Limited Access Document

Requirements for Enterprise Search

December 2, 2004

© Stephen E. Arnold, 2004

ARNOLD
Information
Technology

Agenda for This Section

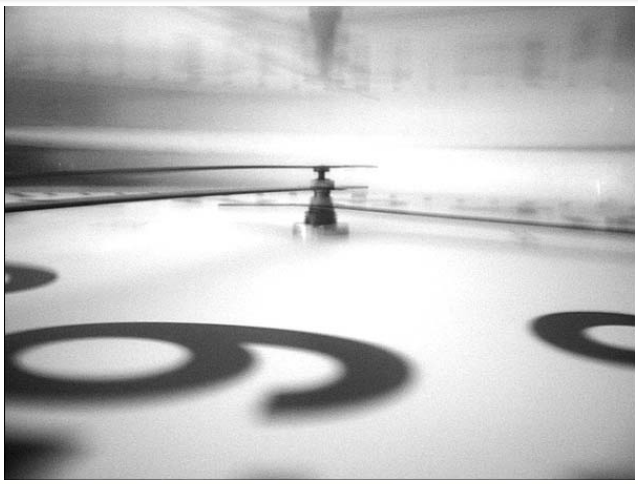
- Background
- Reputation analysis
- Wrap up

Questions at any time…

Why Requirements Get Ignored

- **Time pressure**
- **Management demand**
- **Underestimate the challenges of search**
- **Misunderstand scaling / performance issues**
- **Assume that software can index, classify, and metatag**

Rush Carefully



Why Requirements Get Ignored

- Time pressure
- **Management demand**
- Underestimate the challenges of search
- Misunderstand scaling / performance issues
- Assume that software can index, classify, and metatag

Dilbert Principle

Why Requirements Get Ignored

- Time pressure
- Management demand
- **Underestimate the challenges of search**
- Misunderstand scaling / performance issues
- Assume that software can index, classify, and metatag

Misjudge Task

Why Requirements Get Ignored

- Time pressure
- Management demand
- Underestimate the challenges of search
- Misunderstand scaling / performance issues
- **Assume that software can index, classify, and metatag**

Software… Not Always “Smart”

Why Requirements Are Important

- **Search is complicated even with a "search toaster" from Google or Thunderstone**
- **Indexing and updates can slow a network so no other work can be done**
- **Human involvement is needed. Humans cost money**
- **Scaling is**
  - **Expensive**
  - **Slower than most people believe or accept**

Google "Search Toaster"

GB-1001

GB-5005

GB-8008

No Administration… Just Plug It In

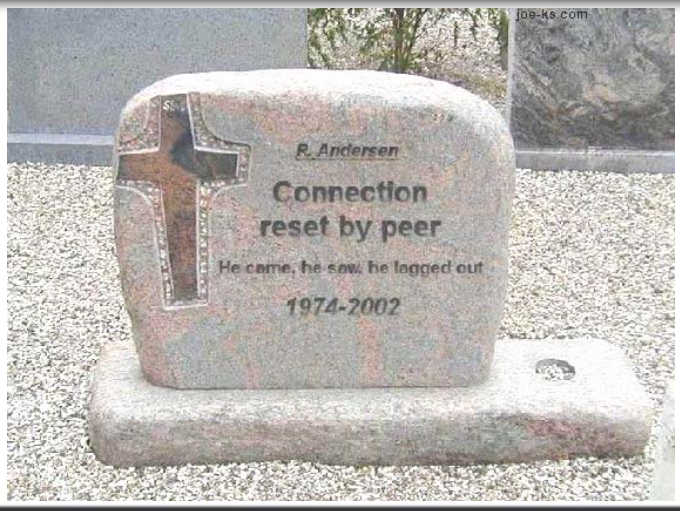
The system administrator's friend… Plug in and go home.

False Drops… No Hits…

Why Requirements Are Important

- Search is complicated even with a “search toaster” from Google or Thunderstone
- **Indexing and updates can slow a network so no other work can be done**
- Human involvement is needed. Humans cost money
- Scaling is
  - Expensive
  - Slower than most people believe or accept

Unresponsive Network



Why Requirements Are Important

- Search is complicated even with a “search toaster” from Google or Thunderstone
- Indexing and updates can slow a network so no other work can be done
- **Human involvement is needed. Humans cost money**
- Scaling is
  - Expensive
  - Slower than most people believe or accept

Humans Needed

Why Requirements Are Important

- Search is complicated even with a “search toaster” from Google or Thunderstone
- Indexing and updates can slow a network so no other work can be done
- Human involvement is needed. Humans cost money
- **Scaling is**
  - **Expensive**
  - **Slower than most people believe or accept**

Plan for a Big System

What Must Be Investigated?

