

# Tracking Service Prototype

(<http://prawn.lanl.gov/tracking>)

---

Xiaoming Liu, Jewel Ward & Herbert Van de Sompel

Share Our Knowledge Session (SOKS)

Research Library, Los Alamos National Laboratory

6 October 2003

# Agenda

- Description of Focused Crawling
- Motivation & Getting There
- Overview of Envisional
- Use Case 1: “LANL Sentiment”
- Use Case 2: “Awareness Service”
- Open issues
- What’s Next
- Q&A

# Focused Crawling

- Web crawlers, robots, or spiders.
  - Web robots are programs that traverse the web automatically. Web crawlers are the basis of web search engines.
- What's a focused crawler?
  - A focused crawler is designed to only gather documents on a specific topic; it's also called a topical crawler.

# Examples of Focus Crawling

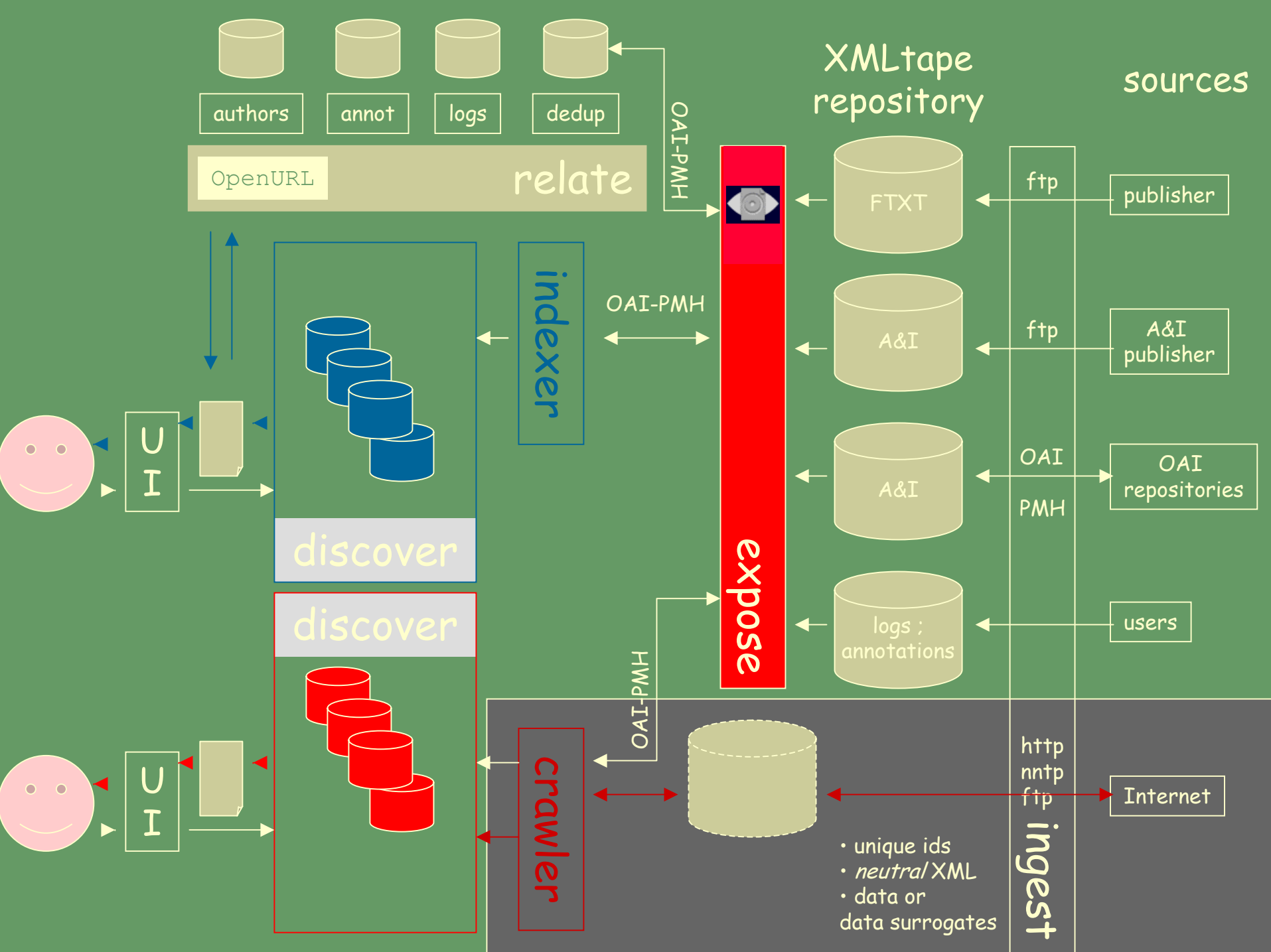
- Personal homepage (ahoy!).
- Computer science research papers in ps/pdf format (ResearchIndex/CiteSeer).
- Email addresses for spam.
- All pages which use Coca-Cola's logo (Envisional).
- Limited-sized collections (20) of high precision with respect to the topic. (Cornell, NSDL).
- Focused crawling under librarian's supervision (Infomine).

