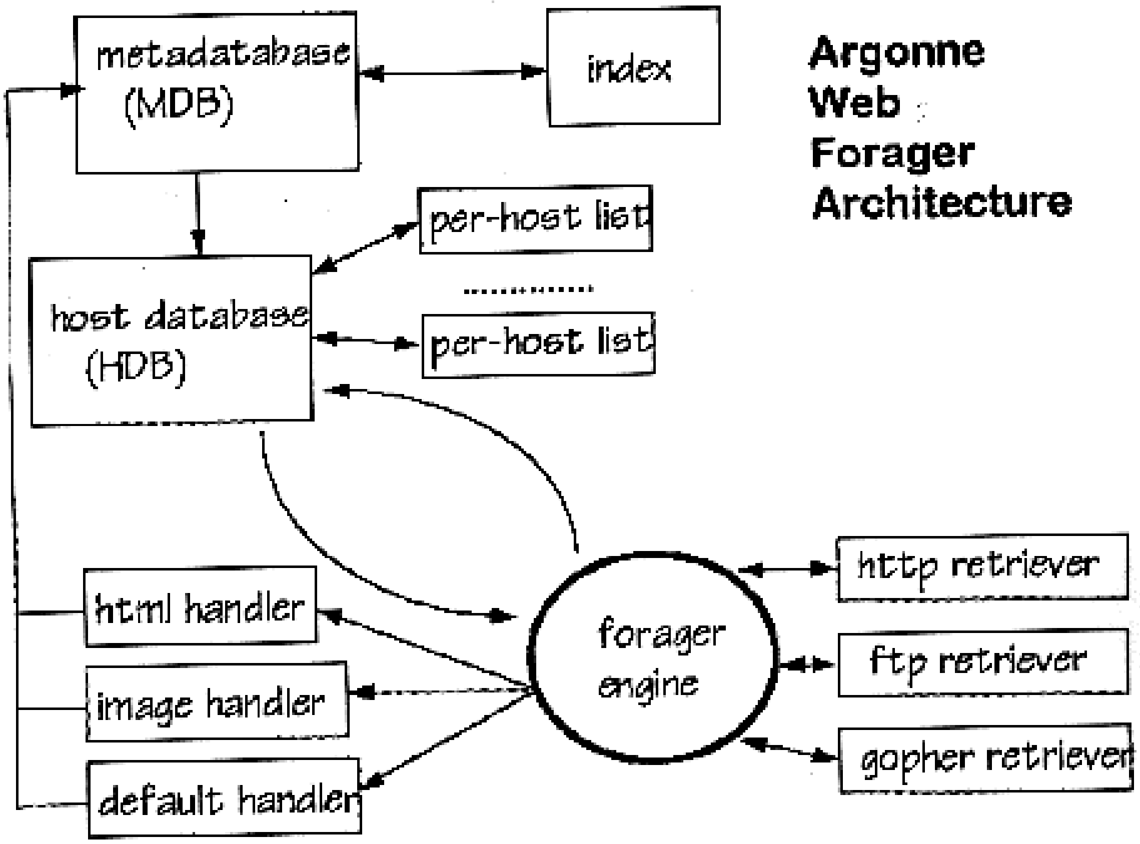
- **Three Levels of Software Development (Pasadena I Workshop)**
  - **Research prototype**
  - **Advanced development prototype**
  - **Commercial product**
- **Focus:**
  - **Move from research to advanced development prototype**
- **Strategy**
  - **Convene outside review panel**
  - **Select as many projects as budget will permit**
- **Ready to carry out plan - Not Funded**

Software Technology Development



Searching

- **Modular web robot**
- **Parallel web indexing engine**
- **Autonomous agents for collecting information on the web**


metadatabase (MDB)
index
Argonne Web Forager Architecture
per-host list
per-host list
host database (HDB)
html handler
image handler
default handler
forager engine
http retriever
ftp retriever
gopher retriever

Measurement

- **Keep statistics on downloads from NHSE pages**
  - **Including userid of request source**
    - **not always possible**
    - **invasion of privacy**
- **Statistical Survey of Unreviewed and Reviewed Software**
  - **Identification of candidates for full review**
- **Systematic Survey of Selected Software**
  - **Application**
  - **Usage pattern**

Summary

- **Initial implementation built on existing technologies**
  - **WWW**
    - **Distributed**
    - **Scalable**
  - **Netlib, etc**
  - **Rapid deployment**
- **Multilevel review and classification scheme**
- **Road Map**
- **Measurement and evaluation**
  - **Statistics**
  - **Evaluations for reviewed**

Summary (continued)

