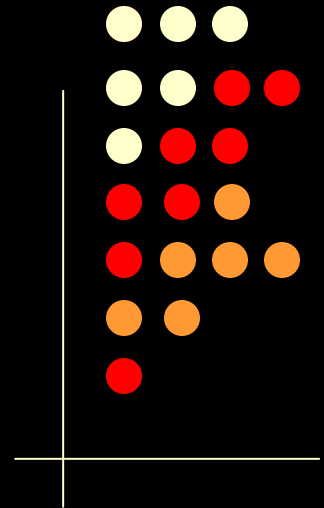




An Interactive Digital Library of Algorithmic Knowledge

*and other Information Science projects at
Cornell*



April 14, 2003 / UC Berkeley

Stuart Allen

Mark Bickford

Robert Constable

Jason Hickey (Cal Tech)

Vitali Khailine (Wyoming)

Richard Eaton

Christoph Kreitz

Lori Lorigo

Jim Caldwell (Wyoming)

Topics

- The Office of Naval Research (ONR) Digital Library Project
 - Goals, Design, Implementation
- Other Information Science Projects
 - National Science Digital Library, NSDL
 - Ranking Function
 - Modeling Bursts of Activity on the Web



ONR Digital Library Project

" ... to create a digital library of algorithms and constructive mathematics useable for program and software construction."



ONR Digital Library Project

" ... to create a digital library of algorithms and constructive mathematics useable for program and software construction."

Goals

- Semantics-based **interactive** infrastructure
- Create, collect and organize **formal content**



Meaning of Formal

"... having a precise meaning or objective criteria of correctness ... ideally computer verifiable, based on syntactic form."



Formalization I

Scheduling Application Formalization

Specify components and behavior

Use mathematical definitions and theorems

Reveals assumptions, increases **assurance**, and **optimizes** task assignments.

Doing A then B then A is the same as doing A.



Formalization II

- Halting problem, Alan Turing - informal

Does a given algorithm ever halt when executed on given arguments?

- no method to solve it for all inputs



Formalization II

- Halting problem, Alan Turing - informal

Does a given algorithm ever halt when executed on given arguments?

- no method to solve it for all inputs

- Formalization

- Create formulas to express halting

- Mechanically prove no method exists for all inputs

$\neg \exists h: \bar{A} \rightarrow \text{Bool}. \forall n: \bar{A}. h(n) = \text{true iff } n \text{ halts}$ ✓



ONR Digital Library Project

" ... to create a digital library of algorithms and constructive mathematics useable for program and software construction."

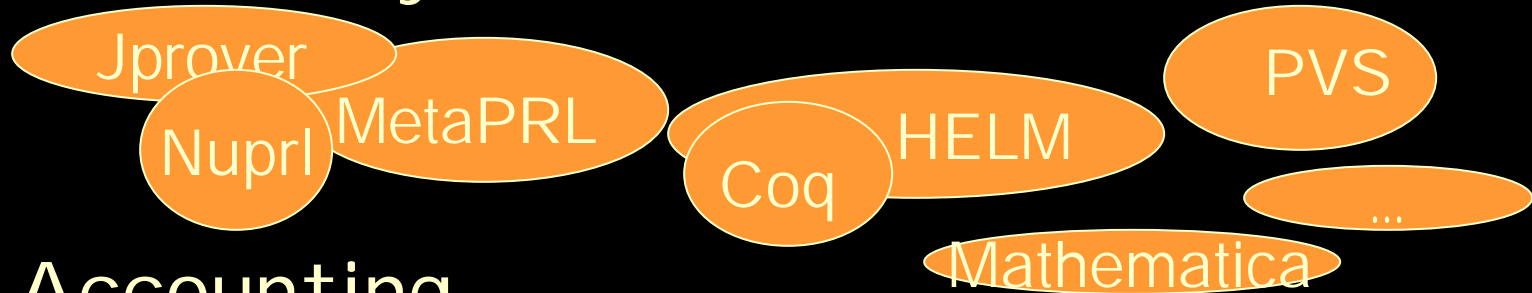
Benefits

- Accelerate **discovery**
- **Access** to formal content
- **Resource** for reliable software and education



Strategy for Meeting ONR Goals

- Community of contributors



- Accounting

- differences and limitations amongst systems, different meanings

- Semantics-based library services

Challenges and Problems

- *Small* and **disconnected** community
- Formal proving is still **hard work**
 - Large expansion in size
 - Shallow resource base
 - Demanding skill set



Challenge Citation

*"PVS is a large and **complex system** and it takes a long while to learn to use it effectively. You should be prepared to invest **six months** to become a moderately skilled user (less if you already know other verification systems, **more if you need to learn logic or unlearn Z**). "*

- PVS website at SRI



Formal Digital Library (FDL)

FDL Project Home - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.nuprl.org/FDLproject/> Go



**MULTIDISCIPLINARY
UNIVERSITY
RESEARCH
INITIATIVE**
a DoD/MURI Project with Cornell University,
CalTech, and the University of Wyoming

"Information Intensive Critical Infrastructure Protection"

FDL Project

Formal Digital Libraries



sponsored by the
Office of Naval Research

Introduction

ONR/MURI Project People FDL Content FDL Prototype

Project Briefings Publications Algorithms Community

Latest News

04-08-03 **PVS collection through Graphs mounted on Web**
Part of the standard [PVS](#) collection of theories (Prelude through Graphs) has been mounted as HTML with links representing references between PVS entities.

03-11-03 **FDL Editor access available via VNC**
VNC (Virtual Network Computing) access to the FDL editor is now available. A password and server name are returned to you which you can then use with your own copy of [VNC](#). Go to [FDL Online](#) to view.

03-11-03 **Search and Lookup FDL Content in XML format**
Two queries to the FDL are now available online for demonstration purposes: a name search and a lemma lookup. Go to [FDL Online](#) to view. The data is returned in XML format, a standard, to allow integration with other applications and tools.

