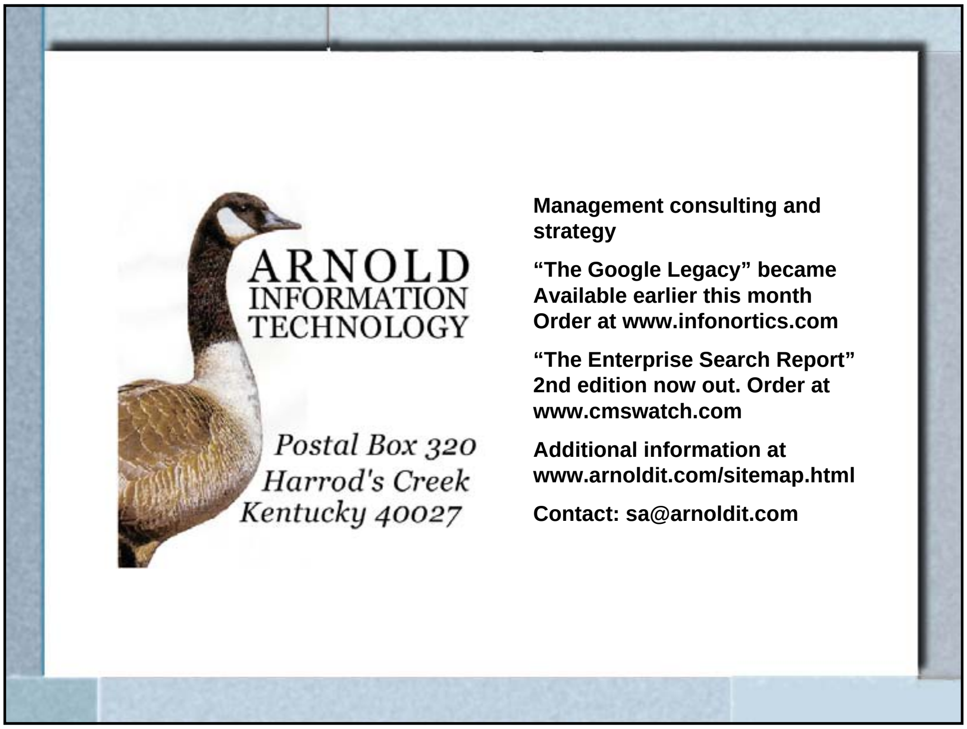
Enterprise Search: What You Must Know about Information Retrieval and the "Google Effect"

© Stephen E. Arnold, Postal Box 320, Harrod's Creek, KY 40027
Email: sa@arnoldit.com – Voice: 502 228 1966



ARNOLD
INFORMATION
TECHNOLOGY

Postal Box 320
Harrod's Creek
Kentucky 40027

Management consulting and strategy

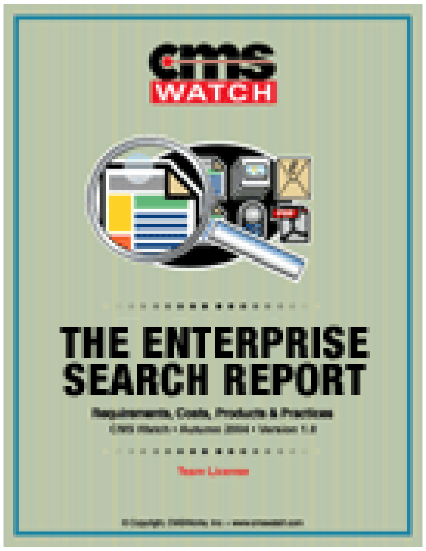
"The Google Legacy" became Available earlier this month Order at www.infonortics.com

"The Enterprise Search Report" 2nd edition now out. Order at www.cmswatch.com

Additional information at www.arnoldit.com/sitemap.html

Contact: sa@arnoldit.com

Limited Access Document



cms WATCH

THE ENTERPRISE SEARCH REPORT

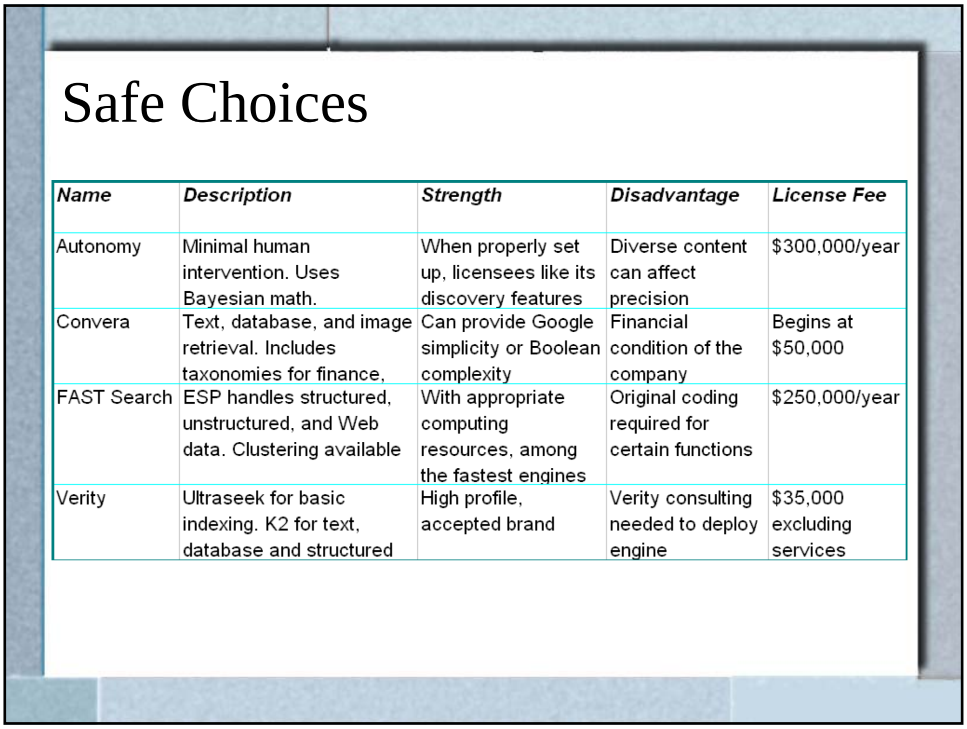
Requirements, Costs, Products & Practices

CMS Watch • Autumn 2004 • Version 1.0

Team License

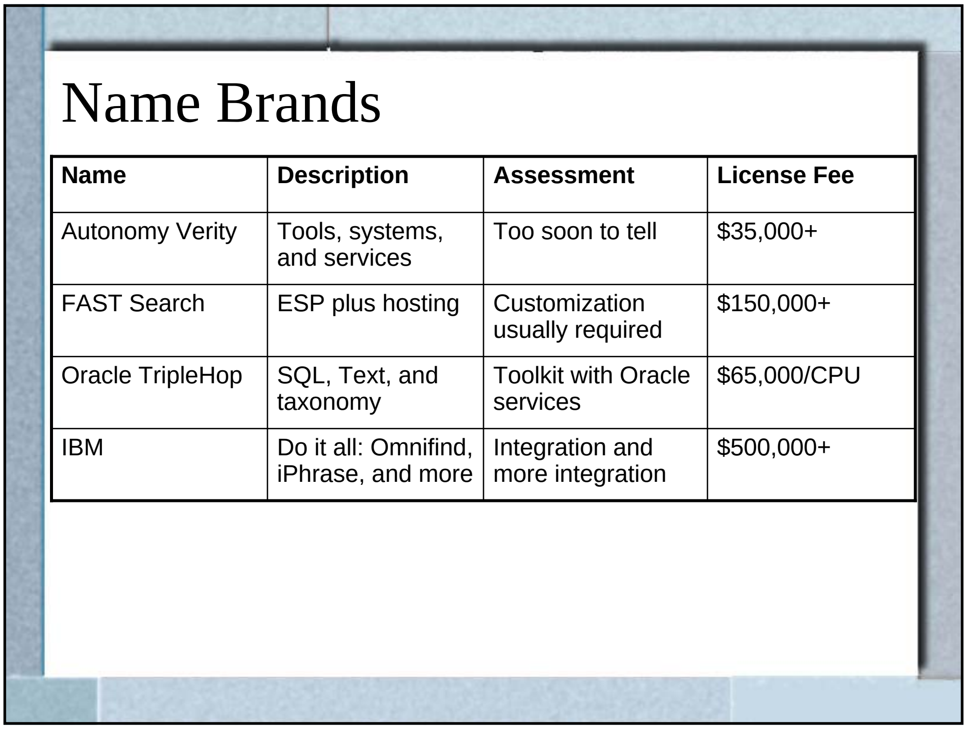
© Copyright CMSWatch, Inc. – www.cmswatch.com