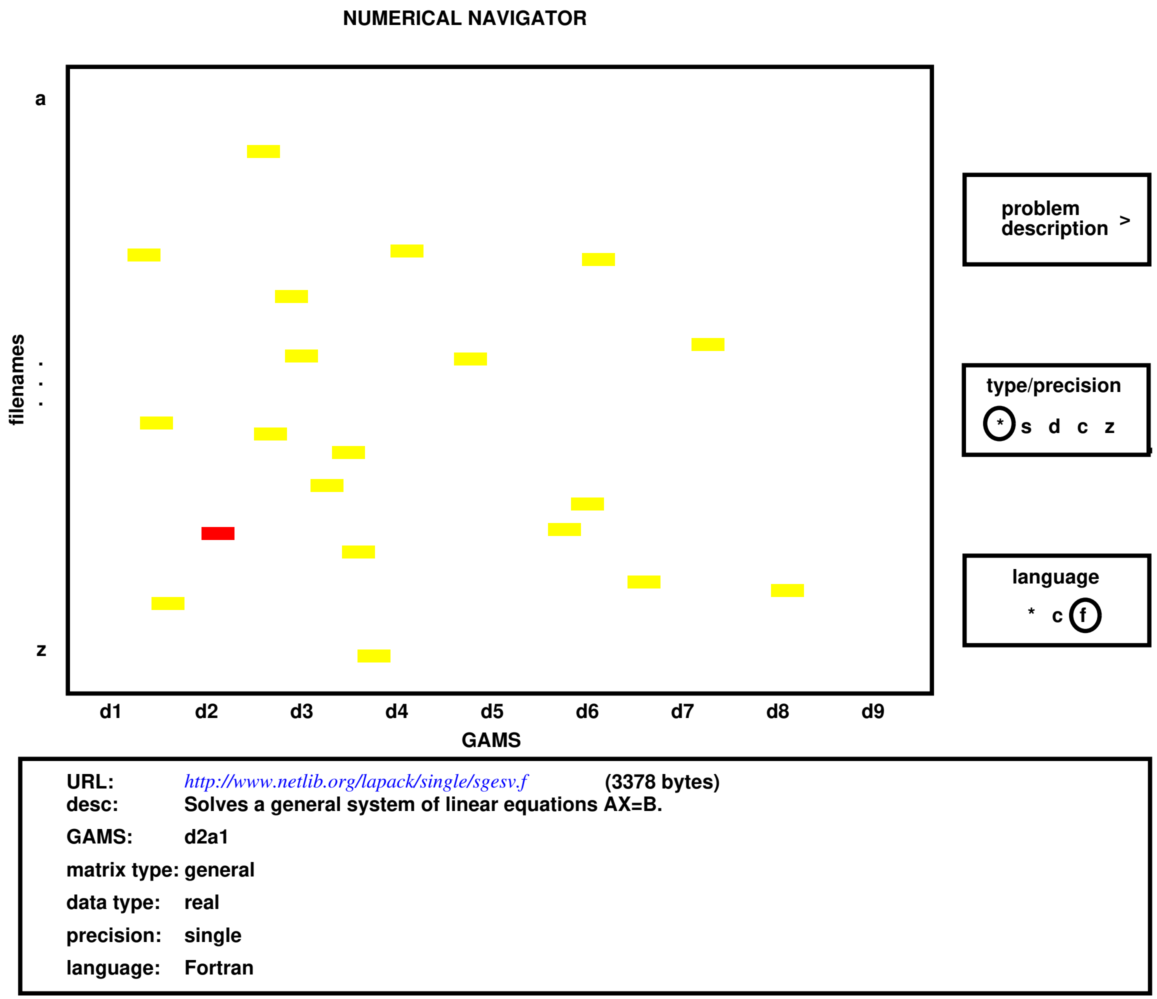
- **Outreach and technology transition**
  - **Educational activities aimed at the user community**
  - **Fostering technology development by industry**
- **Working on standardization within WWW community**
  - **member of RIG, IETF, IESG, WWW consort**

Numerical Navigator

- Applet for visualizing contents of software collection on single screen
- Manipulate display using buttons, slides, and pull-down menu
- Point and click for more information or to download software
- Prototypes in Tcl/Tk and Java
- Prototype for linear algebra software but applicable to other collections

Milestones:

- May 1 - release stable Java version
- Sep 30 - extensions to other collections
- 1997 - release toolkit for collection managers to build their own navigators