02-20-03 **FDL Navigator Manual online**

"The world is engaged in a grand scientific and technological enterprise to build a global information resource. Creating this global resource will be one of the great achievements of information technology. Our research on applied logic, formal methods and automated reasoning will make the emerging information resource more capable -- first by providing a basis for semantic processing of information and for a logical accounting of its structure, and second by including among the resources an interactive digital library of formal computational mathematics. Such a library will bring into being a formal forum that will connect experts and practitioners together in building reliable software systems, educating the information technology work force, empowering the lay scientist, and in nucleating the creation of a broader open library of formally grounded knowledge."

*Prof. Robert Constable, Project PI
Cornell University*



Objects

F			
O	rules	definitions	algorithms, code
R			
M	conjectures	theorems	specifications
A			
L	inferences	certificates	proofs, partial proofs



Objects

F
O
R
M
A
L

rules

definitions

algorithms, code

conjectures

theorems

specifications

inferences

certificates

proofs, partial proofs

I
N
F
O
R
M
A

displays

articles

documentation



Objects

F
O
R
M
A
L

rules

definitions

algorithms, code

conjectures

theorems

specifications

inferences

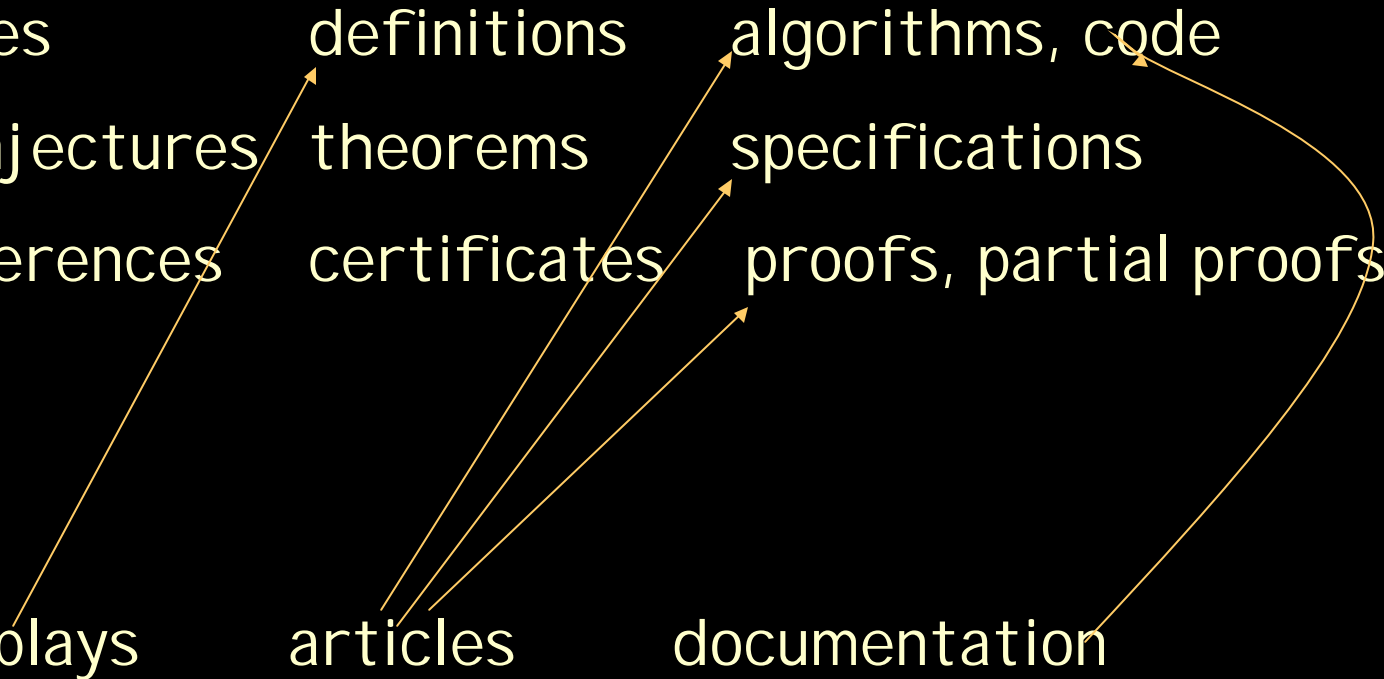
certificates

proofs, partial proofs

displays

articles

documentation

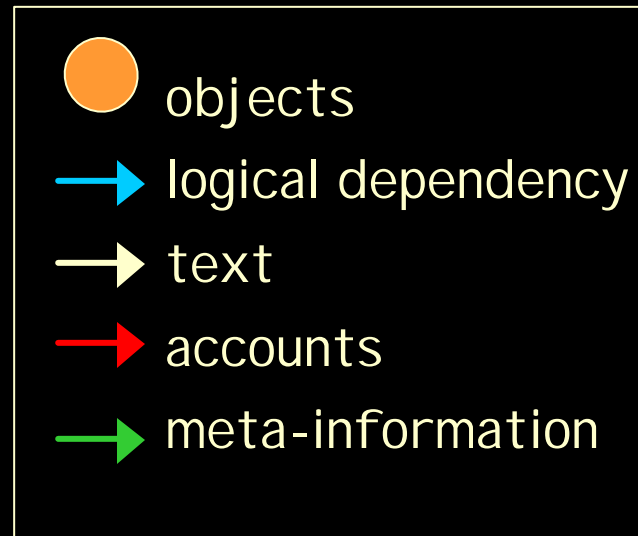
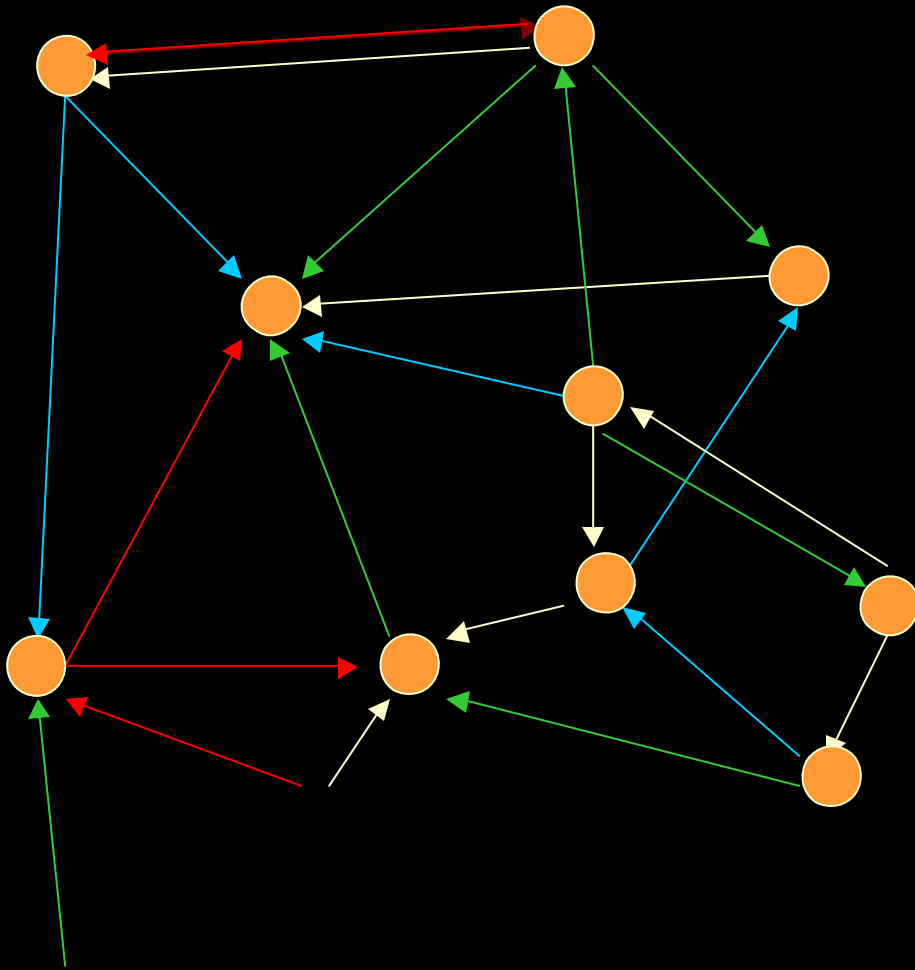