4
Search System Vendors



Safe Choices

Name	*Description*	*Strength*	*Disadvantage*	*License Fee*
Autonomy	Minimal human intervention. Uses Bayesian math.	When properly set up, licensees like its discovery features	Diverse content can affect precision	$300,000/year
Convera	Text, database, and image retrieval. Includes taxonomies for finance,	Can provide Google simplicity or Boolean complexity	Financial condition of the company	Begins at $50,000
FAST Search	ESP handles structured, unstructured, and Web data. Clustering available	With appropriate computing resources, among the fastest engines	Original coding required for certain functions	$250,000/year
Verity	Ultraseek for basic indexing. K2 for text, database and structured	High profile, accepted brand	Verity consulting needed to deploy engine	$35,000 excluding services



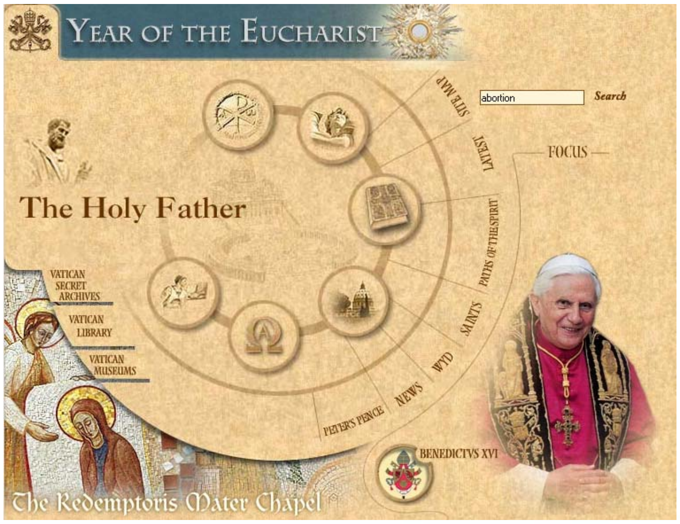
Name Brands

Name	**Description**	**Assessment**	**License Fee**
Autonomy Verity	Tools, systems, and services	Too soon to tell	$35,000+
FAST Search	ESP plus hosting	Customization usually required	$150,000+
Oracle TripleHop	SQL, Text, and taxonomy	Toolkit with Oracle services	$65,000/CPU
IBM	Do it all: Omnifind, iPhrase, and more	Integration and more integration	$500,000+

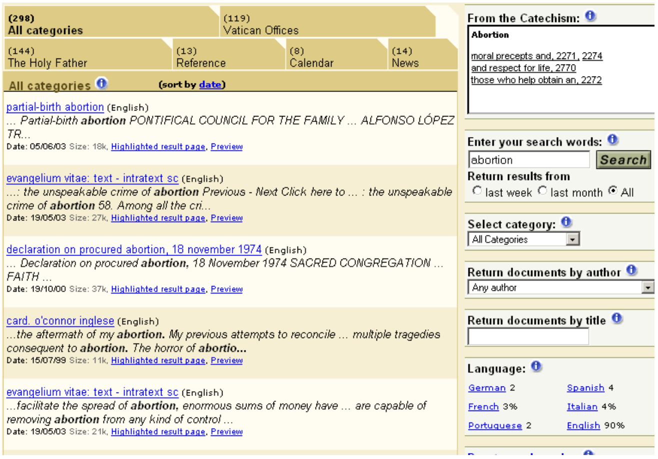
Hot Engines

Name	Description	Assessment	License Fee
Endeca	Search plus taxonomy	Good. Work flow angle	$200,000+
Mondosoft	Search plus taxonomy plus analytics	Robust plus APIs	$50,000+
Vivisimo	Federation and deduplication	Integration solution	$65,000+

Mondosoft “Hybrid” Interface



Hybrid Search: Facets, Hard Coding, Synonym Expansion



(298) **All categories** | (119) Vatican Offices

(144) The Holy Father | (13) Reference | (8) Calendar | (14) News

All categories (sort by date)

partial-birth abortion (English)
... Partial-birth ***abortion*** *PONTIFICAL COUNCIL FOR THE FAMILY ... ALFONSO LÓPEZ TR...*
Date: 05/06/03 Size: 18k, Highlighted result page, Preview

evangelium vitae: text - intratext sc (English)
...: the unspeakable crime of ***abortion*** *Previous - Next Click here to ... : the unspeakable crime of* ***abortion*** *58. Among all the cri...*
Date: 19/05/03 Size: 27k, Highlighted result page, Preview

declaration on procured abortion, 18 november 1974 (English)
... Declaration on procured ***abortion,*** *18 November 1974 SACRED CONGREGATION ... FAITH ...*
Date: 19/10/00 Size: 37k, Highlighted result page, Preview

card. o'connor inglese (English)
...the aftermath of my ***abortion.*** *My previous attempts to reconcile ... multiple tragedies consequent to* ***abortion.*** *The horror of* ***abortio...***
Date: 15/07/99 Size: 11k, Highlighted result page, Preview

evangelium vitae: text - intratext sc (English)
...facilitate the spread of ***abortion,*** *enormous sums of money have ... are capable of removing* ***abortion*** *from any kind of control ...*
Date: 19/05/03 Size: 21k, Highlighted result page, Preview

From the Catechism:

Abortion

moral precepts and, 2271, 2274
and respect for life, 2770
those who help obtain an, 2272

Enter your search words: abortion **Search**

Return results from ○ last week ○ last month ● All

Select category: All Categories

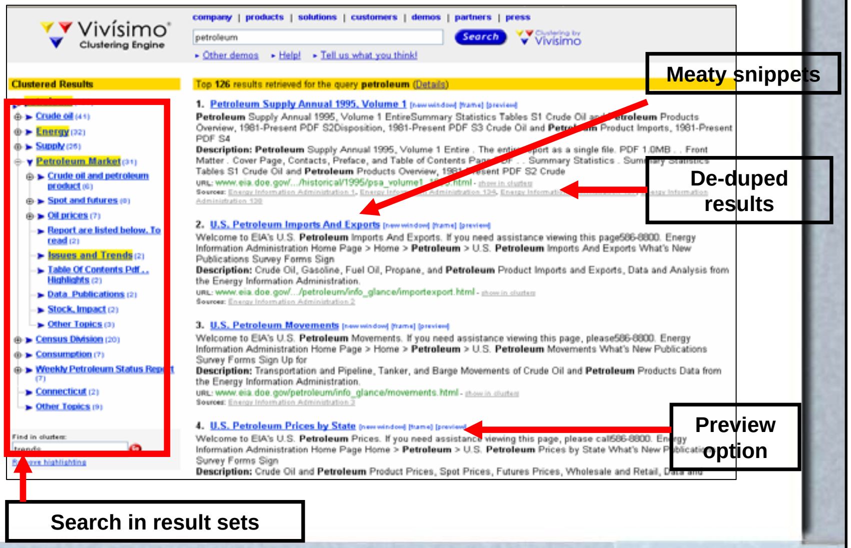
Return documents by author Any author

Return documents by title

Language:

German 2	Spanish 4
French 3%	Italian 4%
Portuguese 2	English 90%

Vivisimo



petroleum

Clustered Results

- Crude oil (41)
- Energy (32)
- Supply (26)
- Petroleum Market (31)
  - Crude oil and petroleum product (6)
  - Spot and futures (8)
  - Oil prices (7)
  - Report are listed below. To read (2)
  - Issues and Trends (2)
  - Table Of Contents Pdf .. Highlights (2)
  - Data, Publications (2)
  - Stock, Impact (2)
  - Other Topics (3)
- Census Division (20)
- Consumption (7)
- Weekly Petroleum Status Report (7)
- Connecticut (2)
- Other Topics (9)