NUMERICAL NAVIGATOR
a
filenames
z
d1
d2
d3
d4
d5
d6
d7
d8
d9
GAMS
problem description >
type/precision
* s d c z
language
* c f
URL: http://www.netlib.org/lapack/single/sgesv.f (3378 bytes)
desc: Solves a general system of linear equations AX=B.
GAMS: d2a1
matrix type: general
data type: real
precision: single
language: Fortran

ApproxWizard

- Applet that helps user select an approximation code
- Interacts with user by doing calculations on sample user data sets
- Calculations done on client and/or remote servers
- Prototypes being developed in Java and Limbo

Milestones:

- April 2 - working demo
- rest of year - evaluate
  - Effectiveness of Java vs. Limbo approaches
  - How to extend to other application areas

cartoon from eric

GAMS II

- **Successor to popular GAMS taxonomy for mathematical software**
- **Combination of taxonomy and thesaurus approaches**
- **Unique set of features for each class**
- **Special reformulation links for mapping between problem classes**
- **Basis for domain-specific expert extensions that help user discriminate between problem-solving modules**

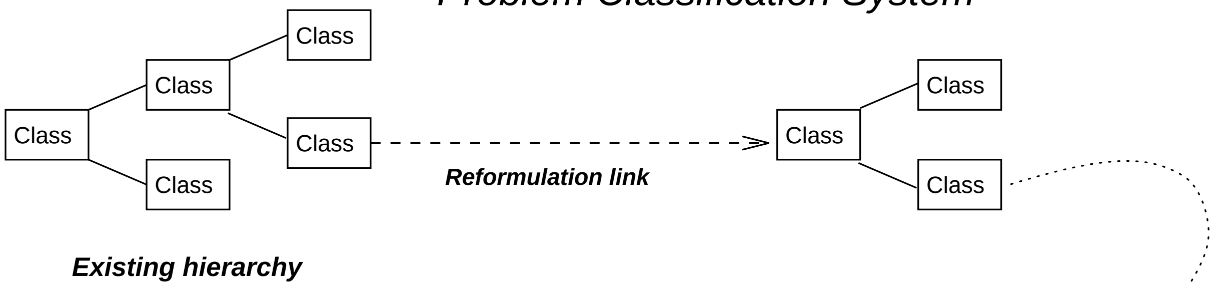
Milestones:

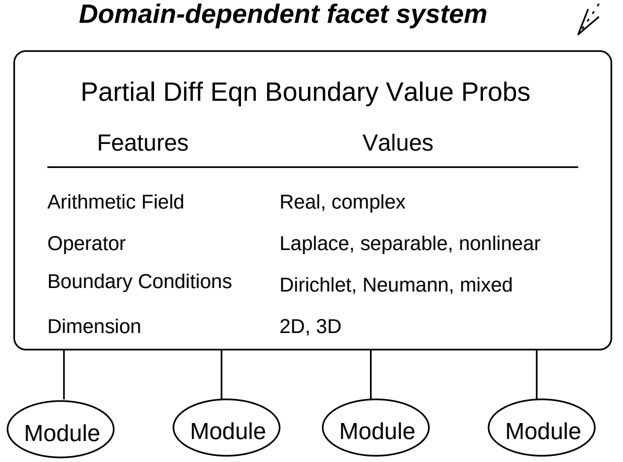
- **Dec 31, 1996 - report on 2nd generation GAMS classification technology**
- **1997 - develop GAMS client in Java with expert advisory extensions**