# Advantages of Focused Crawlers

- Hand-curated collections will not scale with the web.
- Web search engine is not complete.
- Low update rate of web search engine.
- Post-processing.

# Major Challenges

- How to determine whether a downloaded page is on-topic? And, in what order should URLs be visited?
- Focused crawling is very resource demanding.



# Motivation

What services can we provide to our customers using focused crawl results?

- Vertical portal (such as nanotechnology)?
- Personal collection?
  - Awareness Service
- Monitoring a special topic in the web?
  - LANL sentiment

# How to Get There

- Build vs. buy
  - Will we have to build a focused crawl tool?
  - Or, can we buy or “borrow” one?  
Envisional
- Start small, go larger/broader
  - LANL Sentiment
  - Awareness Service
  - Vertical portal (?)

# What is Envisional?

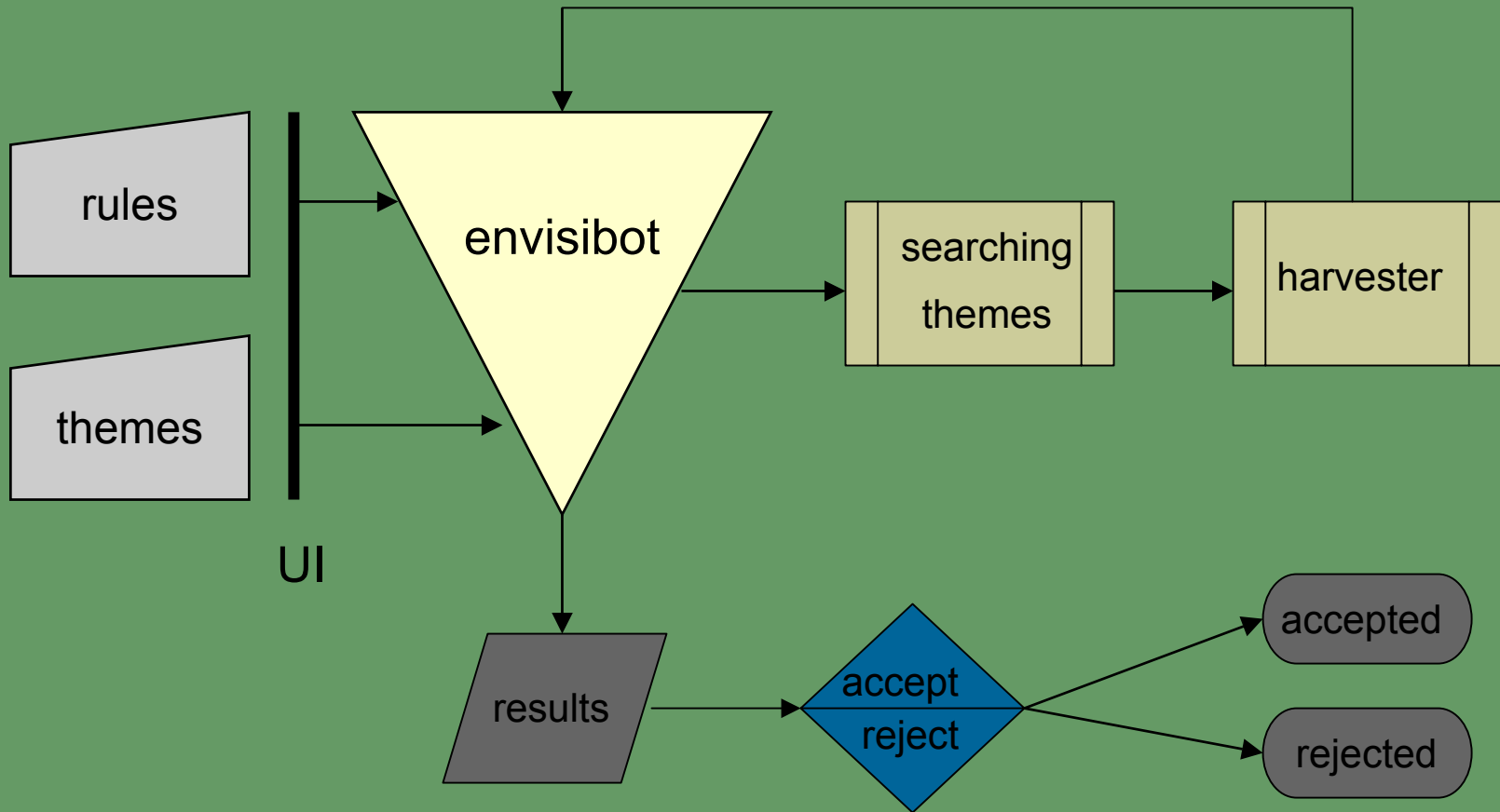
- A software company in the UK
- A type of software (UI and backend)
- A “discovery language” that uses “intelligent agents” to retrieve information
- Combines IR (relevance and precision) with AI (reasoning and knowledge representation)
- Searches for images and text

