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
4
Search System Vendors
**Safe Choices**

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

**Name Brands**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Assessment</th>
<th>License Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomy</td>
<td>Tools, systems, and services</td>
<td>Too soon to tell</td>
<td>$35,000+</td>
</tr>
<tr>
<td>FAST Search</td>
<td>ESP plus hosting</td>
<td>Customization usually required</td>
<td>$150,000+</td>
</tr>
<tr>
<td>Oracle TripleHop</td>
<td>SQL, Text, and taxonomy</td>
<td>Toolkit with Oracle services</td>
<td>$65,000/CPU</td>
</tr>
<tr>
<td>IBM</td>
<td>Do it all: Omnifind, iPhrase, and more</td>
<td>Integration and more integration</td>
<td>$500,000+</td>
</tr>
</tbody>
</table>
### Hot Engines

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Assessment</th>
<th>License Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endeca</td>
<td>Search plus taxonomy</td>
<td>Good. Work flow angle</td>
<td>$200,000+</td>
</tr>
<tr>
<td>Mondosoft</td>
<td>Search plus taxonomy plus analytics</td>
<td>Robust plus APIs</td>
<td>$50,000+</td>
</tr>
<tr>
<td>Vivisimo</td>
<td>Federation and deduplication</td>
<td>Integration solution</td>
<td>$65,000+</td>
</tr>
</tbody>
</table>

### Mondosoft “Hybrid” Interface
Hybrid Search: Facets, Hard Coding, Synonym Expansion

Vivisimo

Search in result sets

Meaty snippets
De-duped results
Preview option
## Special Purpose Tools

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Assessment</th>
<th>License Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inxight</td>
<td>Search plus text mining</td>
<td>A toolkit</td>
<td>$50,000+</td>
</tr>
<tr>
<td>Stratify</td>
<td>Search plus taxonomy</td>
<td>Set up for human input</td>
<td>$75,000+</td>
</tr>
</tbody>
</table>

### Inxight

![Inxight Diagram](image)
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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Autonomy</th>
<th>Verity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td>Bayesian</td>
<td>Linguistics</td>
</tr>
<tr>
<td>Architecture</td>
<td>IDOL server</td>
<td>Distributed servers</td>
</tr>
<tr>
<td>Special features</td>
<td>Discovers</td>
<td>Retrieves</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Autogenerated</td>
<td>Users input categories</td>
</tr>
<tr>
<td>Database support</td>
<td>Requires creating</td>
<td>Can read some database</td>
</tr>
<tr>
<td></td>
<td>documents for the system</td>
<td>files...slowly</td>
</tr>
<tr>
<td></td>
<td>to index</td>
<td></td>
</tr>
<tr>
<td>Strength</td>
<td>Certain processes are</td>
<td>Strong security and work</td>
</tr>
<tr>
<td></td>
<td>fast</td>
<td>flow</td>
</tr>
<tr>
<td>Weaknesses</td>
<td>Hit and miss relevance</td>
<td>Requires much manual</td>
</tr>
<tr>
<td></td>
<td></td>
<td>configuration</td>
</tr>
</tbody>
</table>
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)
- Small -- GB-1001
  two mu (3.0 inches high)
- Medium – GB-5005
- 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
Minimal administrative configuration
- Limited APIs
- Customer expected to resolve technical issues
- Improved security support
- Customization somewhat limited

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”