Find in clusters: trends

Remove highlighting

Top **126** results retrieved for the query **petroleum** (Details)

1. Petroleum Supply Annual 1995, Volume 1 [new window] [frame] [preview]
Petroleum Supply Annual 1995, Volume 1 EntireSummary Statistics Tables S1 Crude Oil and Petroleum Products Overview, 1981-Present PDF S2Disposition, 1981-Present PDF S3 Crude Oil and Petroleum Product Imports, 1981-Present PDF S4
Description: Petroleum Supply Annual 1995, Volume 1 Entire . The entire report as a single file. PDF 1.0MB . . Front Matter . Cover Page, Contacts, Preface, and Table of Contents Page PDF . . Summary Statistics . Summary Statistics Tables S1 Crude Oil and **Petroleum** Products Overview, 1981-Present PDF S2 Crude
URL: www.eia.doe.gov/.../historical/1995/psa_volume1_1995.html - show in clusters
Sources: Energy Information Administration 1, Energy Information Administration 134, Energy Information Administration 127, Energy Information Administration 128

2. U.S. Petroleum Imports And Exports [new window] [frame] [preview]
Welcome to EIA's U.S. **Petroleum** Imports And Exports. If you need assistance viewing this page586-8800. Energy Information Administration Home Page > Home > **Petroleum** > U.S. **Petroleum** Imports And Exports What's New Publications Survey Forms Sign
Description: Crude Oil, Gasoline, Fuel Oil, Propane, and **Petroleum** Product Imports and Exports, Data and Analysis from the Energy Information Administration.
URL: www.eia.doe.gov/.../petroleum/info_glance/importexport.html - show in clusters
Sources: Energy Information Administration 2

3. U.S. Petroleum Movements [new window] [frame] [preview]
Welcome to EIA's U.S. **Petroleum** Movements. If you need assistance viewing this page, please586-8800. Energy Information Administration Home Page > Home > **Petroleum** > U.S. **Petroleum** Movements What's New Publications Survey Forms Sign Up for
Description: Transportation and Pipeline, Tanker, and Barge Movements of Crude Oil and **Petroleum** Products Data from the Energy Information Administration.
URL: www.eia.doe.gov/petroleum/info_glance/movements.html - show in clusters
Sources: Energy Information Administration 3

4. U.S. Petroleum Prices by State [new window] [frame] [preview]
Welcome to EIA's U.S. **Petroleum** Prices. If you need assistance viewing this page, please call586-8800. Energy Information Administration Home Page Home > **Petroleum** > U.S. **Petroleum** Prices by State What's New Publications Survey Forms Sign
Description: Crude Oil and **Petroleum** Product Prices, Spot Prices, Futures Prices, Wholesale and Retail, Data and

Meaty snippets

De-duped results

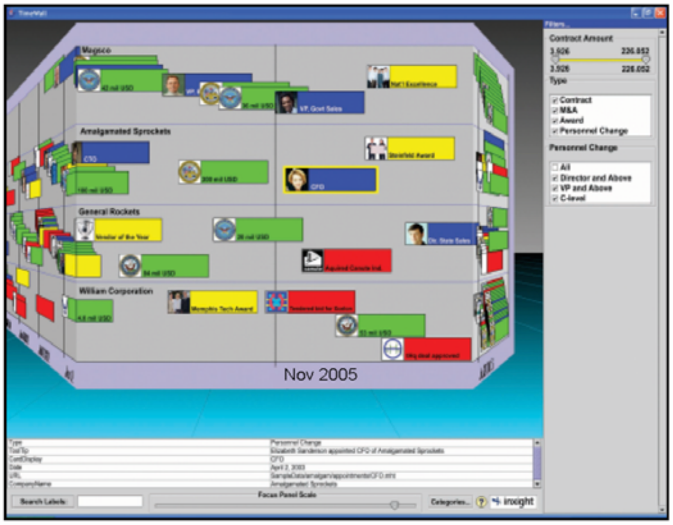
Preview option

Search in result sets

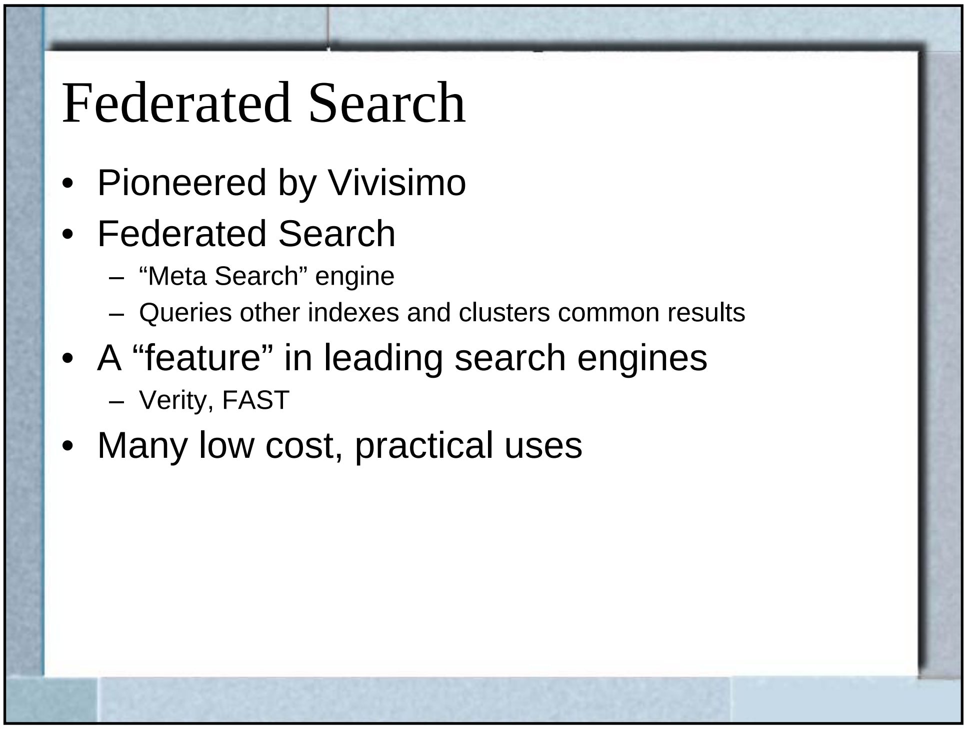
Special Purpose Tools

Name	Description	Assessment	License Fee
Inxight	Search plus text mining	A toolkit	$50,000+
Stratify	Search plus taxonomy	Set up for human input	$75,000+

Inxight

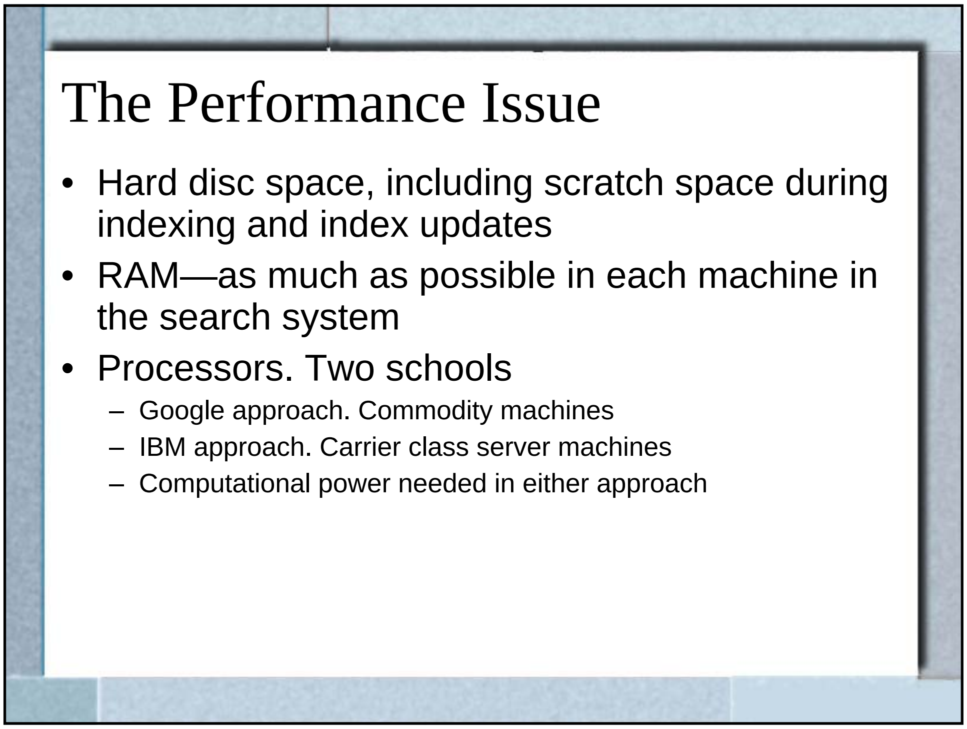


Type	Personnel Change
ToolTip	Elizabeth Sanderson appointed CFO of Amalgamated Sprockets
CardDisplay	CFO
Date	April 2, 2003
URL	SampleData\amalgam\appointments\CFO.mht
CompanyName	Amalgamated Sprockets