I
N
F
O
R
M
A
L



Information Graph of the FDL

Objects have interdependencies and informative links



Proof Example

proof

```
❏ *- PRF : pos_length @edd.lori_10_13 @nuprl4.cs.corne
* top
∀A:U. ∀l:A List. ((¬(l = [])) ⇒ (||l|| ≥ 1))
BY aux_auto(RepeatMFor 2 (D 0) THENM D 2)
* 1
1. A : U
⊢ (¬(l = [])) ⇒ (||l|| ≥ 1)
BY auto(D 0 THENM D (-1))
* 2
1. A : U
2. u : A
3. v : A List
⊢ (¬([u / v] = [])) ⇒ (|[u / v]| ≥ 1)
BY AbReduce 0 THEN aux_auto(D 0)
* 2 1
4. ¬([u / v] = [])
⊢ (||v|| + 1) ≥ 1
BY InstLemma `non_neg_length` ['A'; 'v'] THEN Auto
```

← article

→ accounting

→ theorem

```
∀A:U. ∀l:A List. (||l|| ≥ 0)
```

Accounting Archive

FDL performs archival functions...

- Automath system Auto QE checked the following formalization of **Landau's Grundlagen** (February 22, 1974)
- EQP proved the **Robbins Conjecture**, an open problem for mathematicians since 1933 (October 10, 1996)
- JProver proves "the Butler did not do it" in the this logic-based "**who-done-it**" problem (October 13, 2000)
- CLI presents a formal model of the **Java Virtual Machine** (May 13, 1997)
- ACL2 verifies **Y2K conversion rules** (March 21, 2000)



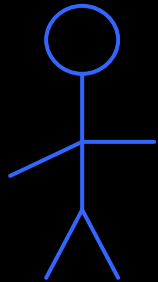
Basic FDL Operations

- **search** by name
- **lookup** object by id
- create an object (empty)
- save
- undo
- "activate" object
- "deactivate" object



Translation Operations

FDL performs translations on collections, or subgraphs.

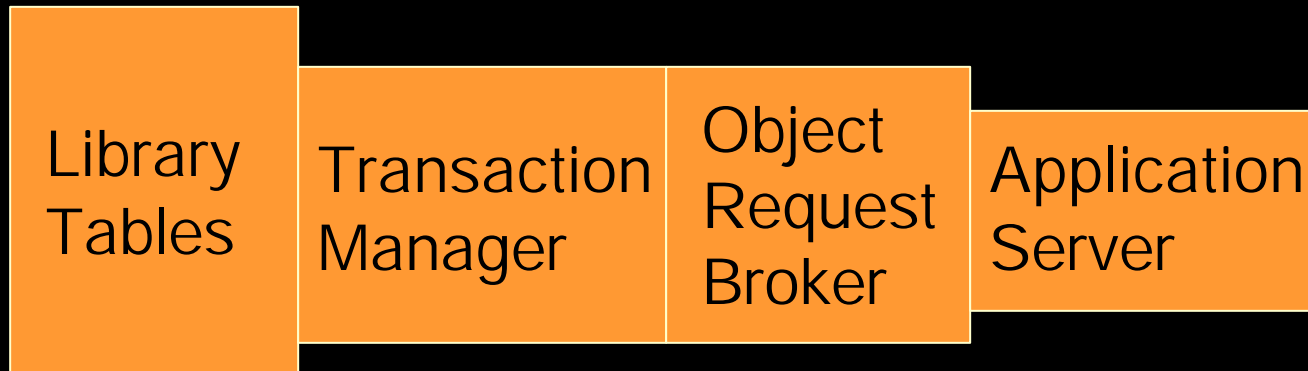


Give me the smallest *closed* set with respect to logical references of objects that contains the proof of Total Order in Distributed System, D.