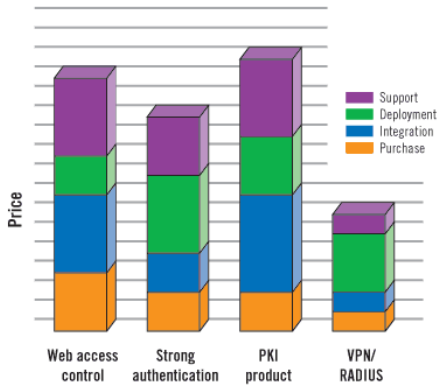
- **Content that will be indexed?**
- Access to that content?
  - Security
  - Work flow
- What user interface?
  - Search box
  - Work flow (saved searches)
  - Yahoo-style facets
- Index update frequency?

Humans Needed

What Must Be Investigated?

- Content that will be indexed?
- **Access to that content?**
  - **Security**
  - **Work flow**
- What user interface?
  - Search box
  - Work flow (saved searches)
  - Yahoo-style facets
- Index update frequency?

Log On

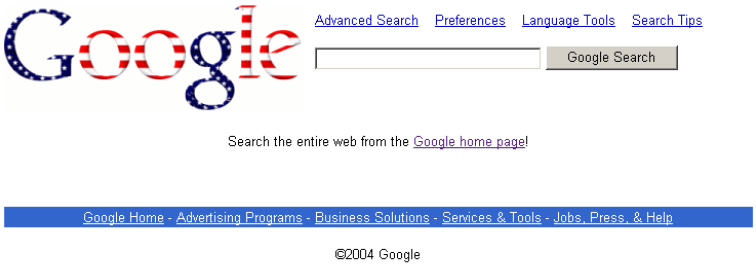


http://www.securecomputing.com/index.cfm?skey=939

What Must Be Investigated?

- Content that will be indexed?
- Access to that content?
  - Security
  - Work flow
- **What user interface?**
  - **Search box**
  - **Work flow (saved searches)**
  - **Yahoo-style facets**
- Index update frequency?

Google

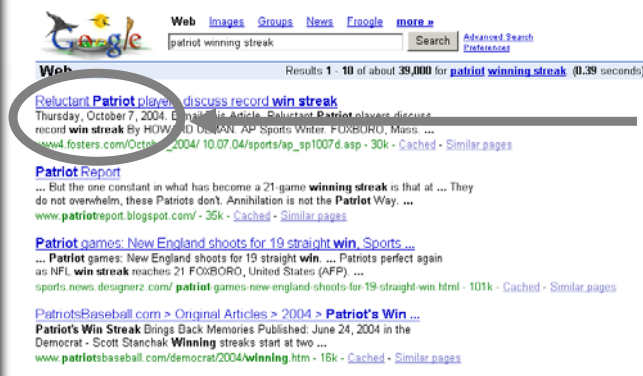


Yahoo

What Must Be Investigated?

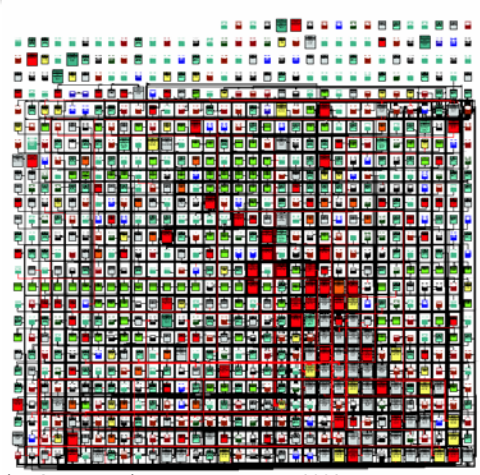
- Content that will be indexed?
- Access to that content?
  - Security
  - Work flow
- What user interface?
  - Search box
  - Work flow (saved searches)
  - Yahoo-style facets
- **Index update frequency?**