# What is Envisional?

## Key terms

- Theme
- Rule
- Envisibot or 'Bot
- Harvester

# What is Envisional?



# Note

The use cases described here for possible services are still “works in progress”.

# Use Case: LANL “Sentiment”

- Purpose
  - News analysis management tool
  - Can determine sentiment of news about Lab
- Code
  - recall vs. precision
  - simple vs. complex

# Use Case: LANL “Sentiment”

- Theme – variations of LANL
- Rules
  - define LANL-related terms to look for
  - positive and negative words, 3 levels
  - downgrade terms, plus remove other labs
  - Brand, probabilities and weighting for score
  - Negative, positive or neutral based on score

# Use Case: LANL “Sentiment”

## Results

- Positive vs. negative vs. neutral
- Major news sources vs. minor news sources

# LANL "Sentiment" Demo

Results for: lanl-sent.s1-20ct03 Positive\_Sentiment - Mozilla

http://prawn.lanl.gov/fcgi/envadmin.fcgi

Search

Home Bookmarks Google Google News Weather MyLib Add2MyLib D-Lib Magazine Ariadne Magazine Oxford English Dictio... Encyclopædia Britan... Association of Resea... TinyURL!

The Jakarta Site - The Jakarta Project ... SunSITE - DSpace for Dummies Using Apa... The Tomcat 4 Servlet/JSP Container - Cla... Results for: lanl-sent.s1-20ct03 Positive... Prototyping Team Website