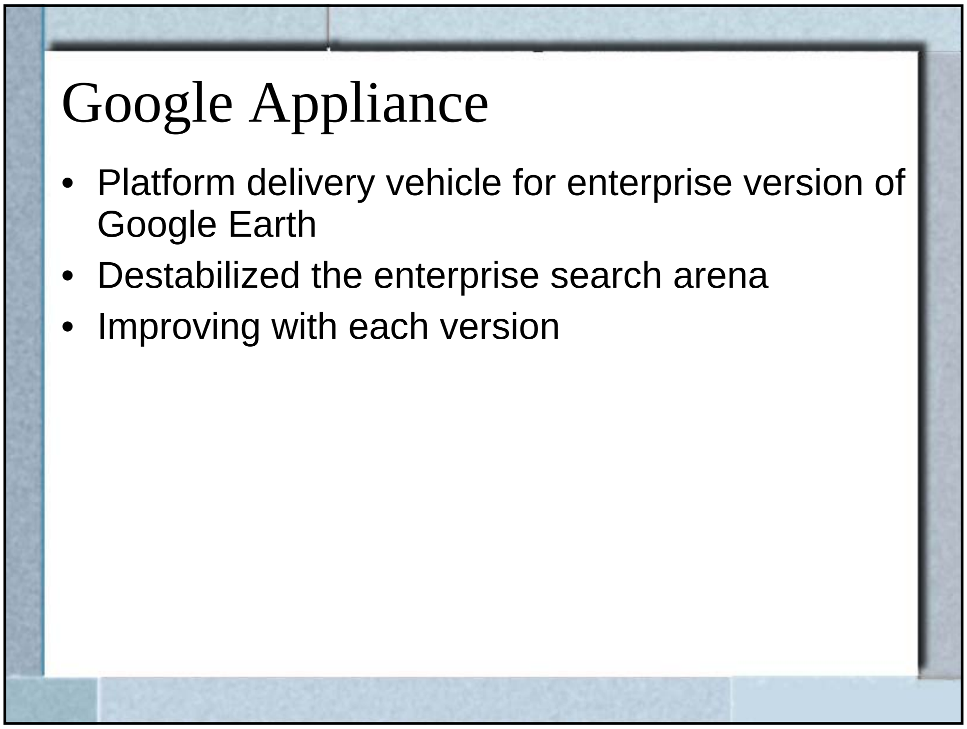
Federated Search

- Pioneered by Vivisimo
- Federated Search
  - “Meta Search” engine
  - Queries other indexes and clusters common results
- A “feature” in leading search engines
  - Verity, FAST
- Many low cost, practical uses



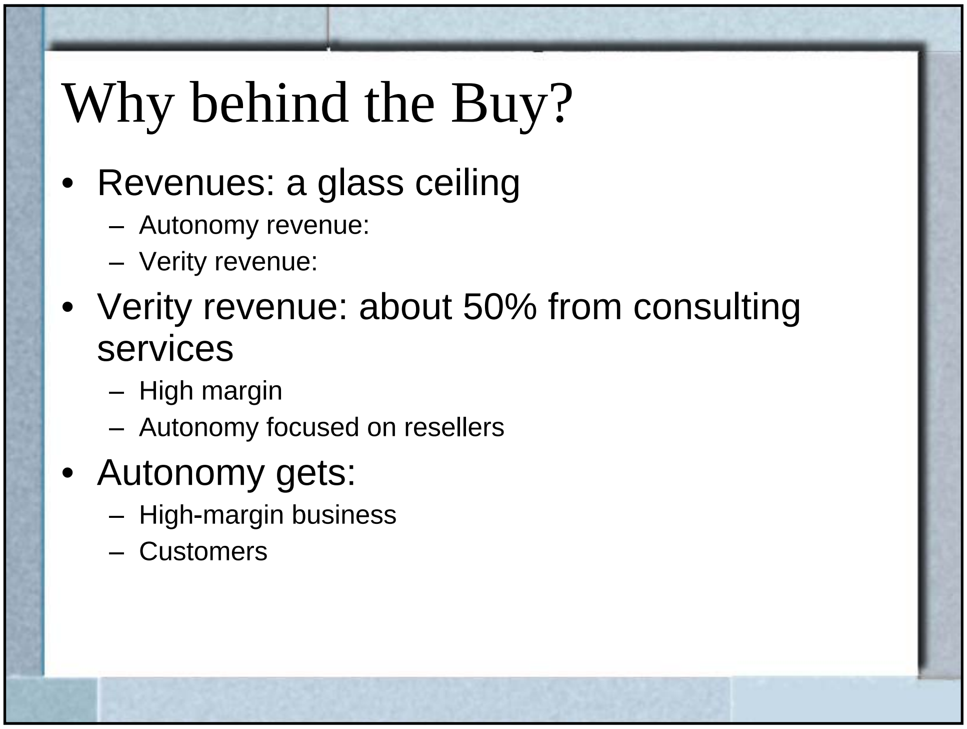
The Performance Issue

- Hard disc space, including scratch space during indexing and index updates
- RAM—as much as possible in each machine in the search system
- Processors. Two schools
  - Google approach. Commodity machines
  - IBM approach. Carrier class server machines
  - Computational power needed in either approach



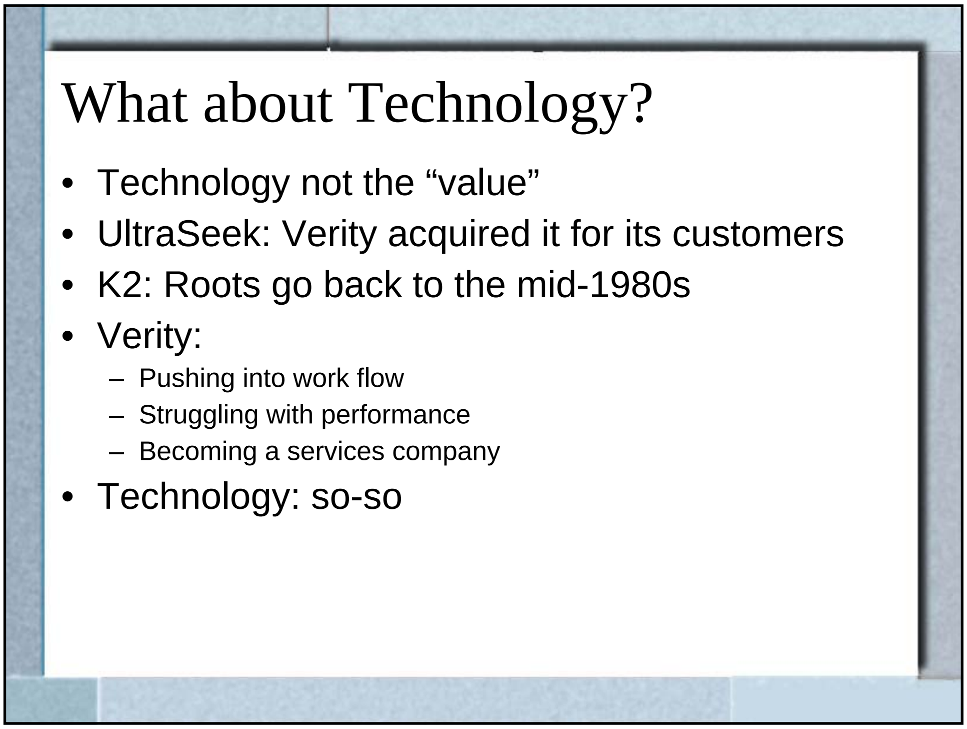
Google Appliance

- Platform delivery vehicle for enterprise version of Google Earth
- Destabilized the enterprise search arena
- Improving with each version