Guide to Available Math Software

Goal : Develop metadata, browser to support intelligent reference interviews

Problem Classification System





New in GAMS II

- revise, update, coarsen
- problem reformulation
- domain-dependent facets

Plans

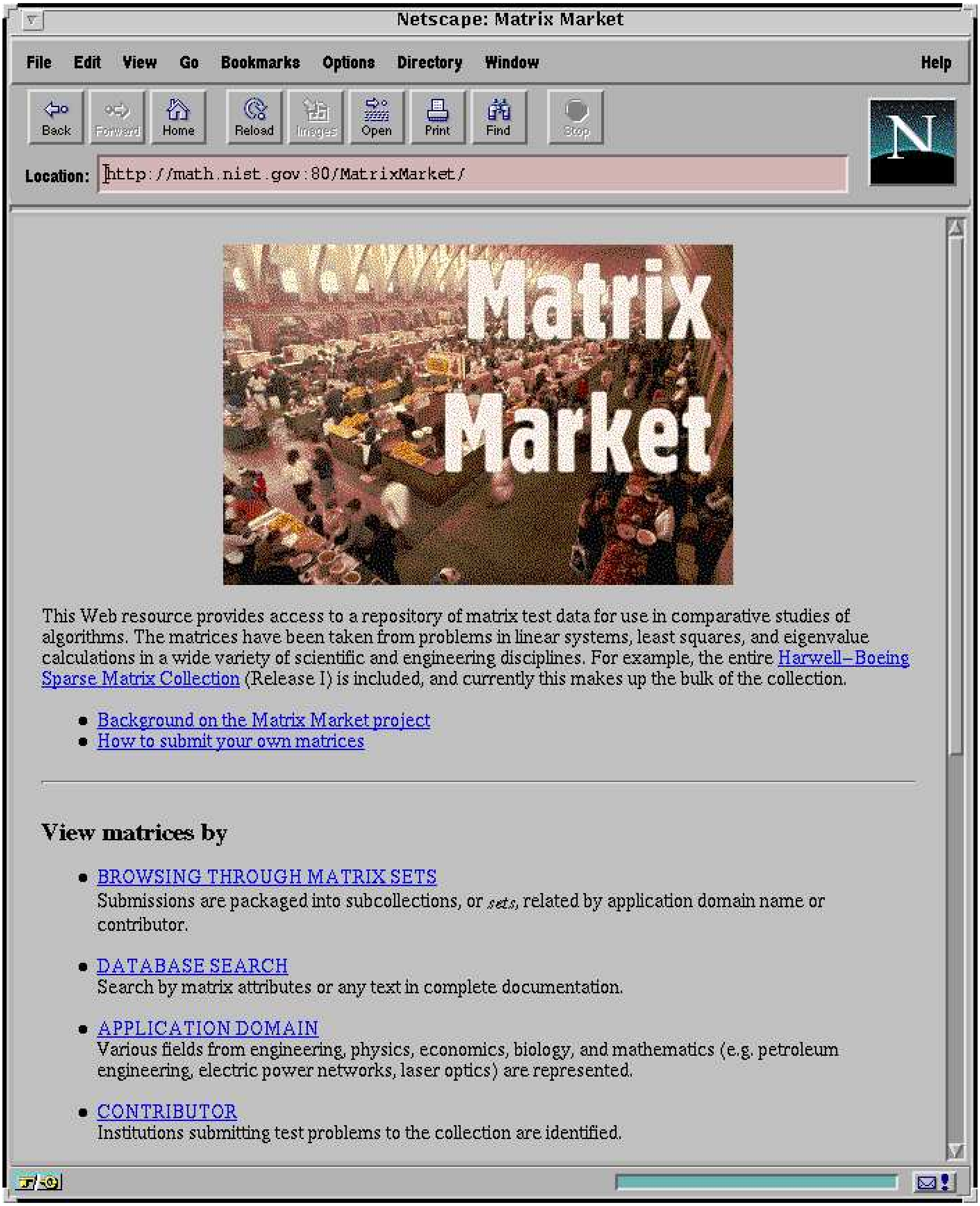
- prototype JavaGAMS (9/96)
- GAMS II design (12/96)
- intelligent JavaGAMS (1997)

MatrixMarket

- **Repository of matrix test data**
- **For comparing algorithms for linear systems, least squares, and eigenvalue calculations**
- **Problems from various fields represented (physics, economics, biology, engineering, etc.)**
- **Browse and search interfaces**
- **User contributions welcome**

Milestones:

- **Mar 15 - release Web interface (http://math.nist.gov/MatrixMarket)**
- **Sep 30 - report on how to extend functionality of MatrixMarket using Java**


Netscape: Matrix Market
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Location: http://math.nist.gov:80/MatrixMarket/
Matrix Market
This Web resource provides access to a repository of matrix test data for use in comparative studies of algorithms. The matrices have been taken from problems in linear systems, least squares, and eigenvalue calculations in a wide variety of scientific and engineering disciplines. For example, the entire Harwell–Boeing Sparse Matrix Collection (Release I) is included, and currently this makes up the bulk of the collection.
Background on the Matrix Market project
How to submit your own matrices
View matrices by
BROWSING THROUGH MATRIX SETS
Submissions are packaged into subcollections, or sets, related by application domain name or contributor.
DATABASE SEARCH
Search by matrix attributes or any text in complete documentation.
APPLICATION DOMAIN
Various fields from engineering, physics, economics, biology, and mathematics (e.g. petroleum engineering, electric power networks, laser optics) are represented.
CONTRIBUTOR
Institutions submitting test problems to the collection are identified.