Give me the largest *closed* set with respect to textual links that contain information about prime numbers.

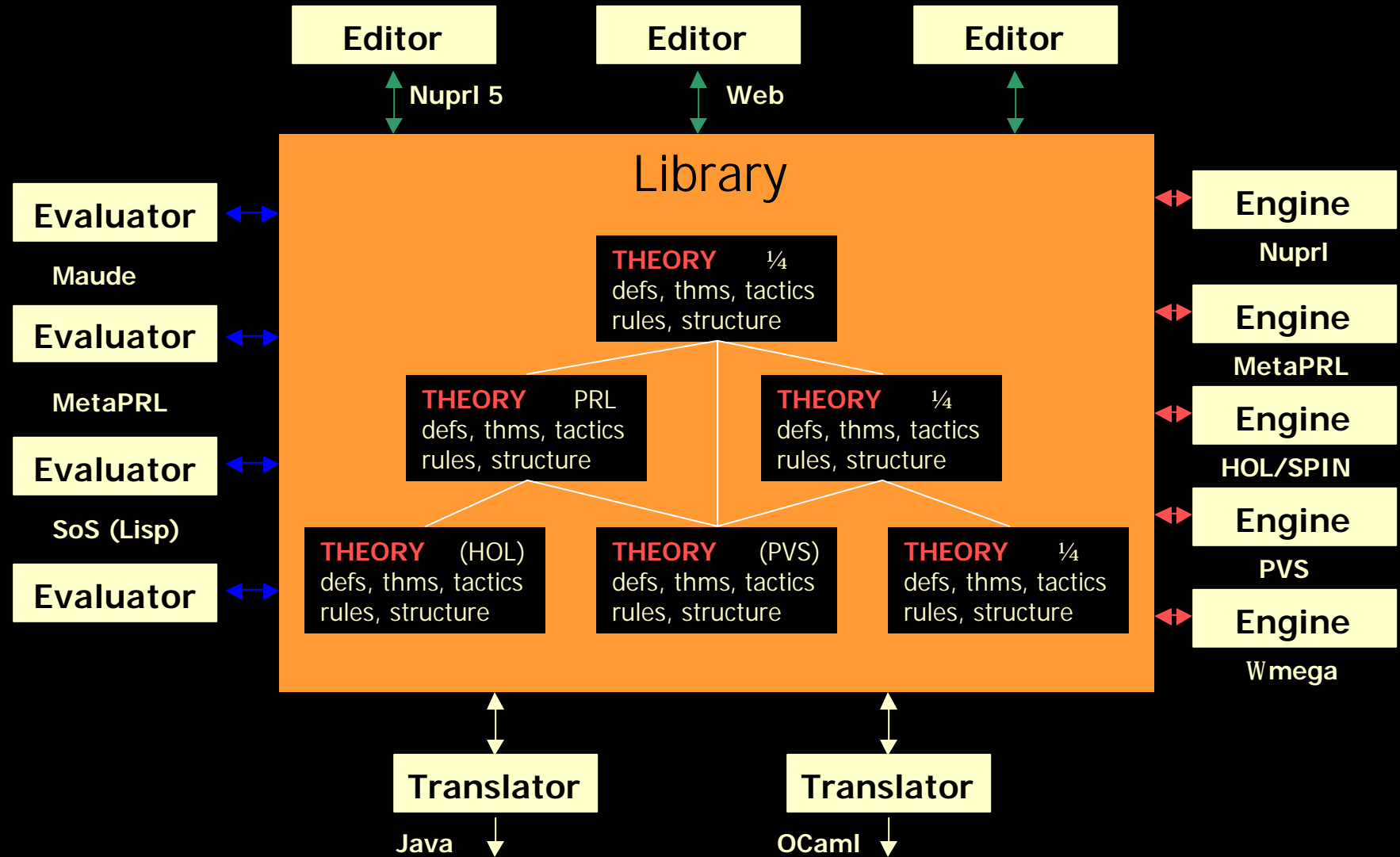


Implementation



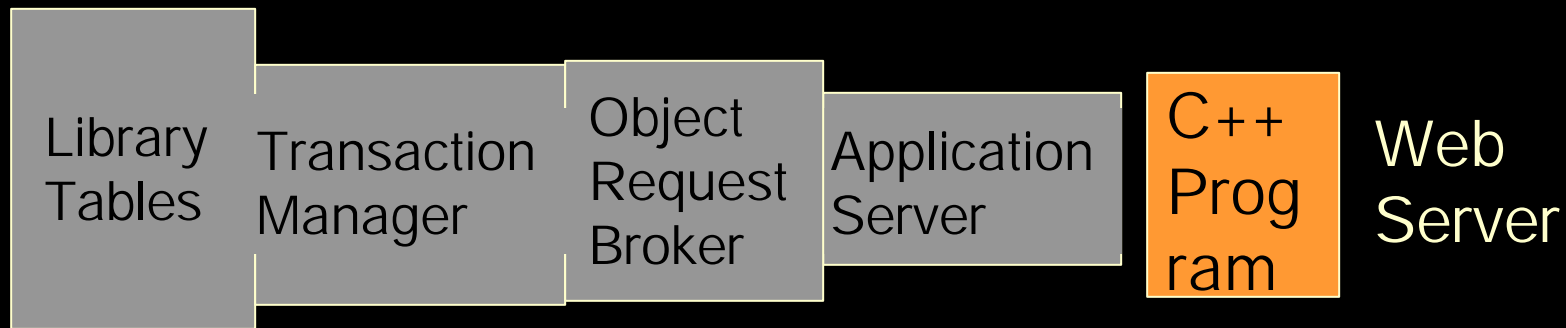
- LI SP/ML based system
 - 6,000 functions
 - 62,000 lines code
 - 22,000 lines comments
- Adapted from Logical Programming Environment and Nuprl