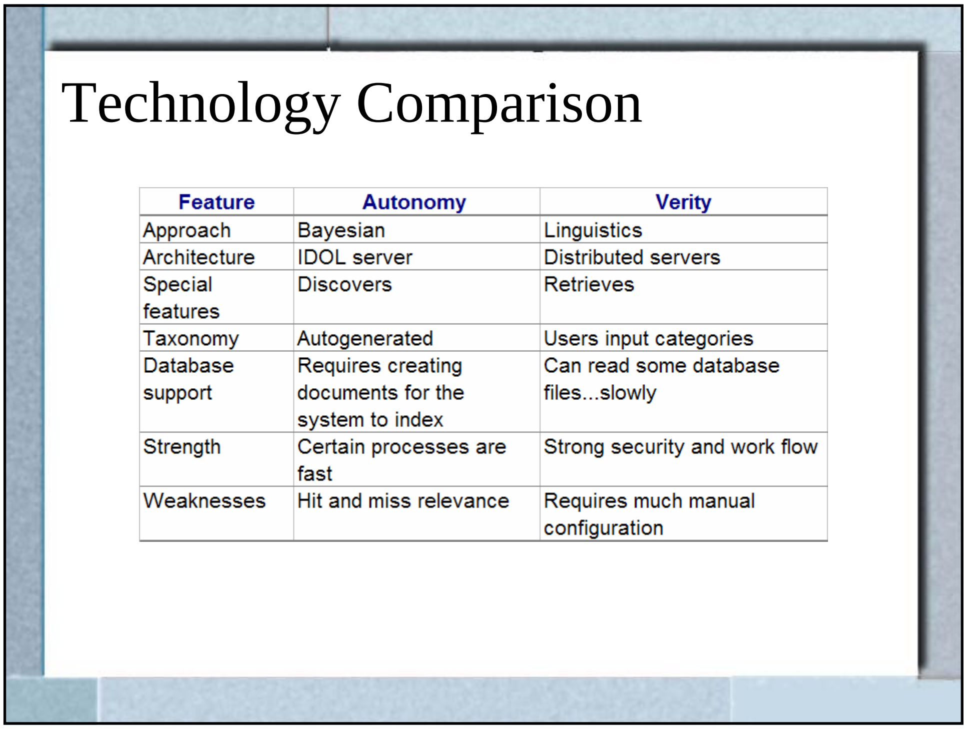
Why behind the Buy?

- Revenues: a glass ceiling
  - Autonomy revenue:
  - Verity revenue:
- Verity revenue: about 50% from consulting services
  - High margin
  - Autonomy focused on resellers
- Autonomy gets:
  - High-margin business
  - Customers



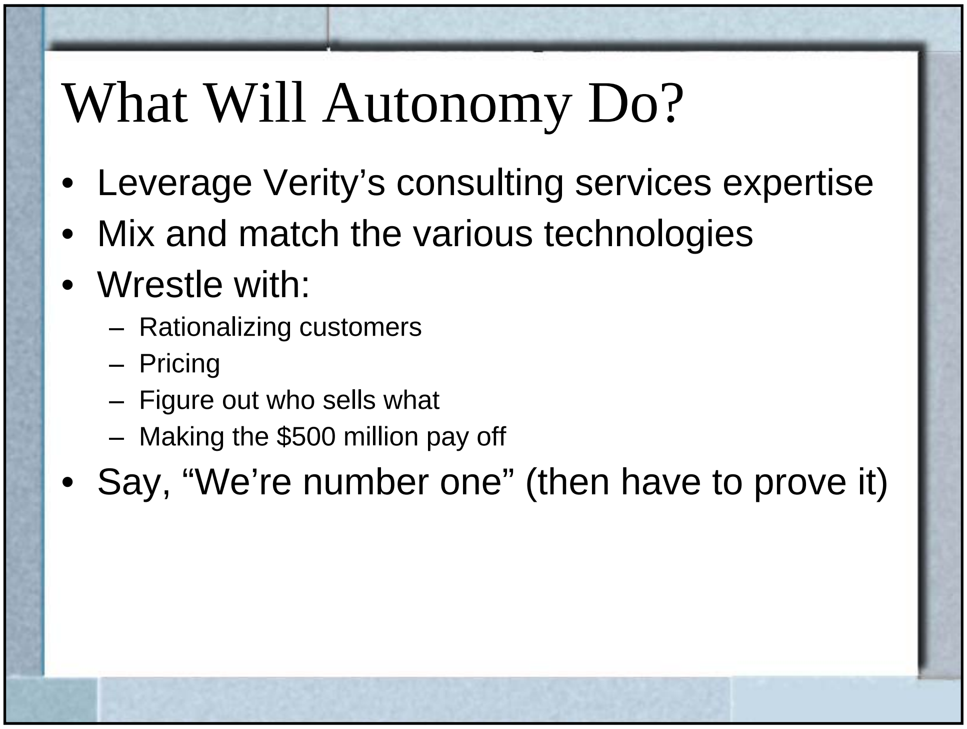
What about Technology?

- Technology not the “value”
- UltraSeek: Verity acquired it for its customers
- K2: Roots go back to the mid-1980s
- Verity:
  - Pushing into work flow
  - Struggling with performance
  - Becoming a services company
- Technology: so-so



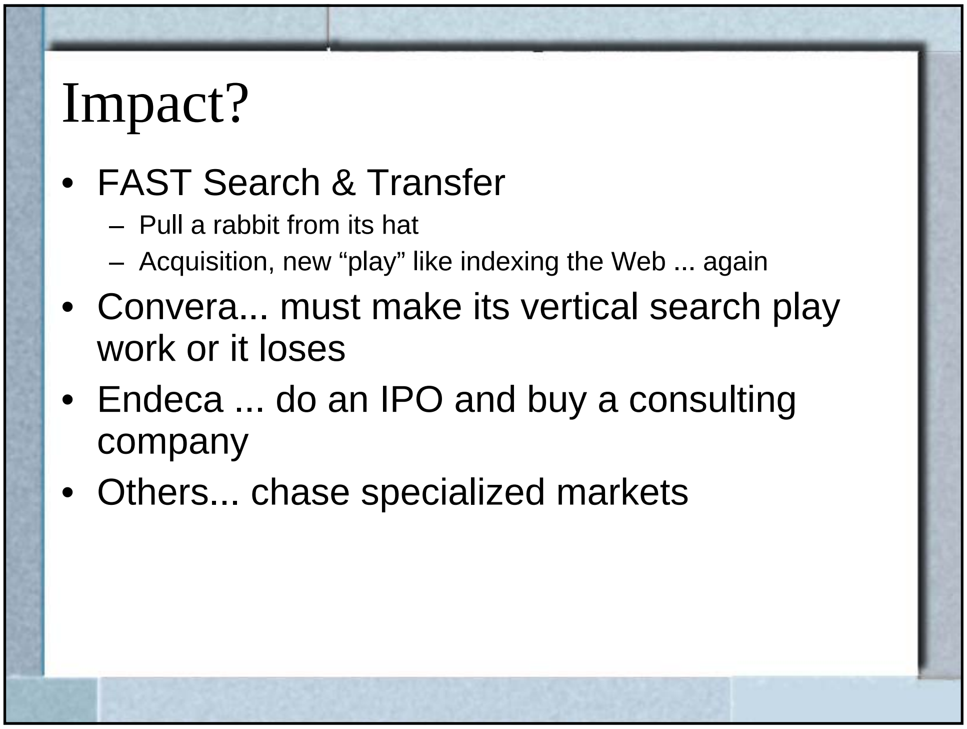
Technology Comparison

Feature	Autonomy	Verity
Approach	Bayesian	Linguistics
Architecture	IDOL server	Distributed servers
Special features	Discovers	Retrieves
Taxonomy	Autogenerated	Users input categories
Database support	Requires creating documents for the system to index	Can read some database files...slowly
Strength	Certain processes are fast	Strong security and work flow
Weaknesses	Hit and miss relevance	Requires much manual configuration



