

Thoughts on Federated and Aggregated Search Architectures

Enterprise Search Summit 2010 NY
Avi Rappoport, Search Tools Consulting
www.searchtools.com / consult2@searchtools.com

Thoughts on Federated and Aggregated Search Architectures

Information Scatter

- Internal
 - Local and remote file shares
 - Email
 - CMS /DMS
 - Application portals
 - Knowledge bases
 - Multimedia, digital assets
- External
 - Research papers and other gated content
 - Public-facing sites
 - The Web

Avi Rappoport / Enterprise Search Summit NY / May 2010 / consult2@searchtools.com

2

Thoughts on Federated and Aggregated Search Architectures

Solution 1: Federate Searching

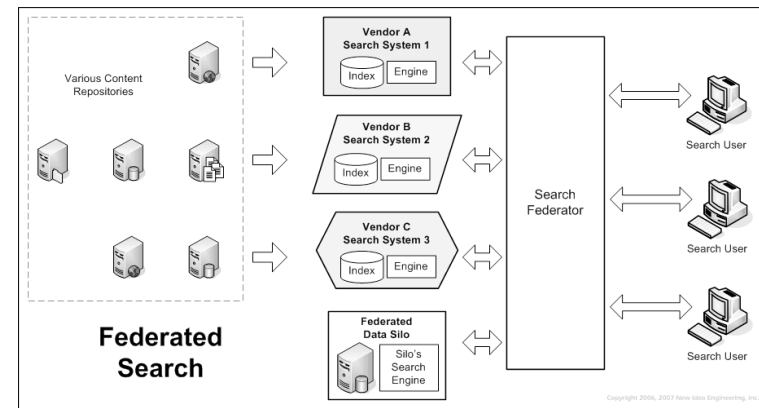
- aka "MetaSearch"
- Single Search Interface
 - Accepts queries and converts to various formats
 - Sends queries to multiple external search engines
 - Includes user authentication
- Collects result lists
 - In external search relevance order
- Collates and sorts by relevance
 - Single list or in panels
 - Can be dynamically updated

Avi Rappoport / Enterprise Search Summit NY / May 2010 / consult2@searchtools.com

3

Thoughts on Federated and Aggregated Search Architectures

Federated Search Diagram



by New Idea Engineering, ideaeng.com

Avi Rappoport / Enterprise Search Summit NY / May 2010 / consult2@searchtools.com

4

Thoughts on Federated and Aggregated Search Architectures

Federated Results: Apple Site

Store Items

Product info

Support pages

Thoughts on Federated and Aggregated Search Architectures

Federated Results: Science.gov

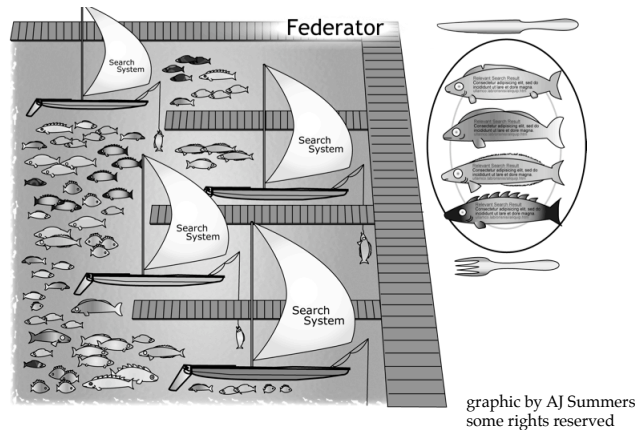
Special sources

Main Results

Dynamic facets from results clustering

Thoughts on Federated and Aggregated Search Architectures

Federate at Search Time



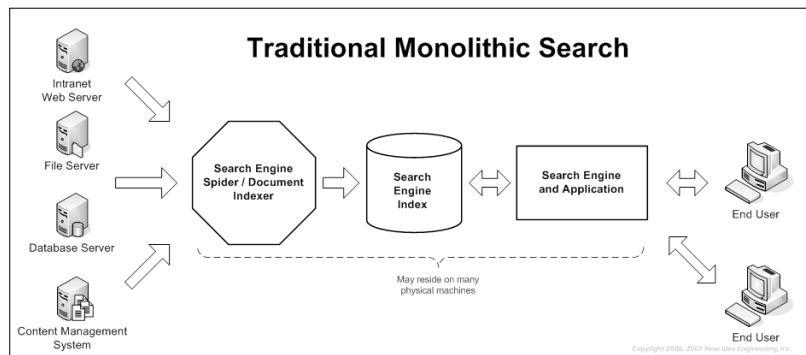
Thoughts on Federated and Aggregated Search Architectures

Solution 2: Aggregate Indexing

aka "Unified Information Access"

