Upgrading Enterprise Search^{*}

R. McDunn

Stanford Linear Accelerator Center, Stanford University, Stanford, California 94309

Abstract

This presentation will describe the process we went through this past year to upgrade our enterprise search tool from a very old version of Inktomi to the latest version of Verity Ultraseek. We started with requirements gathering and then compared requirements against several available products to determine which product to choose. After purchasing the product, we worked through several defined phases of implementation and customization, with initial rollout late January 2004. Finally, we will show you where we are today and describe future search plans.

Invited talk presented at Interlab 2004

Oak Ridge, TN

October 27-29, 2004

^{*} Work supported by Department of Energy contract DE-AC02-76SF00515.

InterLab 2004 :: October 28, 2004

Upgrading Enterprise Search

Ruth McDunn

Web Information Manager

mcdunn@slac.stanford.edu

Stanford Linear Accelerator Center

Search at SLAC

- Started with Infoseek product in 1998
- License shared with Stanford University
- Limited by the decisions that Stanford made:
 - Document count (500K limit)
 - Security issues (no facility to index secure content)
- In 2003, Stanford decided to change:
 - They moved to Google.
 - Stanford departments could go or stay.
 - What were we to do?

Infoseek product name changed to Inktomi, then Ultraseek.

Our Options

1. Follow Stanford to Google.
 2. Purchase our own license for a search product.
 3. Stay with the existing Inktomi product (ver 4x).
- Our decision needed to deal with:
 - Growing dissatisfaction with the quality and timeliness of the existing search implementation
 - Multiple Web servers on different OS and file systems
 - We weren't indexing intranet/secure content
 - SLAC is heavily indexed by outside agents and some users had better luck with outside search tools

Search Goal

- We needed an enterprise search facility:
 - “With the speed and accuracy of an electron beam.”
 - “Yielding results as relevant as a Nobel prize winning experiment.”
 - “That covered the breadth of our document universe.”
 - “That would not try to index black holes.”
 - “That did not require an Einstein to use (or administer).”
- This lead us to a process to develop requirements and analyze vendors.

Requirements Development

- Brainstorming
 - Requirement description
 - Problems to solve
- Determined user of requirement/feature
 - Searcher, Web Author, Search Administrator
- Accessed value of requirement/feature
 - Must Have, Desirable, Future Need, Don’t Need

Vendor Review

- We looked at five vendors:
 - Google
 - Ultraseek
 - ht://dig
 - Swish
 - Your Amigo
- Eliminated open source products (ht://dig and Swish)
 - We don’t have available in-house programming skills
- Eliminated Your Amigo
 - Indexing method used processes that may not be approved by our computer security group

Vendor Analysis

- For all **Must-Have** and **Desirable** requirements, we reviewed Google and Ultraseek to determine if the product addressed a requirement:

- Out of the box
- Can be done
 - What would it take to address the requirement in terms of time, hardware, money, programming
- Not a feature

Typical Analysis

Analysis Details

Requirement Name:	Sorted search results (64)
Group Ranking:	Must Have
Current Problem:	Currently users cannot sort search results very effectively.
Requirement Description:	Capability of sorting search results by a variety of methods. Examples are date, group, category, type, relevancy, and so on. If search categories are defined, then capability of sorting results by category. Ability to apply multiple criteria to a search (i.e., by category then date).
Submitted by:	mcdunn on 6/10/2003

Vendor	Availability	Analysis	
UltraSeek	Out of the box	Supports sorting of search results by date or by relevancy.	Bebo on 6/4/2003
Google	Out of the box	The Google Search Appliance supports sorting of search results by relevancy or date.	Google Rep on 6/4/2003

Fields include: Requirement Name, Group Ranking, Current Problem, Requirement Description, Submitted by.
Vendor analysis includes vendor name, availability, analysis, and who performed the analysis.