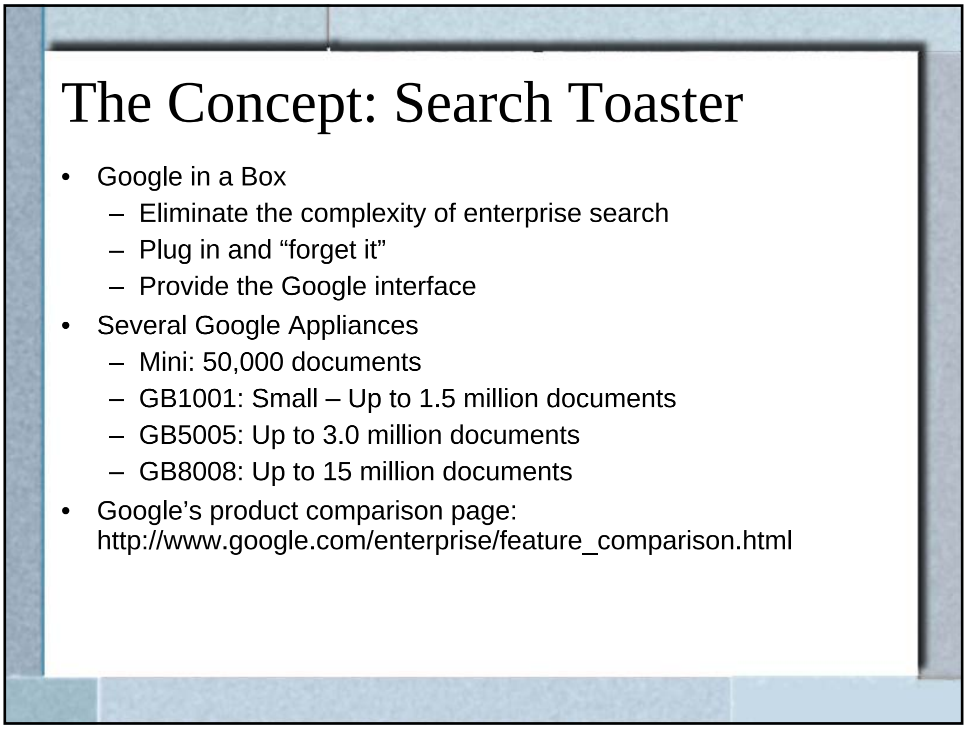
What Will Autonomy Do?

- Leverage Verity’s consulting services expertise
- Mix and match the various technologies
- Wrestle with:
  - Rationalizing customers
  - Pricing
  - Figure out who sells what
  - Making the $500 million pay off
- Say, “We’re number one” (then have to prove it)



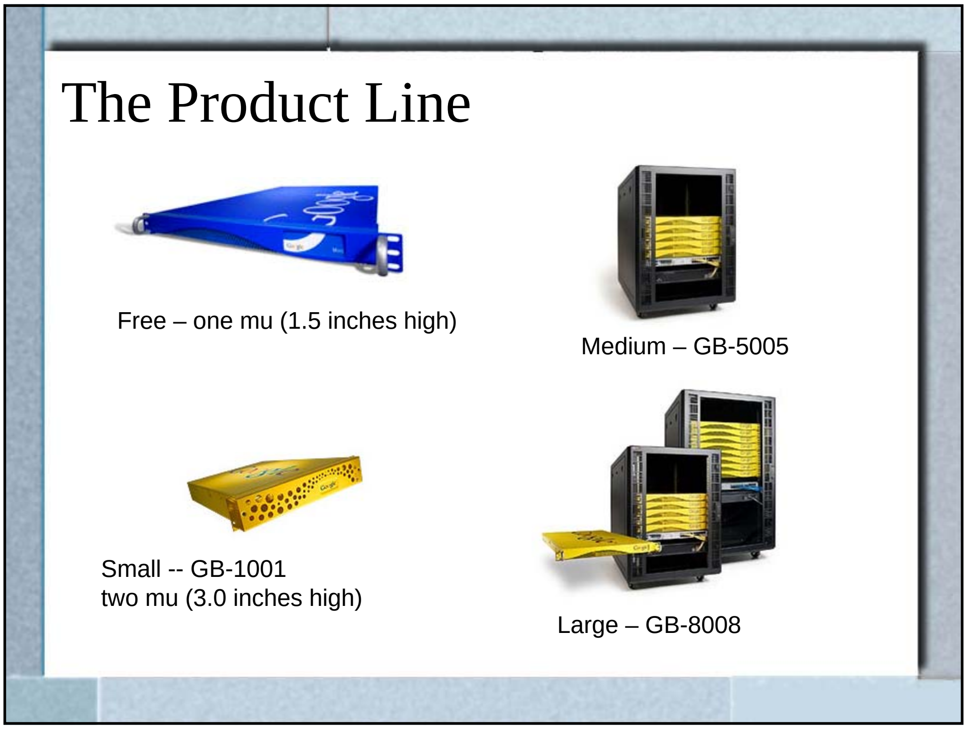
Impact?

- FAST Search & Transfer
  - Pull a rabbit from its hat
  - Acquisition, new “play” like indexing the Web ... again
- Convera... must make its vertical search play work or it loses
- Endeca ... do an IPO and buy a consulting company
- Others... chase specialized markets



The Concept: Search Toaster

- Google in a Box
  - Eliminate the complexity of enterprise search
  - Plug in and “forget it”
  - Provide the Google interface
- Several Google Appliances
  - Mini: 50,000 documents
  - GB1001: Small – Up to 1.5 million documents
  - GB5005: Up to 3.0 million documents
  - GB8008: Up to 15 million documents
- Google’s product comparison page: http://www.google.com/enterprise/feature_comparison.html



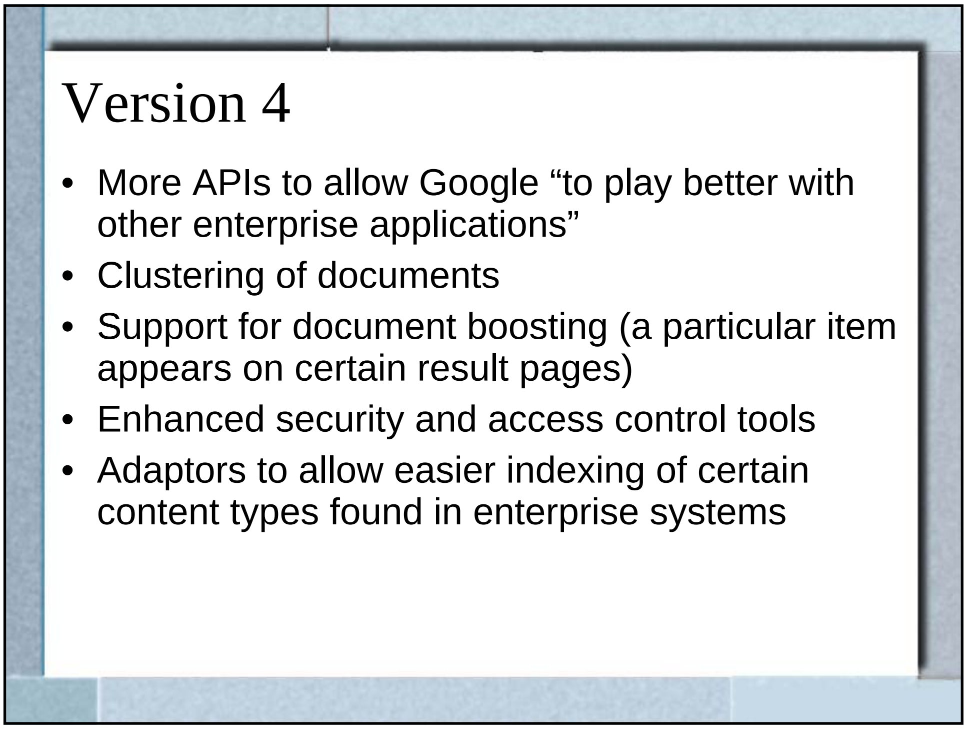
The Product Line

Free – one mu (1.5 inches high)