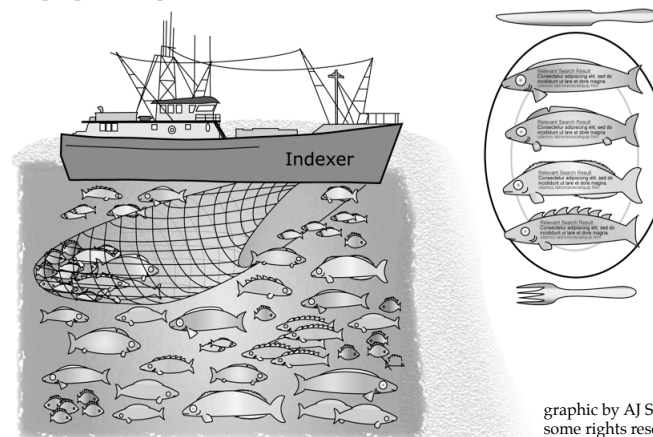
- Gather all possible data
 - Robot crawlers on intranets
 - RSS blog feeds
 - Automated connectors
 - Custom scripts
- Store in a single index
 - Include access control information
- Simple to search all at once

Aggregated Search Diagram



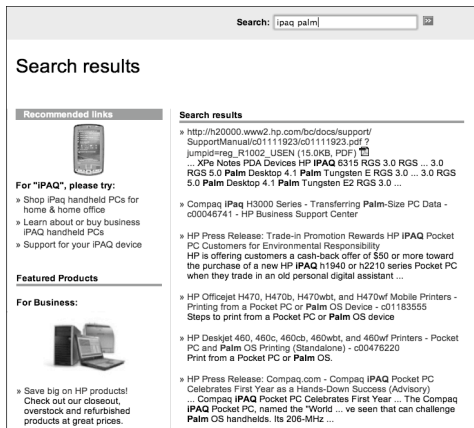
by New Idea Engineering, ideaeng.com

Aggregate At Index Time



graphic by AJ Summers
some rights reserved

Aggregate Results: HP.com



Sources of Content

Federating

- Lotus Notes
- News feeds and archives
- Legal: Westlaw, Lexis
- Government Documents
 - Patents, Census
 - Multi-national materials
- Academic journal portals
- Large social networks
- The Web

Aggregating

- Enterprise intranets
- File servers
- Sharepoint
- CMS/DMS
 - usually have awful search
- Data warehouses
 - Current CRM
- Legal discovery

Preparation Process

Federating

- Analyze sources
 - Test search connectors
- Store source information
 - Open Archives Initiative
 - Taxonomy
- Minimal bandwidth

Aggregating

- Index
 - Match data connectors
 - Open each file or record
 - Tokenize, stem words
 - De-duplicate
 - Store words and documents
- Scale issues
 - Hardware and software
 - Bandwidth requirements

Keeping Content Current

Federating:

- Source internal updates
 - Near-real-time
- Content may change
 - Between queries
 - No notification

Aggregating:

- Depend on connectors
 - Frequent polling
 - Automated notification
 - Programmatic triggers
 - Re-crawling
- Merging Updates
 - Scale issues
- Content may change
 - Between index runs
 - Some notification

Security & Access Control

Federating

- Send user credentials with query
 - Depends on source security
 - Automation can be hard
- Always current

Aggregating

- Early-binding
 - Index and store ACL info
 - Update index on changes
- Late-binding
 - For each result
 - Send authorization request
 - OK if item is allowed
 - Repeat until 10 items are allowed

Search and Retrieval Process

Federating

- Convert & send query
 - z39.50, RDW
 - HTTP, Web Services
 - OAI - Open Archives
 - Custom connectors
- Network and source speed
- Collect results
 - Standardized formats, XML
 - Screen-scraping
- Cache frequent results

Aggregating

- Single syntax
- No delay
- Results in standard format
- Cache frequent results

Relevance Ranking

Federating

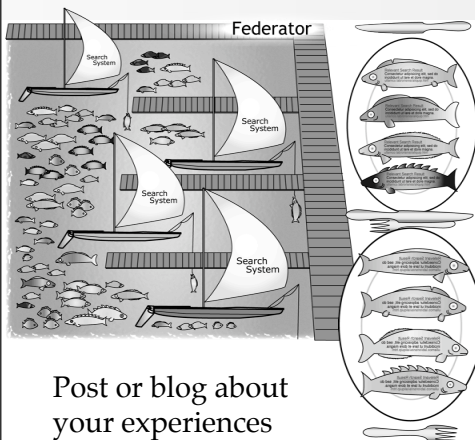
- Combine results listings
 - Duplicate detection here
- Overall source relevance
- May re-rank
 - Source ranking is quirky
 - Based on metadata
 - User activities

Aggregating

- One results listing
 - Still may need de-duping
- Get single relevance rank
 - Very fast
 - IDF: Inverse Document Frequency
 - Rare words in index
 - Boost for more matches

Checklist Per Data Source

- Federation
 - Source interaction at search time
 - External content, databases, channels
 - Always-current content and access control
 - Slower response time, tricky relevance
- Aggregation
 - Source interaction at index time
 - Very large index files
 - Content and access control updates trickier
 - Fast response time, straightforward relevance



Be open-minded,
analyze the benefits
of each approach for
each data source.

Post or blog about
your experiences