Architecture



FDL Online Demo

- **search** by name
- **lookup** object by id



- **vnc** into editor

FDL Online Demo

The screenshot shows a web browser window with the following elements:

- Browser address bar: Search, Sign In, My Yahoo!, News, Health, Yahoo!
- Page Title: **FDL XML Data**
- Header: *Building Interactive Digital Libraries of Formal Algorithmic Knowledge*
- Text: *Returns a list of object identifiers matching your query, i.e. enter "list" for a list of identifiers of lemmas in the FDL about lists.*
- Form: **Name Search:**
- Text: Click to submit your search.
- Text: *Returns the lemma with the given object identifier, which you can get by using the name search above. i.e. try "%NOr10" to get the XML format of a list object.*
- Form: **Lemma Lookup by ID:**
- Text: Click to submit your search.

```
<fdl:term>
<fdl:param_list>
  <fdl:param>
    <fdl:paramtype>t</fdl:paramtype>
    <fdl:paramvalue>implies</fdl: ...>
  </fdl:param>
</fdl:param_list>
<fdl:term_list>
  <fdl:term>
    <fdl:param_list>
      <fdl:param>
        <fdl:paramtype>t</fdl:paramtype>
        <fdl:paramvalue>equal</fdl: ...>
      </fdl:param>
    </fdl:param_list>
  ...

```

Current Contents

- Nuprl and PVS Content
 - 422 theories (collections)
 - 1699 definitions
 - 7351 theorems
 - 3867 proofs
- MetaPRL (Cal Tech)
- In Progress – import many more in xml and mathml formats



Is Data Format Sufficient?

- **Term: Operator (Op) x Term List**
- XML (DTD for HELM)
- OMDoc
- MathBus
- Ascii (compressed and uncompressed)

Op = Opname{params}

op(Term List)

add{}(x;y)

x=natural{1}()

y=mult(z;r)

Contribution Methods

System	Method
Jprover	MathBus
C++ Program	Ascii uncompressed
MetaPRL	MathBus
OMDoc	XSLT
PVS	LISP structures
Larch from ATC	In progress



Related Work

- HELM (Hypertextual Electronic Library of Mathematics) - <http://helm.cs.unibo.it/>
- MathWeb - <http://www.mathweb.org>
- Mathematical Knowledge Management (MKM) - <http://imps.mcmaster.ca/na-mkm-2002/presentation>
- Wolfram Research MathWorld - <http://mathworld.wolfram.com/>
- Living Book - <http://www.slicing.de/books>



Questions

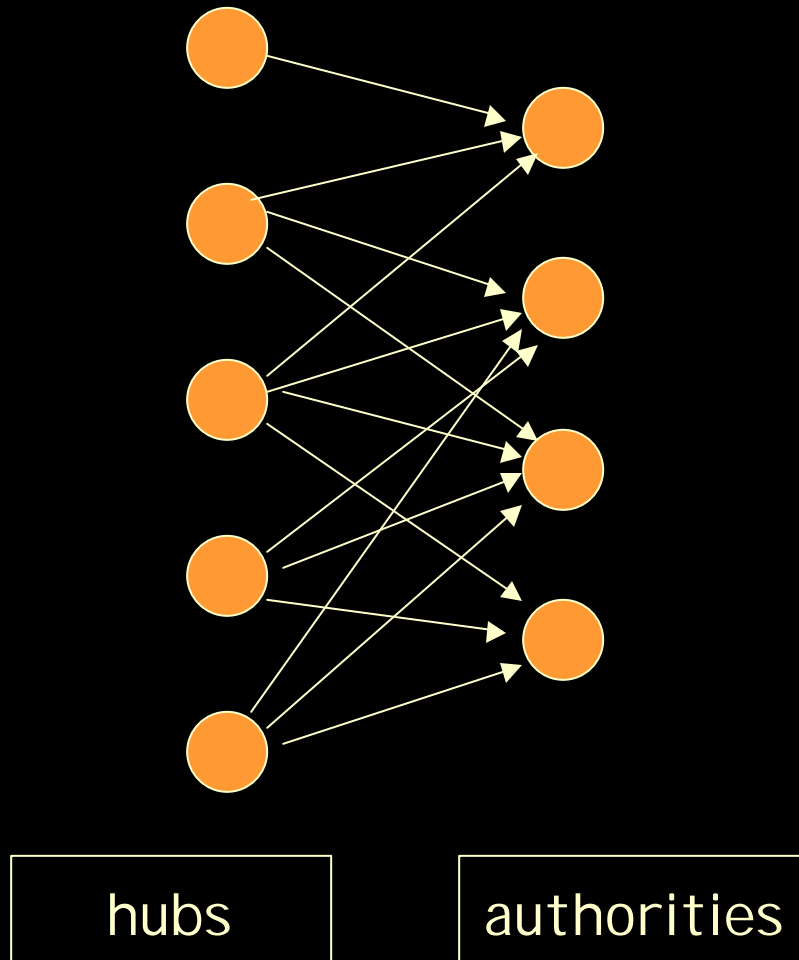
- What communities to target?
- How to most effectively search?
- Data format justification?