Program Builder

- Agent-based approach
- Select self-consistent set of software modules meeting user-specified criteria
- Customize, install, and test software on user's platform
- Performance evaluation and tuning
- Automatically notify user of updates

Milestones:

- Dec 31, 1996 - Report describing program builder concept and requirements (including security)
- 1997 - Prototype implementation using NHSE as testbed

Keith on RCDS

Safe Execution Environment

- System within which untrusted code from remote sources may be executed locally
- Varying degrees of trust, depending on authentication of code origin
- Program interpretation and run-time support
- Relocation and communication services
- Tradeoffs between
  - security and functionality
  - security and performance

Security Issues

Threats:

- **Trojan horses**
- **Viruses and worms**
- **Eavesdropping**
- **Denial of service**

Proposed security measures:

- **Restricted access to files and commands**
- **Asking user approval before accessing files or sending email**
- **Strong type checking**
- **Untrusted envirnoments calling operations exported from trusted environment**
- **Public-key encryption for authentication and integrity**
- **Enforcement of resource limits**

Security Profile

Thorough analysis of security aspects of environment that is to be secured.

Includes:

- **Realistic usage scenarios**
- **Analysis of threats and vulnerabilities**
- **Security objectives**
- **Requiring for meeting objectives**

Possible scenarios:

- **Agent-based collaborative technologies**
- **Intelligent information retrieval**
- **Web-based high performance computing**

Safe Execution Environment Milestones

- Security profile
  - Mar 15 - draft of scenarios document
  - May 1 - comments and feedback on scenarios document
  - Sep 1 - vulnerability analysis and security objectives
  - Dec 31 - draft of entire security profile
- 1997 - Release of prototype server safe execution environment

Collaborations:

- NCSA (Ken Rowe)
- OSF Java project (Mike Weiss)
- Dartmouth Agent TCL (Bob Gray)
- IBM Itinerant Agents (David Chess)

Repository Interoperability

- DLib Working Group on Repository Interactions
- Reuse Library Interoperability Group (RIG)
  - Participating in Web-based interoperability experiment using SGML and HTML binding of RIG BIDM
  - Chair of newly formed technical committee on intellectual property rights labeling
- National HPCC Software Exchange (NHSE)
  - Uniform interface to domain-specific repositories
  - Uniform approach to intellectual property rights protection
  - Repository in a Box toolkit for repository maintenance and interoperation

RIG BIDM

RIG stands for Reuse Library Interoperability Group (http://www.rig.org/

Basic Interoperability Data Model (BIDM)

- Minimal set of information about assets that reuse libraries should be able to exchange
- Class hierarchy with attributes and relationships
- Approved as IEEE Standard P1420.1 in December 1995
- Basis of current RIG interoperability experiments

Repository Interoperability Milestones

DLib Milestones:

- **Mar 11-12 - attend DLib Repository Interactions Working Group meeting**

RIG Milestones:

- **Mar 15 - ballot RIG Proposed Standard for SGML and HTML bindings of RIG BIDM**
- **Jul 15 - start IEEE ballot process for SGML and HTML bindings**
- **May 15 - draft of RIG BIDM annex for IPR labeling**
- **Fall 96 - ballot RIG Proposed Standard for IPR labeling**

NHSE Repository in a Box

- Resource center for software repository issues
  - security
  - intellectual property rights
- Customizable toolkit
  - Publishing tool for creating and maintaining software catalog records
  - Tools for creating and using classification schemes and controlled vocabularies
  - Local search setup and link to global NHSE search interface

NHSE Interoperability Milestones

- **Apr 1 - identify and establish contact with candidate domain-specific repositories**
- **Apr 1 - release NHSE software cataloging standard (superset of RIG BIDM)**
- **Jun 1 - release Repository in a Box**
- **Oct 1 - bring prototype repositories online**
- **Oct 1 - integrate repository management with RCDS**
- **Dec 1 - common search interface to interoperating repositories**
- **Dec 31 - recruit additional discipline-oriented repositories**

Distributed Searching

- Construction of high-quality databases
- Intelligent gathering from multiple sites
- Meaningful summary of database contents
- Intelligent database selection
- Leverage high-quality manually constructed information
- Combination of multiple search results
- Relevance feedback across multiple databases

Latent Semantic Indexing (LSI)

- Low-rank approximation to term-document matrix
- Retrieval of relevant documents based on statistical word co-occurence
- Research areas
  - Efficient updating and downdating for dynamic collections
  - Index distribution and partitioning for parallel and distributed processing
  - User interfaces for experts and end users
- Prototype at http://www.netlib.org/cgi-bin/lsiBook/

Searching Milestones

- **Apr 1 - design of distributed search architecture**
- **Sep 1 - prototype interactive LSI-based gathering tool**
- **Sep 1 - LSI interface to Harvest Broker**
- **Dec 1 - evaluate LSI gatherer and indexer using NHSE testbed**
- **1997 - interfaces for multiple database search**