Medium – GB-5005

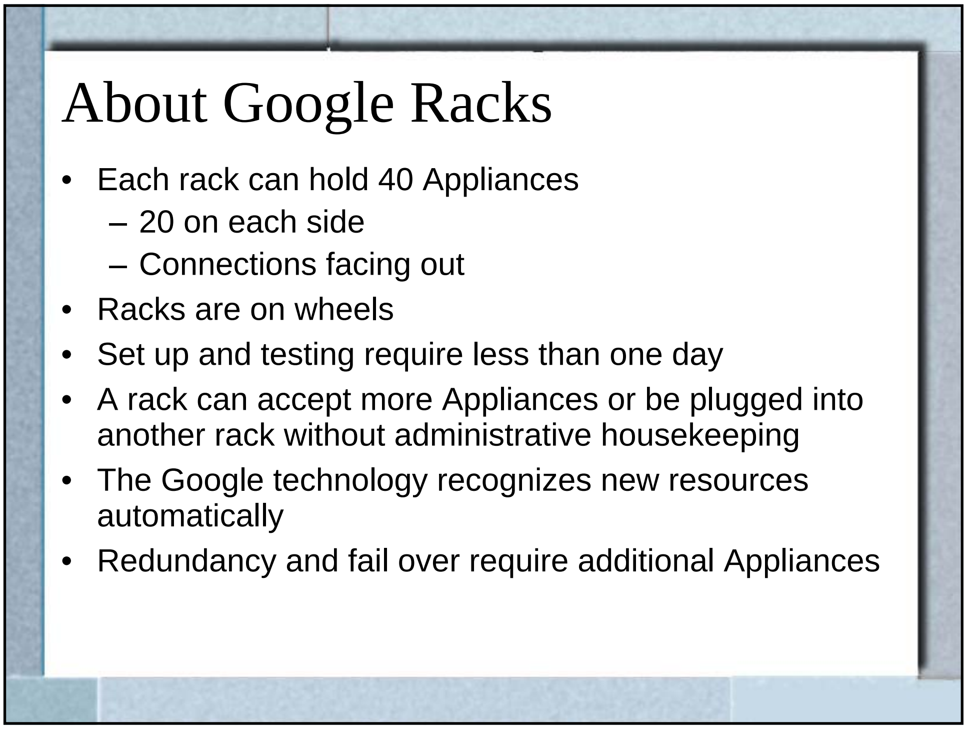
Small -- GB-1001
two mu (3.0 inches high)

Large – GB-8008



Version 4

- More APIs to allow Google “to play better with other enterprise applications”
- Clustering of documents
- Support for document boosting (a particular item appears on certain result pages)
- Enhanced security and access control tools
- Adaptors to allow easier indexing of certain content types found in enterprise systems



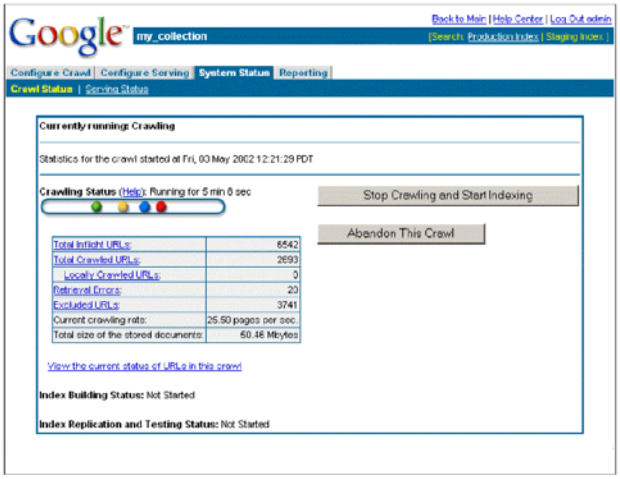
About Google Racks

- Each rack can hold 40 Appliances
  - 20 on each side
  - Connections facing out
- Racks are on wheels
- Set up and testing require less than one day
- A rack can accept more Appliances or be plugged into another rack without administrative housekeeping
- The Google technology recognizes new resources automatically
- Redundancy and fail over require additional Appliances

Pricing

- Begins at $3,000 for $100,000 documents support for second year is $995
- Your cost: depends on the number of documents
- An enterprise installation: $250,000 up to $1 million or more
- Resellers in the U.S. and Europe perform:
  - Pre-sale document analysis
  - Set up assistance (basically clicking on the folders that will be indexed)
  - Define collections
- Collections--document limit per collection

Typical Administrative Screen


Google
my_collection
Back to Main | Help Center | Log Out admin
[Search: Production Index | Staging Index]
Configure Crawl
Configure Serving
System Status
Reporting
Crawl Status | Serving Status
Stop Crawling and Start Indexing
Abandon This Crawl

Currently running: Crawling

Statistics for the crawl started at Fri, 03 May 2002 12:21:29 PDT

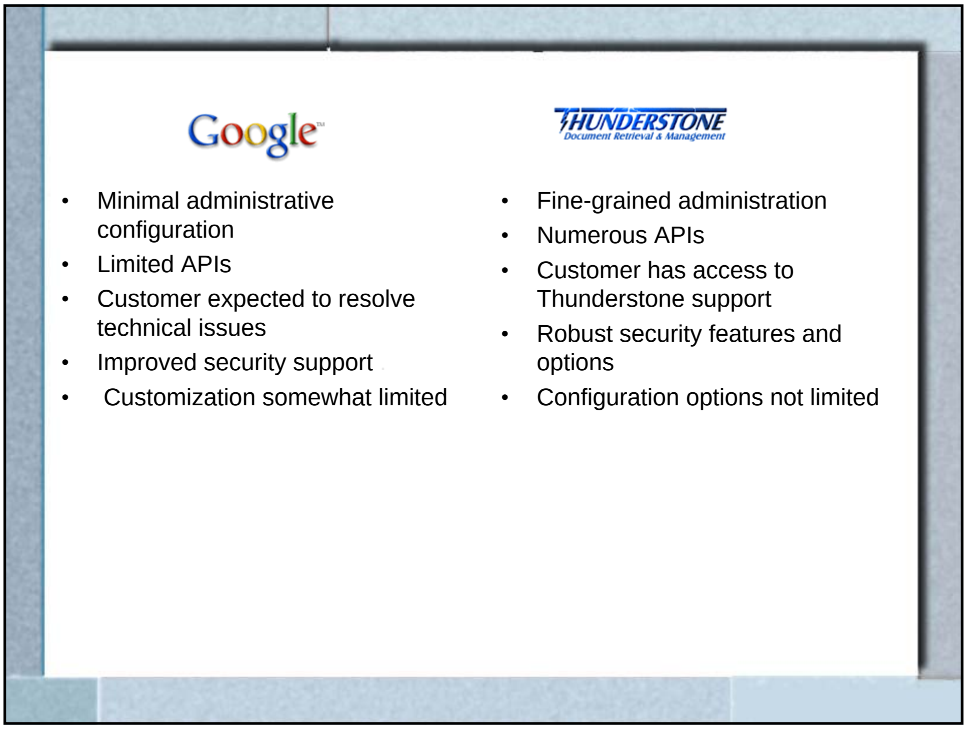
Crawling Status (Help): Running for 5 min 8 sec

Total Inflight URLs:	6542
Total Crawled URLs:	2693
Locally Crawled URLs:	0
Retrieval Errors:	20
Excluded URLs:	3741
Current crawling rate:	25.50 pages per sec.
Total size of the stored documents:	50.46 Mbytes