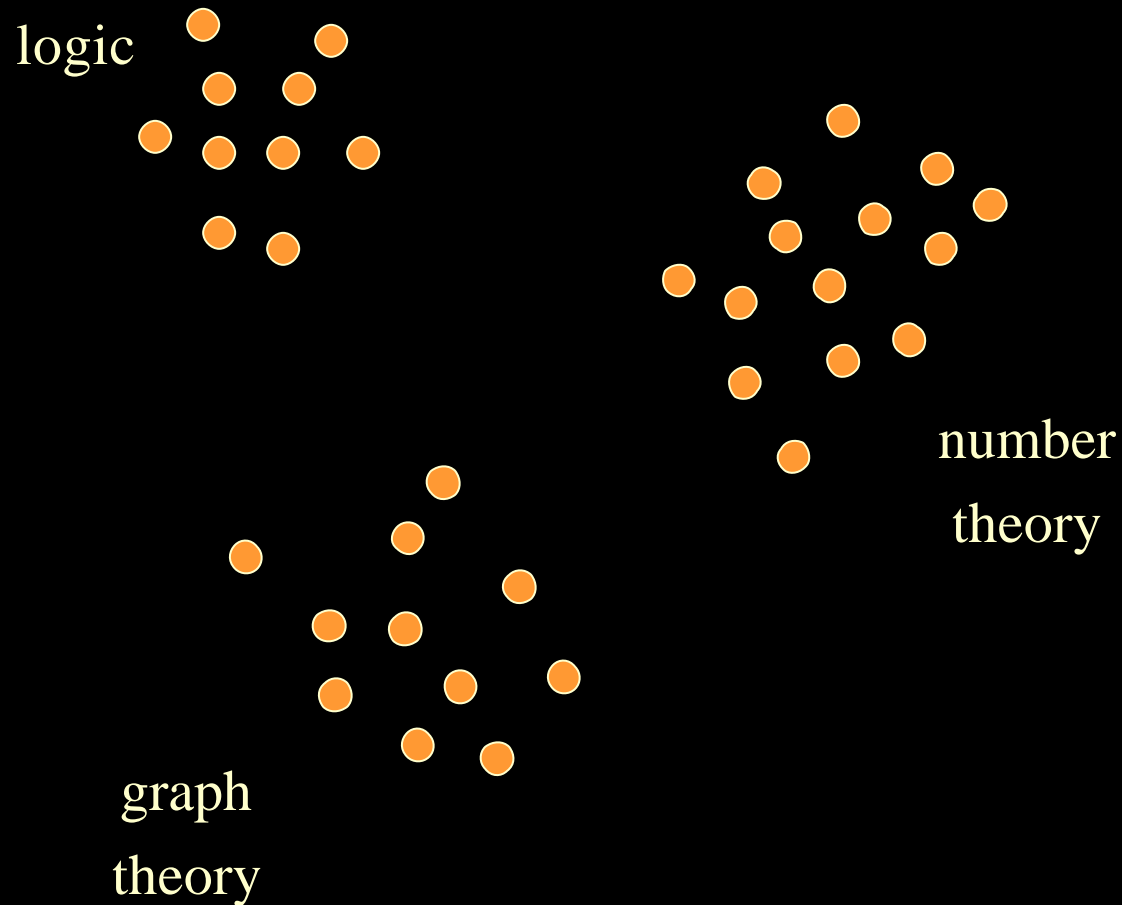
Questions

- What communities to target? Education, Code
- How to most effectively search? dependency
- Data format justification? MathML future?
- What usability studies and evaluations can we perform?
- Can we learn from semantic knowledge?

Learning from the Information Graph



Classifying by Eigenvectors



Topics

- The Office of Naval Research (ONR) Digital Library Project
 - Goals, Design, Implementation
- Other Information Science Projects
 - National Science Digital Library, NSDL
 - Ranking Function
 - Modeling Bursts of Activity on the Web



NSF NSDL

digital library: "A managed environment of multimedia materials in digital form, designed for the benefit of its user population, *structured* to facilitate access to its contents, and equipped with aids to navigate the global network ... with users and holdings totally distributed, but *managed as a coherent whole*."

--Mel Collier, *International Symposium on Research, Development, and Practice in Digital Libraries 1997*



NSF NSDL

digital library: "A managed environment of multimedia materials in digital form, designed for the benefit of its user population, *structured* to facilitate access to its contents, and equipped with aids to navigate the global network ... with users and holdings totally distributed, but *managed as a coherent whole*."

--Mel Collier, *International Symposium on Research, Development, and Practice in Digital Libraries 1997*