Update One Hour after Upset

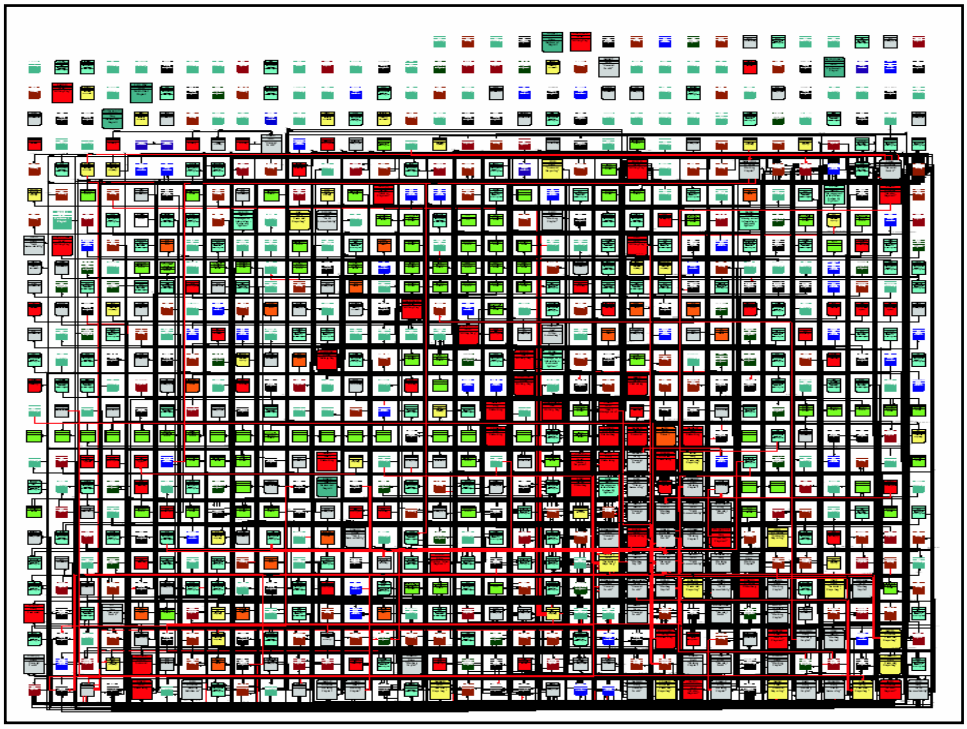


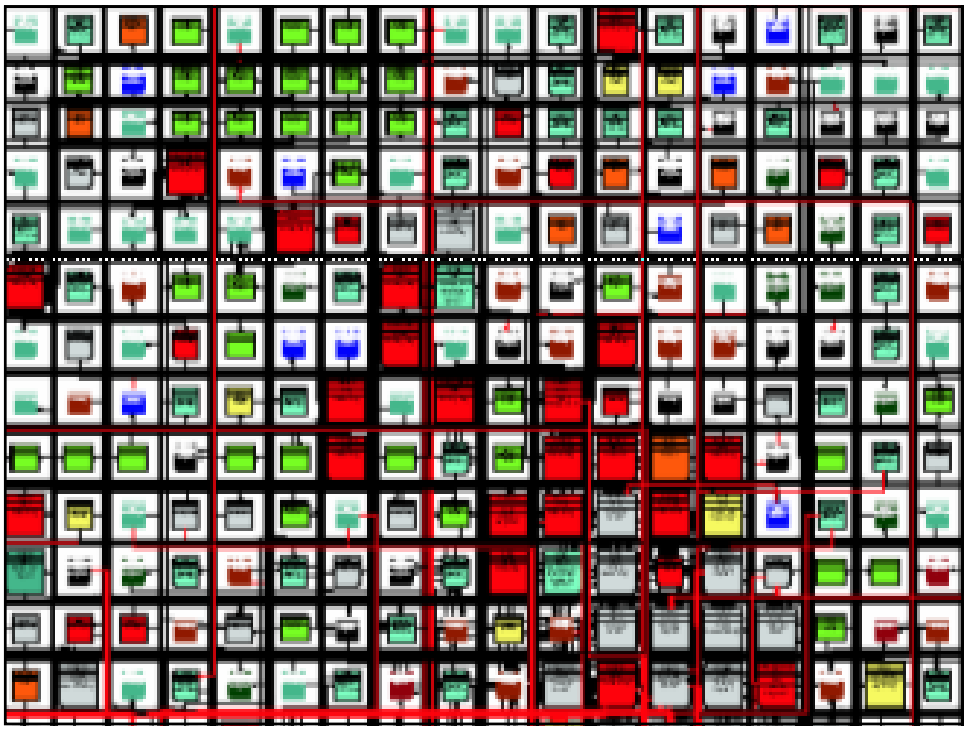
No update within one hour of Patriots' loss.