Purchased Ultraseek

- It was a good, but not perfect, match to our requirements
- It leveraged our existing experience
- Avoided a “hybrid” search environment – public vs. private
- Minimized impact to existing search pages
 - No change in page URL or collection names
- New features: Quick Links, spellchecking, highlight word
- New administrative features: Reports, error messages, log files
- Supported future searching models (CCE)

Phase 1

- Why a phased rollout?
 - <<<1 FTE to manage the entire search function
- Phase 1
 - Install and replicate existing collections (January 2004)
 - Add new and special collections (pubs, SLAC Today, intranet)
 - Implement three search pages - on site, off site, public
 - Customize the search results pages
 - Python coding - [EditPlus](#)
 - Silent rollout - simply substituted one IP for another
 - Start to publicize changes and integrate intranet collections
 - February 20, 2004 The Interaction Point article “[SLAC Web Search Improved!](#)”

Server Configuration

- Hardware
 - Dell PowerEdge 1650 with Windows 2000 (previous server was Solaris)
 - 60 GB data disk
- Software
 - Ultraseek 5.1
 - CCE
 - SSL Security Module

Three Search Pages

- If internal collection names are revealed, a hacker could see the search results (URL and summary) from our internal collections

- Use of passage-based summaries may increase information exposure
- A temporary solution:
 - Public collections searched through public search page and search forms
 - Public and intranet collections searched through two internal pages and search forms on internal pages
 - One page restricted by IP address (onsite users)
 - One page restricted by SSL/Windows Authentication (offsite users)

Customized Search Results Page

[Skip to content](#) | [Back to Search SLAC Public Web](#)

Search the SLAC Public Web

Start new search [Search these results](#)

Collections to Search:

☒ SLAC Public ☒ SLAC Today ☒ SSRL Public ☒ SLAC Publications ☒ BaBar Public

More collections: ([onsite computer](#) or [offsite computer](#)).

theory of quarks

search

[Help](#) [Advanced Search](#)

Tip: You can restrict your search to the title of a document.

Example: title:"Holiday Schedule"

162913 results found, top 500 sorted by relevance

[score using date](#) [hide summaries](#) [group by location](#)

1-10

Theory: Color Charge

... a whimsically named attribute of **quarks** and gluons that cannot be ... the predictions of the gluon **theory of quark** interactions. This **theory** is called QCD, short for ...

<http://www2.slac.stanford.edu/wc/theory/colorchrg.html> - 11.5KB - SLAC Public

70%

05 May 03

[Find Similar](#)
[Highlight](#)

SLAC-PUB-8267 -- Quarks, Gluons and Frustrated Antiferromagnets

... the frustrated antiferromagnet Hamiltonian, the **theory** becomes a **theory of quarks** interacting with color gauge-fields ...

<http://www.slac.stanford.edu/pubs/slacpubs/8000/slac-pub-8267.html> - 3.4KB - SLAC Publications

66%

04 Apr 01

[Find Similar](#)
[Highlight](#)

ssi77 001

... We Have a Successful Gauge **Theory of Quarks** and Leptons? 5.2 Experimental ... involving even larger numbers of **quarks** and leptons. The proliferation of **quarks** and leptons has been sudden ...

http://www.slac.stanford.edu/pubs/slacpubs/ssi77/ssi77_001.pdf - 6510.0KB - SLAC Publications

58%

03 Jan 00

[Find Similar](#)
[Highlight](#)

Customized the style sheets, added common header and footer, added collection selection, added links to full search pages, rewrote tips to include SLAC content, removed the Verity logo, highlighted the collections in the results.

Phase 2

- Upgraded to version 5.2 (March 2004)
- Researched security issue (content revealed in results summaries), possible solutions, and estimated costs
- Analyze reports and implemented Quick Links
- Push SLAC searchers toward the internal, full search pages
- Start thinking about taxonomies and CCE usage
- Plan and initiate user and author training
 - [What's New in Searching the SLAC Web](#)
 - SLAC Web Best Practices: Make web pages search-friendly

Report: Top Queries

- Information to build a taxonomy

Top Queries - 4/1/2004 to 9/30/2004

Rank	Query	Times	Avg. Clicks	Null Results	Avg. Page Views
1	cafeteria	582	0.29	0	1.00
2	citrix	443	0.95	0	1.00
3	travel	381	0.65	0	1.00
4	map	361	0.77	0	1.03
5	cafe	322	0.34	0	1.00
6	library	280	0.99	0	1.00
7	flea	253	1.01	0	1.00
8	xweb	195	0.67	0	1.05
9	family day	174	0.97	0	1.11
10	mcdunn	151	0.12	0	1.02

11	sluo	150	1.00	0	1.00
12	maps	145	0.78	0	1.00
13	security	132	0.23	0	1.03
14	vpn	131	0.88	0	1.03
15	ssrl	131	0.86	0	1.00
16	babar	130	1.01	0	1.08
17	housing	130	1.12	0	1.01