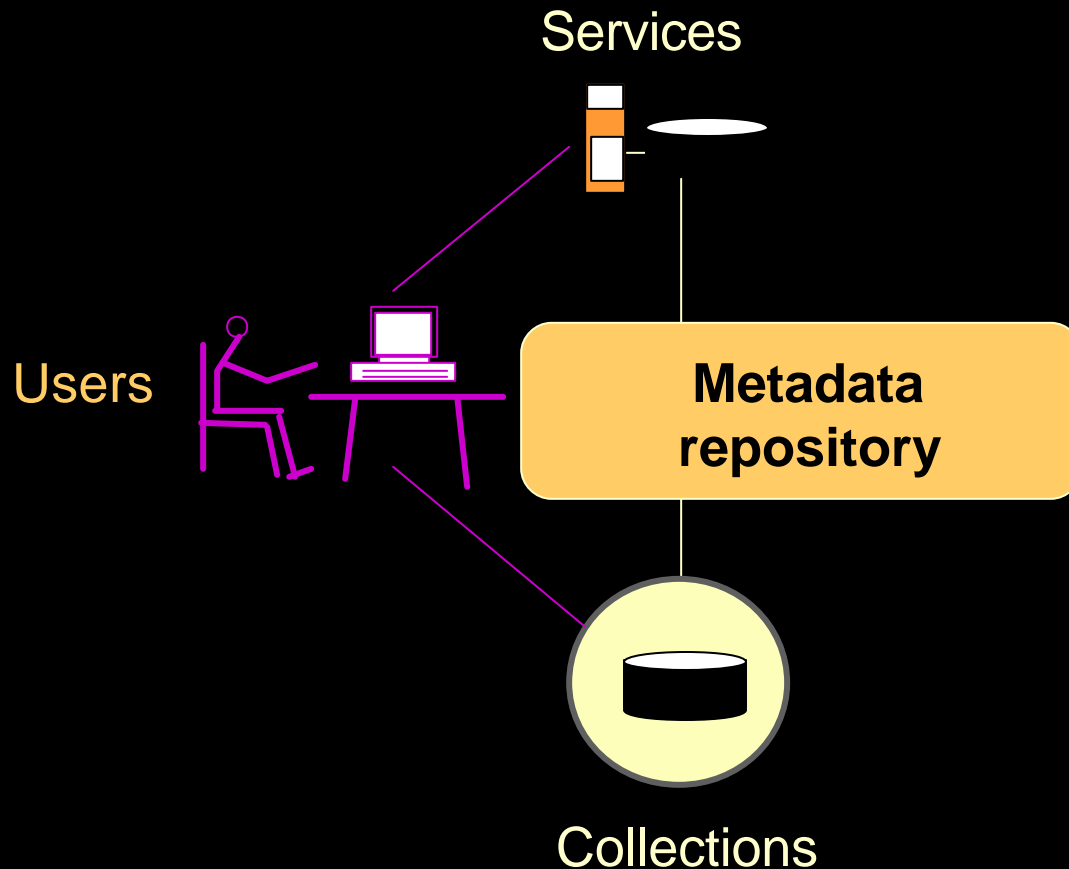


Bill Arms (director), Naomi Dushay, Carl Lagoze, Sandy Payette, Carol Terrizi, Vicky Weissman, <more>



Metadata Repository

The Open Archives Initiative Protocol for Metadata Harvesting: Los Alamos, Nasa, and Cornell



Cornell's NSDL Portal

The screenshot shows a Netscape 6 browser window displaying the NSDL Portal Prototype. The address bar shows the URL <http://nsdlib.nsdlib.cornell.edu/nsdl/portal/index2.html>. The page features a navigation menu with links for 'SITE for Students', 'SITE for Teachers', 'My SITE', 'News', 'Tours', 'Tools', and 'Science Pictures'. A search bar is located in the top right corner. The main content area is divided into three sections: 'EARTH SCIENCE PICTURE OF THE DAY (EPOD)', 'PALEONTOLOGICAL RESEARCH INSTITUTION', and 'THE ALSOS DIGITAL LIBRARY FOR NUCLEAR ISSUES'. Each section includes a photograph and a brief description. The 'EPOD' section features a photograph of the Earth from space, captioned 'Earthshine!'. The 'PALEONTOLOGICAL RESEARCH INSTITUTION' section features a photograph of a fossil, captioned 'Eurypterid'. The 'THE ALSOS DIGITAL LIBRARY FOR NUCLEAR ISSUES' section features a photograph of a document. The page also includes a sidebar with a navigation menu and a footer with contact information and logos for NSF and Cornell University.

NSDL Portal Prototype - Netscape 6

File Edit View Search Go Bookmarks Tasks Help

[http://nsdlib.nsdlib.cornell.edu/nsdl/portal/index2.html](#)

Home Bookmarks

Site for science Register/Login SEARCH:

HOME SITE for Students SITE for Teachers My SITE News Tours Tools Science Pictures

INDEX OF
Home

ABOUT SITE FOR SCIENCE

- ▶ WELCOME
- ▶ GETTING STARTED
- ▶ BROWSER COMPATIBILITY
- ▶ ABOUT NSF-NSDL PROGRAM
- ▶ ABOUT CORNELL 'SITE'
- ▶ NSDL WHITEBOARD REPORT
- ▶ COLLECTIONS

NEWS

- ▶ NSF SCIENCE NEWS
- ▶ EARTH SYSTEM SCIENCE EDUCATION NEWS HEADLINES
- ▶ NAVGLOBE

TOPICS

- ▶ ERIC TOPICS
- ▶ GEM SUBJECTS
- ▶ LIBRARY OF CONGRESS SUBJECT HEADINGS
- ▶ DEWEY DECIMAL CLASSIFICATION

EXHIBIT
Items of interest from the SITE library.

EARTH SCIENCE PICTURE OF THE DAY (EPOD)

The Earth Science Picture of the Day (EPOD) collects and archives photos, imagery, graphics, and artwork with short explanatory captions and links exemplifying features within the Earth system.

[More information](#)

Browse EPOD by general topics
Over forty general topics lead to EPOD images used since September 2000.

Earth Science Image Links
EPOD maintains a short but quality list of other earth science image sites.

Take a look!
Take a look at a photograph of a sun pillar ...

PALEONTOLOGICAL RESEARCH INSTITUTION

PRI is Ithaca, New York's Museum of the Earth, and a resource for fossil enthusiasts and educators the world over. We're home to the US's 7th largest collection of fossils, and 5th largest collection of type specimens. Our educational programs reach teachers and students from around the country.

[More information](#)

Collections at PRI
PRI has a world-class collection.

Exhibits at PRI
Get information about what is currently being displayed.

Museum of the Earth
A new public natural history exhibit facility for one of the nation's largest and finest fossil collections.

THE ALSOS DIGITAL LIBRARY FOR NUCLEAR ISSUES

The Alsos digital library is a Web-based collection of references to resources that offer a broad, balanced perspective of topics relating to the origin, function, and legacy of the atomic bomb.

The original Alsos Project
How close were the Germans to developing the atomic bomb?

Project Information
How and why the Alsos Project was started.

"SITE for Science" is a core integration system prototype for a digital library. The National Science Foundation provides funding for this project which is part of the National Science, Mathematics, Engineering, and Technology Education Digital Library Program.