Red=Sensitive, Green=Not Sensitive, Other=?



Source: Catherine Santana, Director FMMP, DoD, 2002





What Must Be Investigated?

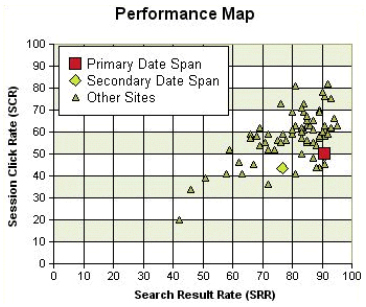
- **Who can access the system?**
- What usage tracking is needed?
- What are the file types to be indexed?
- What legacy systems' content must be indexed?
- Will source documents be served from:
  - A data warehouse
  - The server or machine where documents reside?

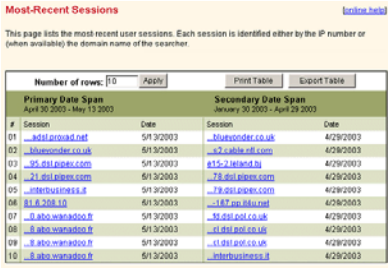
Jobs, Not Individuals

What Must Be Investigated?

- Who can access the system?
- **What usage tracking is needed?**
- What are the file types to be indexed?
- What legacy systems' content must be indexed?
- Will source documents be served from:
  - A data warehouse
  - The server or machine where documents reside?

Metrics

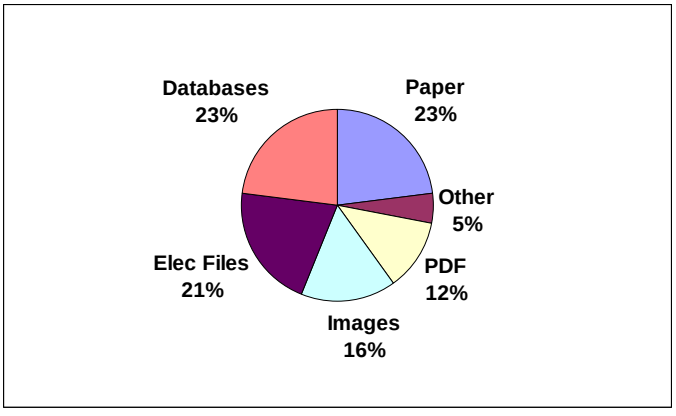




What Must Be Investigated?

- Who can access the system?
- What usage tracking is needed?
- **What are the file types to be indexed?**
- What legacy systems' content must be indexed?
- Will source documents be served from:
  - A data warehouse
  - The server or machine where documents reside?