Report: Top Queries with No Results

- Add/edit
 - A Quick Link, especially for intranet sites
 - Content on the web site to improve findability and reindex

Top Queries with No Results - 4/1/2004 to 9/30/2004

Rank	Query	Times
1	3DIsst	22
8	Cyterski	7
9	+url:www.slac.stanford.edu/econf Ouyed	7
10	+url:www-group.slac.stanford.edu/hr globie	7
11	+url:www-group.slac.stanford.edu/hr aina cohen	6
12	url:www2.slac.stanford.edu/vvc Mailstop74	6
13	pseudomyxoma peritonei	6
14	slac-puc-2631	6
15	www.fidelity.com	6
16	url:www.slac.stanford.edu/esh url:www-group.slac.stanford.edu/esh JHAM	6

Report: Top Requested Documents

Top Requested Documents - 4/1/2004 to 9/30/2004

Rank	URL	Times
1	http://www.slac.stanford.edu/slac/fleamarket/ <i>Top Queries:</i> fleamarket (28); slac fleamarket (2); mazda (2); flea (252); Flea Market (4); slac flea market (7); market (6); jiri@slac moving home-staff (3); flee market (12); flea market (94); Make a Quick Link	428
2	http://www.slac.stanford.edu/slac/map/ <i>Top Queries:</i> map (184); slac map (38); slac maps (6); area maps (2); site map (9); Maps (11); slac site map (5); building numbers (2); campus map (6); maps (94); site maps (2); Make a Quick Link	392
3	http://www2.slac.stanford.edu/computing/windows/services/citrix/ <i>Top Queries:</i> citrix + mac (1); citrix web (1); farm (2); citrix at slac (3); citrix from unix (1); linux citrix (1); citrix unix (1); citrix (255); metaframe (1); citrix logon (1); citrix xp (1); terminal server (1); citrix with OSX (1); citrix linux (2); ica server (1); Citrix MetaFrame XP Farm (1); citrix csa xp (1); server (1); citrix internet (1); Citrix access (1); citrix v7 (1); citrix farm (1); slac citrix (8); citrix farm web page (1); Citrix (15); citrix install (1); wts citrix (1); Make a Quick Link	306
4	http://www-group.slac.stanford.edu/travel/slac_travel.html <i>Top Queries:</i> currency converter (1); sensitive countries (1); travel (235); travel reimbursement (1); Travel Department (1); travel department (1); foreign sensitive subjects (1); travel reimbursement form (1); slac travel request (1); slac travel form (1); travel office (3); local travel (1); slac travel (5); Travel (7); slac travel health insurance (1); Make a Quick Link	261
5	http://www.slac.stanford.edu/comp/net/email/index.html <i>Top Queries:</i> mailing list (3); exchange (4); web mail (48); email (66); Web mail (3); mailing lists (3); mail (45); email forwarding (4); Email (3); .forward (3); Make a Quick Link	236

Report: Quick Links

Keywords	Result		
cafeteria, cafeteria menu, menu, slac menu, cafe, slac cafe, slac cafeteria, café, cafe men, cafeteria hours, slac cafeteria menu, lunch, breakfast, dinner, snacks, coffee, place to eat	SLAC Cafe Menu	edit	delete
maps, map, building number, campus map, room, map of slac, building map, slac map, slac area, site map, area map, meeting rooms, stanford area, san francisco area, auditorium, panofsky auditorium, slac local, slac site map, sitemap	Site and Area Maps	edit	delete
travel form, travel office, travel, slac travel, travel department, travel reimbursement, travel request, lodging, currency converter, per diem	SLAC Travel Reimbursement	edit	delete
dhcp form, visitor tap, dhcp, dhcp registration, dhcp czar, mac address, dhcp czar list	DHCP @ SLAC	edit	delete
citrix, citrix terminal, windows terminal, citrix slac, metaframe xp, citrix for linux, citrix account, citrix at home, metaframe, citrix client, new citrix farm, citrix metaframe, how do i map, citrix help, citrix farm, windows terminal server, citrix ica wts, wts, ica, setting up citrix, citrix + mac, citrix and osx, linux citrix	Citrix MetaFrame XP Farm @ SLAC	edit	delete

Report: Quick Link Detail

- Modified URL allows tracking of Quick Link usage
- Custom Summary allows us to include intranet sites