View the current status of URLs in this crawl

Index Building Status: Not Started

Index Replication and Testing Status: Not Started

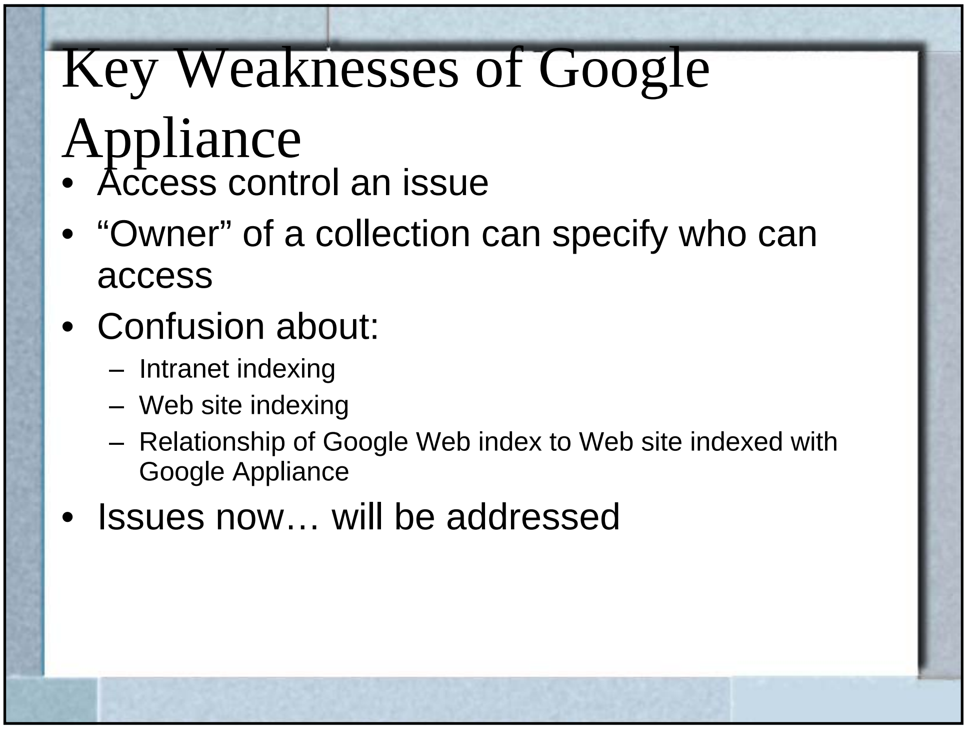


Google

- Minimal administrative configuration
- Limited APIs
- Customer expected to resolve technical issues
- Improved security support
- Customization somewhat limited

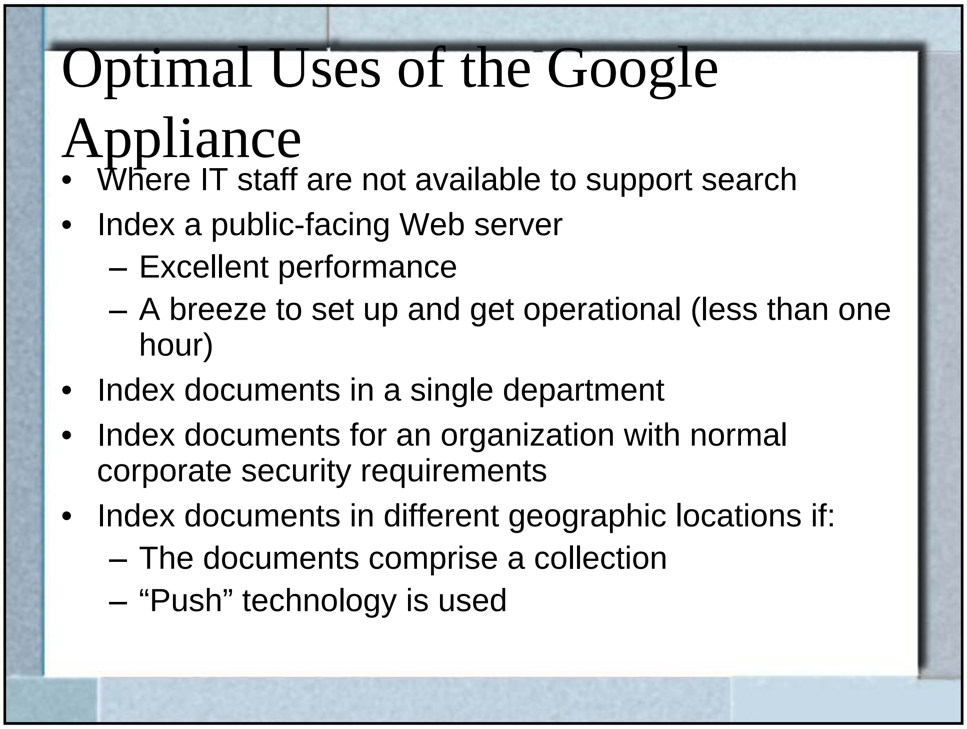
THUNDERSTONE
Document Retrieval & Management

- Fine-grained administration
- Numerous APIs
- Customer has access to Thunderstone support
- Robust security features and options
- Configuration options not limited



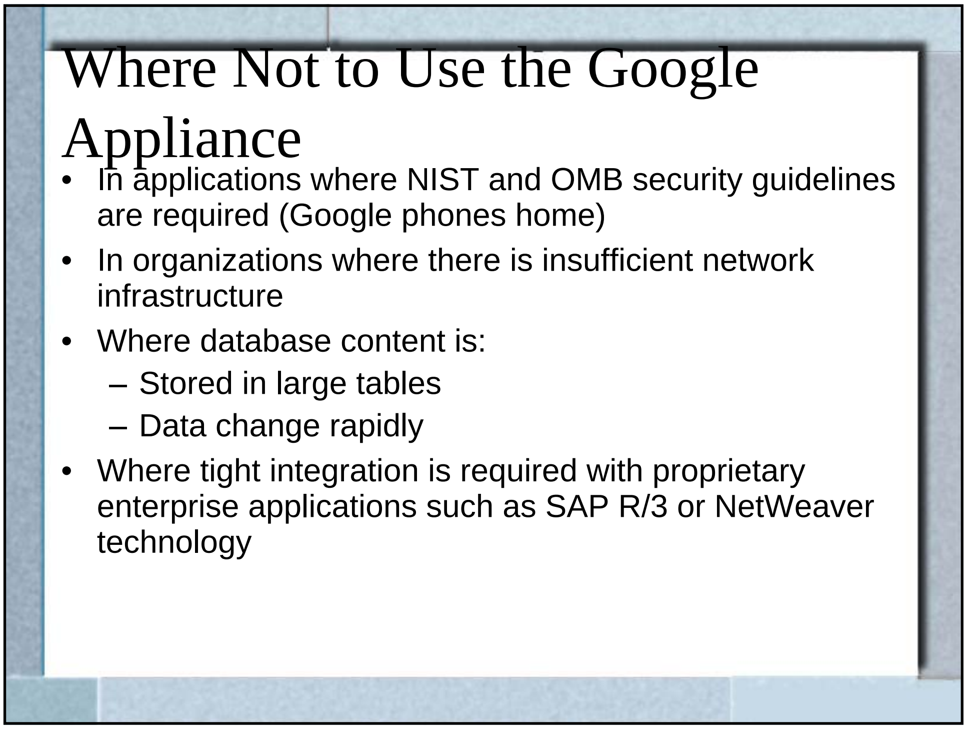
Key Weaknesses of Google Appliance

- Access control an issue
- “Owner” of a collection can specify who can access
- Confusion about:
  - Intranet indexing
  - Web site indexing
  - Relationship of Google Web index to Web site indexed with Google Appliance
- Issues now… will be addressed



Optimal Uses of the Google Appliance

- Where IT staff are not available to support search
- Index a public-facing Web server
  - Excellent performance
  - A breeze to set up and get operational (less than one hour)
- Index documents in a single department
- Index documents for an organization with normal corporate security requirements
- Index documents in different geographic locations if:
  - The documents comprise a collection
  - “Push” technology is used



Where Not to Use the Google Appliance

- In applications where NIST and OMB security guidelines are required (Google phones home)
- In organizations where there is insufficient network infrastructure
- Where database content is:
  - Stored in large tables
  - Data change rapidly
- Where tight integration is required with proprietary enterprise applications such as SAP R/3 or NetWeaver technology

What’s the Significance of the Appliance

- Extends the “keep it simple” philosophy of Google
- Puts Google in the product business
- Opens a new, potentially lucrative market where user dissatisfaction with existing products may be evident
- Demonstrates the flexibility of the core search technology
- Puts Google in the reseller business
- Linkage with enterprise desktop search possible

Match candidates to your requirements... then test in a “bake off”