Document Formats

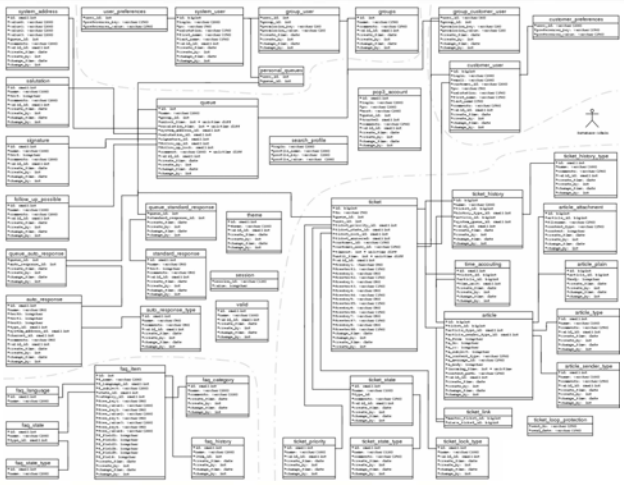


Source: AMR Research, September 2004

Databased Information… High Value

2.3	.52	.54	.80	1.1	.29	1.1	2.2	1.0	.26	1.6	.89	2.1	3.1	1.3	4.0	.72
2.8	1.2	2.9	2.2	1.7	.94	1.8	2.9	4.0	.23	4.5	1.9	4.0	4.3	4.5	7.0	.79
.60	.33	.27	.67	.51	.58	.10	1.1	.76	.41	2.0	1.0	1.1	1.7	.49	1.1	1.0
.17	.36	.70	.95	.60	.29	.61	.62	1.3	.47	.49	.49	.99	1.2	.48	1.7	.41
1.6	.03	.77	.15	.82	.99	.84	2.3	.86	1.4	1.7	.98	1.2	2.7	.18	1.7	1.6
.02	1.5	2.6	1.3	2.1	.80	.50	1.5	1.5	.89	4.4	1.7	1.0	1.1	1.7	1.9	.67
.83	.08	.31	.22	.17	.48	.02	.82	.17	.24	.63	.27	.59	.68	.05	.87	.50
1.2	.26	.76	.12	.12	.35	.12	.43	.56	.87	.49	.61	.29	1.3	.67	.94	1.0
1.5	.04	.72	.35	.40	.87	.44	1.3	.83	1.5	2.4	.63	.73	2.0	.20	.45	1.1
.86	.14	.00	.41	.29	.17	.64	.64	.43	.10	.53	.28	.61	1.4	.02	2.1	.60
.33	.03	.33	.35	.07	.48	.14	.43	.86	.10	.70	.84	.23	.19	.04	.79	.46
.75	.02	.88	.77	.42	.24	1.1	1.0	1.9	.40	.21	.58	2.4	1.2	.86	2.0	.07
1.1	.07	.14	.61	.48	.08	.89	1.2	1.0	.97	.05	.71	1.4	2.1	.34	3.5	.85
3.0	.74	1.2	.91	1.6	.13	2.8	2.7	1.9	1.3	.57	1.5	3.5	5.2	1.4	6.6	1.5
1.2	.02	.70	.15	.14	.30	.40	1.2	.05	.90	.01	.43	.93	1.8	.15	1.2	1.1
.30	.76	1.1	.75	1.1	.51	.40	.14	.82	.22	2.0	1.2	1.0	1.2	.98	2.2	.19
1.4	.10	.08	.70	.81	.93	1.3	2.1	.63	1.8	.88	1.2	2.3	2.9	1.0	3.1	1.4

Structured and Unstructured Needed



What Must Be Investigated?

- Who can access the system?
- What usage tracking is needed?
- What are the file types to be indexed?
- **What legacy systems' content must be indexed?**
- Will source documents be served from:
  - A data warehouse
  - The server or machine where documents reside?

Legacy Systems

What Must Be Investigated?

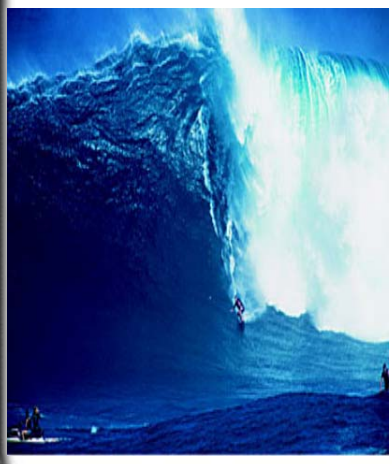
- Who can access the system?
- What usage tracking is needed?
- What are the file types to be indexed?
- What legacy systems' content must be indexed?
- **Will source documents be served from:**
  - **A data warehouse**
  - **The server or machine where documents reside?**

Storage, Not a Single Hard Drive

Challenges

- Data waves, not floods
- Requirements versus budget
- “Nice to have” versus “must have”
- Knowledge versus assumptions
- Experience versus inexperience
- Time versus expectations
- Vendor’s promises versus system realities
- Existing infrastructure versus search systems’ storage and computational hunger

Data: Waves, not Floods



- Large volumes of data
- Fast moving
- Changes
  - Quite important
  - Major challenge

Requirements Pitfalls

- **Users explain search in terms of Google, not their specific work-related needs**
- Customization
- Resources limited; therefore, risk increases
- Too little research; therefore, decisions are guesses
- Ignore / underestimate:
  - Copyright issues
  - Access procedures
  - Indexing issues

The Google Effect

Requirements Pitfalls

- Users explain search in terms of Google, not their specific work-related needs
- **Customization**
- Resources limited; therefore, risk increases
- Too little research; therefore, decisions are guesses
- Ignore / underestimate:
  - Copyright issues
  - Access procedures
  - Indexing issues