Parameters
Users
Doc types
XML Mappings
Quick Links

Keywords:
security, ssec, site security, safeguards, main gate, sector 30 gate, site entr

URL:
http://www-search.slac.stanford.edu:8765/cs.html?url=https://www-internal.sl

Title:
☐ Automatic Title
☒ Custom Title:

Summary: ☐ Automatic Summary
☒ Custom Summary (HTML is allowed):

SLAC Safeguards & Security
Administrative Offices provide multiple services to SLAC Employees, Contractors, Users and Visitors. Security Assistance: Main Gate - 24 hrs./day, 365 days/year. Ext. 2551. *SLAC Windows authentication required to view this page.*

Quick Links

Search the SLAC Public Web

Start new search [Search these results](#)

Collections to Search:

- ☒ SLAC Public ☒ SLAC Today ☒ SSRL Public ☒ SLAC Publications
☒ BaBar Public

More collections: ([onsite computer](#) or [offsite computer](#)).

cafe

search

[Help](#) [Advanced Search](#)

Tip: Use lowercase, otherwise, an exact match is used.

Example: BaBar, TIP, SPIRES

Quick Links:

[SLAC Cafe Menu](#)

SLAC Cafe serves breakfast from 6:30 - 9:00 am and lunch from 11:30 - 1:30 pm and is located in Building 42, next to Panofsky Auditorium. Use the links on this page to view the daily menu choices.

97 results found, sorted by relevance

[score using date](#) [hide summaries](#)
[group by location](#)

1-10

[SLAC Cafe - Home](#)

... SLAC Cafe - Home ... SLAC Cafe ...

<http://www2.slac.stanford.edu/cafe/> - 2.4KB - **SLAC Public**

78%

05 Nov 04

[Find Similar](#)
[Highlight](#)

Phase 3 (now)

- Upgraded to Version 5.3 (September 2004)
- Continue to improve in collection definition
- Evaluate effectiveness of Quick Links
- Evaluate new features
 - Hit-level authentication for results display issue
 - Page Expert for templated sites (TIP, ES&H)
- Start taxonomy development based on query reports and explore CCE implementation
- “Optimize Your Website for Search”
 - Web author training based on the best practices

Where are we now?

- Analyzed the search web server logs
- Through September this year, 200K page views or 5,500 page views per week on average
- Would like to see usage increase with time

What search forms are being used?

Top Referring Sites % Page Views

Detailed home page	30.1	59K
Welcome page	17.8	35K
BaBar Public Search	10.8	21K
BaBar Private Search	4.6	9K
SLAC Today	3.1	6K
Search - Public	2.1	4K
Computing	1.8	4K
Search - Internal	1.3	3K

- 3% of the Detailed home page hits are Beck and myself
- Some of the users of the welcome page box should be retrained to the internal search box

Quick Link Use Tracking

- Change Quick Link URL
 - <http://www-search.slac.stanford.edu:8765/cs.html?url=http://www2.slac.stanford.edu/cafe/&qt=clickthrough&col=quicklink&n=0>
- Access Log File
 - 216.175.120.217 - - [24/Oct/2004:05:56:15 -0700] "GET /cs.html?url=<http://www2.slac.stanford.edu/cafe/>&qt=clickthrough&col=quicklink&n=0 HTTP/1.1" 302 186 "" "Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; Win 9x 4.90)"
- Clickthrough Log
 - <http://www2.slac.stanford.edu/cafe/> 2004/10/24 05:56:15 click quicklink clickthrough 0

Quick Links Usage (10/1 - 10/24)

Quick Link Description		Click throughs
Job Hazard Analysis & Mitigation		119
Cafeteria Menu		96
Site Maps		63
Meeting Room Map		38
Email @ SLAC		31
Site Engineering & Maintenance		30
Citrix (remote computer access)		23
Business Information Services		21
Travel Reimbursement		17
Property Management		17

Quick Links Usage (2)

Quick Link Description		Click throughs
XWeb (Software Installation)		15
Holiday Schedule		12
Meeting Room Contacts		11
SLAC Race		11
Safeguards & Security		11
Remote Computer Access		8
The Interaction Point		8
Guest House		7
Education & Outreach		3
SLAC Tours		3

Questions?

- S5
 - Simple Standards based Slide Show System courtesy of Eric Meyer and downloadable at <http://www.meyerweb.com/eric/tools/s5/>.