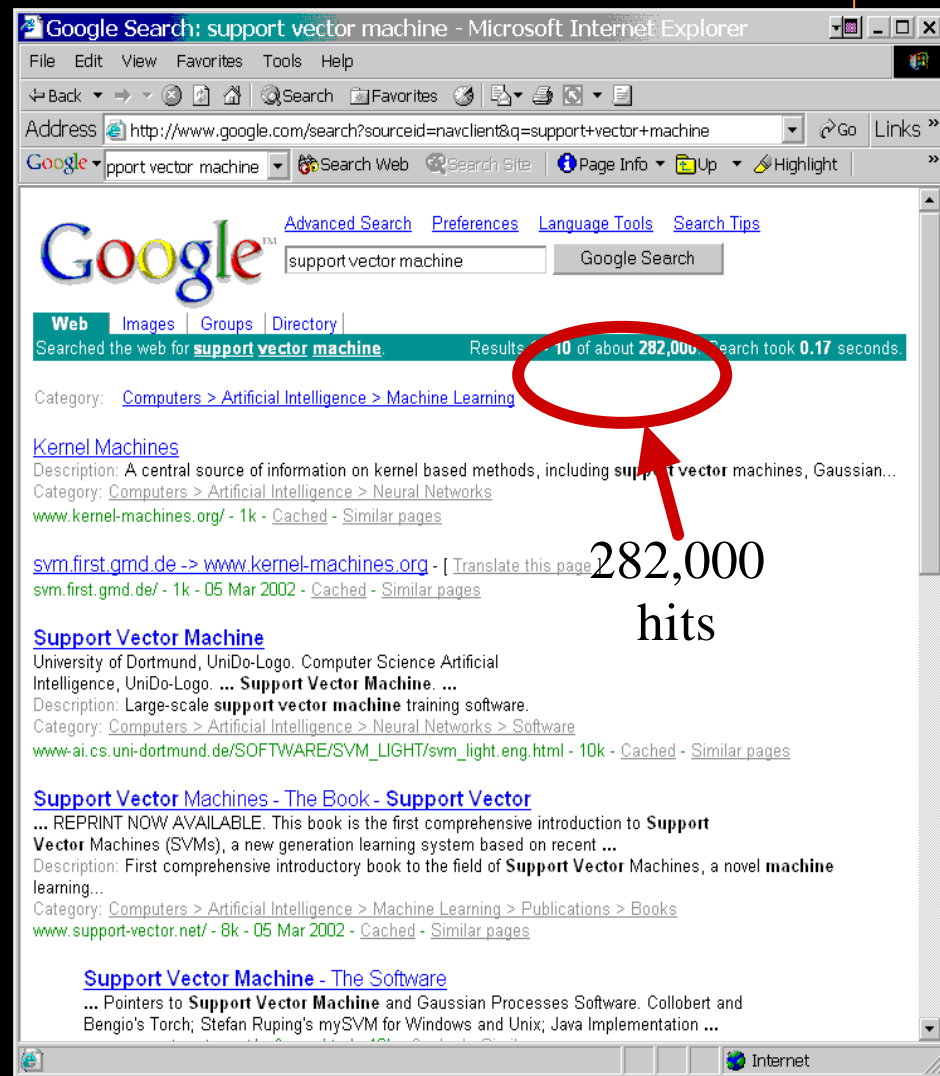
ABOUT CONTACT NSF CORNELL

Document Done (7.501 secs)



Using Machine Learning in Ranking

- T. Joachims
 - Support Vector Machines
- Classification
 - "A is good", "B is bad"
 - Often not appropriate
- Ranking
 - "A is better than B"
 - Information Retrieval
- Example
 - Improve Search Engine using Machine Learning



A screenshot of a Microsoft Internet Explorer browser window displaying a Google search for "support vector machine". The search results page shows the Google logo, the search query, and the number of results: "10 of about 282,000". A red circle highlights the number "282,000", and a red arrow points to it from the text "282,000 hits" written below the circle. The search results list several links related to support vector machines, including "Kernel Machines", "svm.first.gmd.de", "Support Vector Machine" from the University of Dortmund, "Support Vector Machines - The Book", and "Support Vector Machine - The Software".

Experiment

- **Training**- 260 queries from 20 users
- **Learning**- ranking function, r
- **Comparison**- Google, MSN, TopRank
- **Results**- r shows improvement with a 95% significance using two-tailed binomial sign test
 - Google: in 13/88 queries users clicked on more Google links, and in 29, more from r



Modeling Bursty Streams, J. Kleinberg

- What are Bursts of Activity? – sharp rises in frequency of certain features of the stream

Email
Messages

Conference
Titles

Current
Affairs

Web Usage
Data

- Bursts are modeled as state transitions in a weighted automaton model

ACM SIGKDD '02



US Presidential State of the Union Address (1790 to 2002)

Word	Interval of burst
gentlemen	1790 - 1800
whilst	1857 - 1860
emancipation	1862 - 1864
coinage	1877 - 1886
depression	1930 - 1937
war	1942 - 1945
peacetime	1945 - 1947
inflation	1971 - 1980
women	1981 - 1984
college	1995 -
communities	1995 -
america	1996 -



"studies digital information in its human and social context."

Dean Robert Constable



- *Information systems*
- *Social systems*
- *Human centered systems*



[Research Programs](#)

[Undergraduate Education](#)

[Courses](#)

[Concentration/Minor/Major](#)

[Graduate Education](#)

[Faculty and Researchers](#)

[Information Science Building](#)

[IS Seminar Series](#)

INFORMATION SCIENCE MAJOR JUST APPROVED



Reference Cites

- FDL www.nuprl.org/FDLproject
- NSDL www.nsd1.org/
- Rank www.cs.cornell.edu/People/tj/
- Bursts www.cs.cornell.edu/home/kleinber
- Info Science www.cis.cornell.edu/

Thank you!

lolorigo@cs.cornell.edu