Customization Costs

Requirements Pitfalls

- Users explain search in terms of Google, not their specific work-related needs
- Customization
- **Resources limited; therefore, risk increases**
- Too little research; therefore, decisions are guesses
- Ignore / underestimate:
  - Copyright issues
  - Access procedures
  - Indexing issues

Adequate Resources… Essential

Requirements Pitfalls

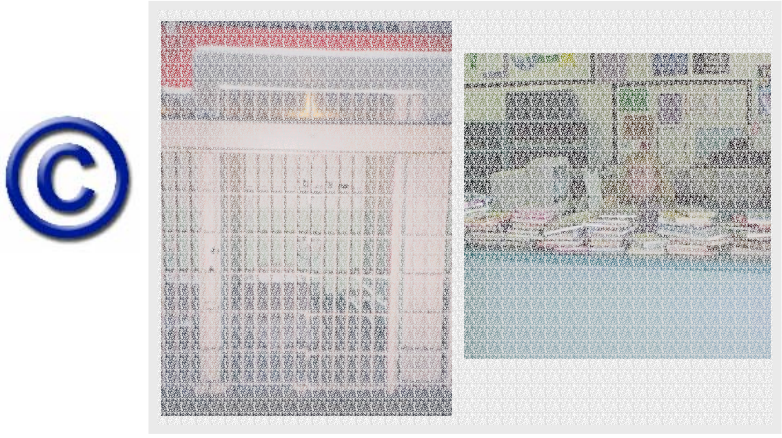
- Users explain search in terms of Google, not their specific work-related needs
- Customization
- Resources limited; therefore, risk increases
- **Too little research; therefore, decisions are guesses**
- Ignore / underestimate:
  - Copyright issues
  - Access procedures
  - Indexing issues

Guesses Not Good

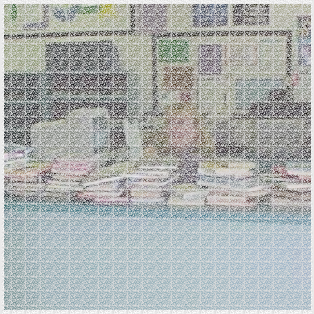
Requirements Pitfalls

- Users explain search in terms of Google, not their specific work-related needs
- Customization
- Resources limited; therefore, risk increases
- Too little research; therefore, decisions are guesses
- **Ignore / underestimate:**
  - **Copyright issues**
  - **Access procedures**
  - **Indexing issues**

Copyright… Access… Manual Work



Copyright… Access… Manual Work



Copyright… Access… Manual Work

Lessons Learned

- **No free.lunch… search can and will top $1 million in its first year in mid-sized organiztions**
- **Fruitcake data… data harmonization requires manual effort plus software**
- **Silos of data… update and security issues persist**
- **Costs go up faster than inflation; for example, 212 percent over an 18 month period**
- **Technical problems can be expensive to solve**

The Key to Rocket Science

Lessons Learned

- No freelunch… search will easily top $1 million in its first year in mid-sized organiztions
- **Fruitcake data… data harmonization requires manual effort plus software**
- Silos of data… update and security issues persist
- Costs go up faster than inflation; for example, 212 percent over an 18 month period
- Technical problems can be expensive to solve

Data Fruitcake

Lessons Learned

- No freelunch… search can easily top $1 million in its first year in mid-sized organiztions
- Fruitcake data… data harmonization requires manual effort plus software
- **Silos of data… update and security issues persist**
- Costs go up faster than inflation; for example, 212 percent over an 18 month period
- Technical problems can be expensive to solve

One Data Silo Means…

More Silos… Not Integration

Lessons Learned

- No freelunch… search can easily top $1 million in its first year in mid-sized organiztions
- Fruitcake data… data harmonization requires manual effort plus software
- Silos of data… update and security issues persist
- **Costs go up faster than inflation; for example, 212 percent over an 18 month period**
- Technical problems can be expensive to solve

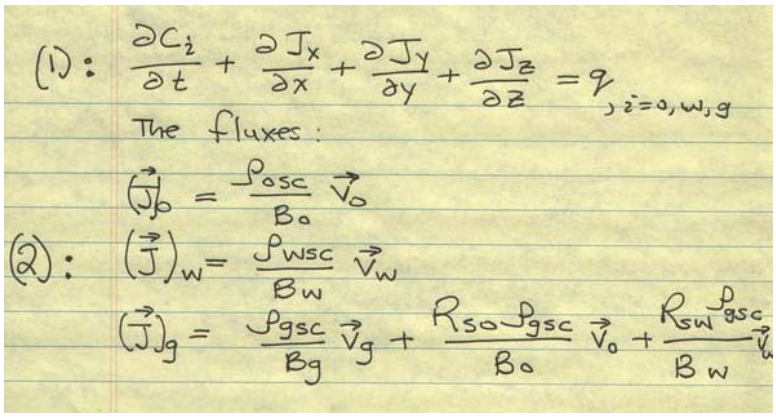
Costs Can Explode

Nomar / A-Rod / Jeter

Lessons Learned

- No freelunch… search can easily top $1 million in its first year in mid-sized organiztions
- Fruitcake data… data harmonization requires manual effort plus software
- Silos of data… update and security issues persist
- Costs go up faster than inflation; for example, 212 percent over an 18 month period
- **Technical problems can be expensive to solve**

Complicated Problems



Requirements Tips

- **Form a search "team". Work as a team.**
- **Identify the key stakeholders. Discuss their search needs.**
- **Use a Web survey for general information.**
- **Communicate to keep expectations realistic.**
- **If search will be embedded in work flow, concentrate data collection where search touches existing processes.**
- **If resources are available, consider outside requirements specialists.**

A Team … Search Has Many Parts

Requirements Tips

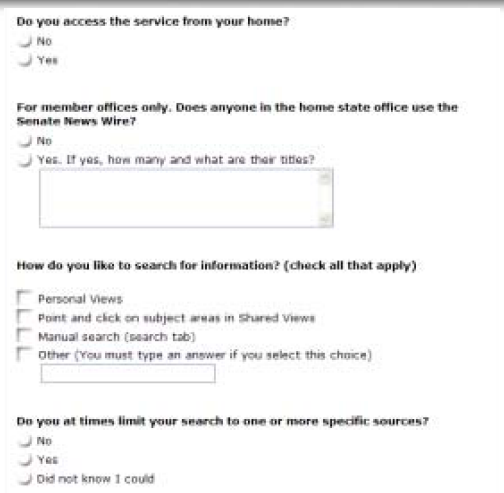
- Form a search "team". Work as a team.
- **Identify the key stakeholders. Discuss their search needs.**
- Use a Web survey for general information.
- Communicate to keep expectations realistic.
- If search will be embedded in work flow, concentrate data collection where search touches existing processes.
- If resources are available, consider outside requirements specialists.

Find the Stakeholders

Requirements Tips

- Form a search "team". Work as a team.
- Identify the key stakeholders. Discuss their search needs.
- **Use a Web survey for general information.**
- Communicate to keep expectations realistic.
- If search will be embedded in work flow, concentrate data collection where search touches existing processes.
- If resources are available, consider outside requirements specialists.

Use a Web Survey



Do you access the service from your home?
- No
- Yes

For member offices only. Does anyone in the home state office use the Senate News Wire?
- No
- Yes. If yes, how many and what are their titles?

How do you like to search for information? (check all that apply)
- Personal Views
- Point and click on subject areas in Shared Views
- Manual search (search tab)
- Other (You must type an answer if you select this choice)

Do you at times limit your search to one or more specific sources?
- No
- Yes
- Did not know I could

Requirements Tips

- Form a search “team”. Work as a team.
- Identify the key stakeholders. Discuss their search needs.
- Use a Web survey for general information.
- **Communicate to keep expectations realistic.**
- If search will be embedded in work flow, concentrate data collection where search touches existing processes.
- If resources are available, consider outside requirements specialists.

Search… No Pot of Gold

Requirements Tips

- Form a search “team”. Work as a team.
- Identify the key stakeholders. Discuss their search needs.
- Use a Web survey for general information.
- Communicate to keep expectations realistic.
- **If search will be embedded in work flow, concentrate data collection where search touches existing processes.**
- If resources are available, consider outside requirements specialists.

Focus Where Search Hits Work Flow

Requirements Tips

- Form a search “team”. Work as a team.
- Identify the key stakeholders. Discuss their search needs.
- Use a Web survey for general information.
- Communicate to keep expectations realistic.
- If search will be embedded in work flow, concentrate data collection where search touches existing processes.
- **If resources are available, consider outside requirements specialists.**

Consider Consultants… Save Time

Tomorrow…the “Semantic Web”?