Delete Hide from Classification Hide from All Results

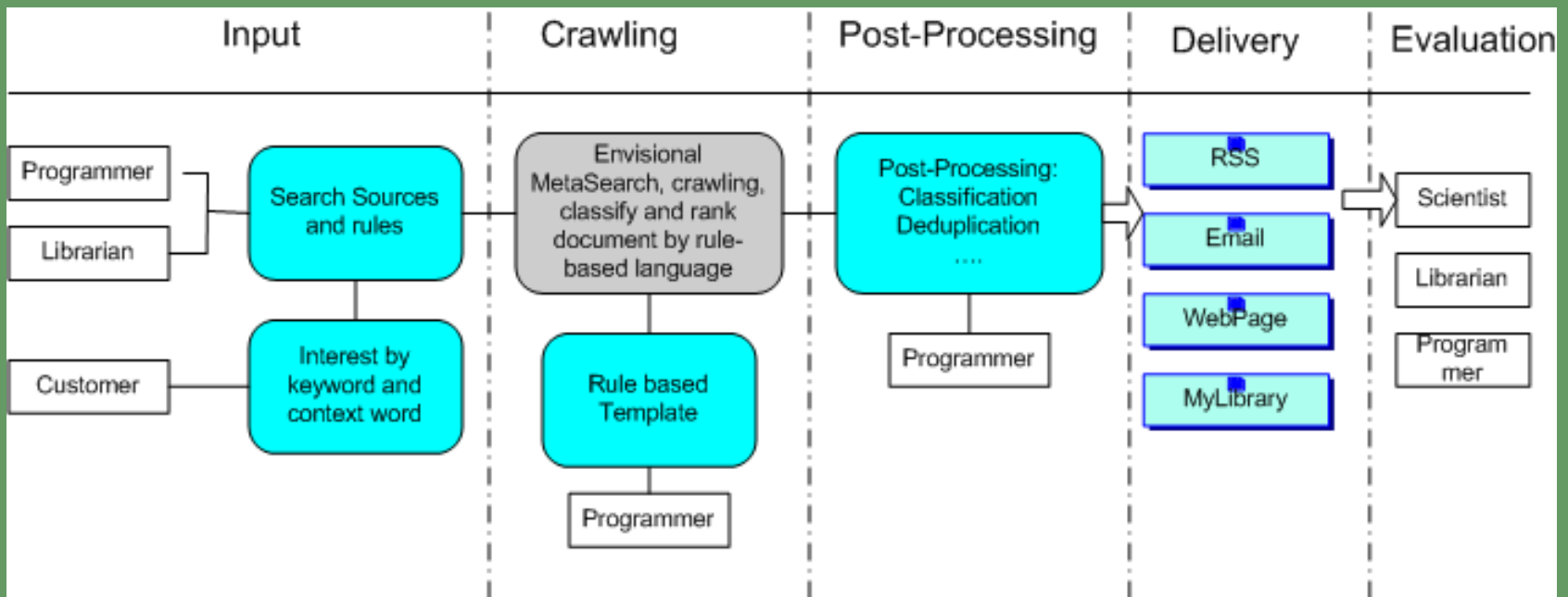
No.	Score	Title / Link (Group by host)
1.	3552	<a href="http://www.dailyca.org/article.asp?id=12952">The Daily Californian (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.dailyca.org/article.asp?id=12952 Document ID:7741687759450790481
2.	2974	<a href="http://www.signonsandiego.com/news/education/20031002-9999_1n2dynes.html">SignOnSanDiego.com &gt; News &gt; Education -- New UC system president has left an innovative mark (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.signonsandiego.com/news/education/20031002-9999_1n2dynes.html Document ID:8620734111718169217
3.	2381	<a href="http://www.kobtv.com/index.cfm?viewer=storyviewer&amp;id=5063&amp;cat=NIMTOPSTORIES">HTML_BeginTemplate_Templates.kobhome.dwt.HEAD.BeginEditable (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.kobtv.com/index.cfm?viewer=storyviewer&id=5063&cat=NIMTOPSTORIES Document ID:1235956622737398322
4.	2335	<a href="http://www.collegiatetimes.com/index.php?ID=2118">Collegiate Times (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.collegiatetimes.com/index.php?ID=2118 Document ID:7502434029246732906
5.	1967	<a href="http://www.guardian.co.uk/Archive/Article/0,4273,4503478,00.html">Guardian Unlimited   Archive Search (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.guardian.co.uk/Archive/Article/0,4273,4503478,00.html Document ID:2679501040797633014
6.	1931	<a href="http://kobtv.com/index.cfm?viewer=storyviewer&amp;id=5057&amp;cat=BUSINESS">HTML_BeginTemplate_Templates.kobhome.dwt.HEAD.BeginEditable (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://kobtv.com/index.cfm?viewer=storyviewer&id=5057&cat=BUSINESS Document ID:7644297417508903540
7.	1880	<a href="http://www.spaceref.com/news/viewpr.html?pid=12496">Los Alamos Hosts Gamma-Ray Burst Anniversary Conference   SpaceRef - Your Space Reference (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.spaceref.com/news/viewpr.html?pid=12496 Document ID:2863726413055112363
8.	1816	<a href="http://austin.bizjournals.com/austin/stories/2003/09/22/daily14.html?st=b_in_hl">Los Alamos lab uses TippingPoint systems - 2003-09-23 - Austin Business Journal (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://austin.bizjournals.com/austin/stories/2003/09/22/daily14.html?st=b_in_hl Document ID:2470083658125269623
9.	1626	<a href="http://www.guardian.co.uk/Archive/Article/0,4273,4017035,00.html">Guardian Unlimited   Archive Search (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.guardian.co.uk/Archive/Article/0,4273,4017035,00.html Document ID:6258173894696913396
10.	1603	<a href="http://www.tallahassee.com/mld/democrat/news/local/6878676.htm">Tallahassee Democrat   09/28/2003   Mag lab works for new grant (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.tallahassee.com/mld/democrat/news/local/6878676.htm Document ID:7704251587548273287
11.	1546	<a href="http://news.bbc.co.uk/1/hi/world/americas/800444.stm">BBC News   AMERICAS   'No evidence' of Los Alamos spying (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://news.bbc.co.uk/1/hi/world/americas/800444.stm Document ID:5565886189476660971
12.	1526	<a href="http://releases.usnewswire.com/GetRelease.asp?id=153-09302003">U.S. Newswire - DOE Awards Key Cleanup Contracts to Small Businesses; Section 8 (a) Businesses to Perform Cleanup Work at Small Sites (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://releases.usnewswire.com/GetRelease.asp?id=153-09302003 Document ID:2858096913520908956
13.	1513	<a href="http://www.abqtrib.com/archives/news03/092303_news_prince.shtml">Albuquerque Tribune Online (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.abqtrib.com/archives/news03/092303_news_prince.shtml Document ID:2016627470644402787
14.	1357	<a href="http://www.nytimes.com/2003/09/30/obituaries/30ROSE.html?ex=1065585600&amp;en=d380ff80f74bfc04&amp;ei=5006&amp;partner=ALTAVISTA1">M. N. Rosenbluth, 76, an H-Bomb Developer, Dies (Cache) (Edit) (Email Link) (Host Lookup) (Source)</a> ✓ http://www.nytimes.com/2003/09/30/obituaries/30ROSE.html?ex=1065585600&en=d380ff80f74bfc04&ei=5006&partner=ALTAVISTA1 Document ID:5277515075837352671

Done

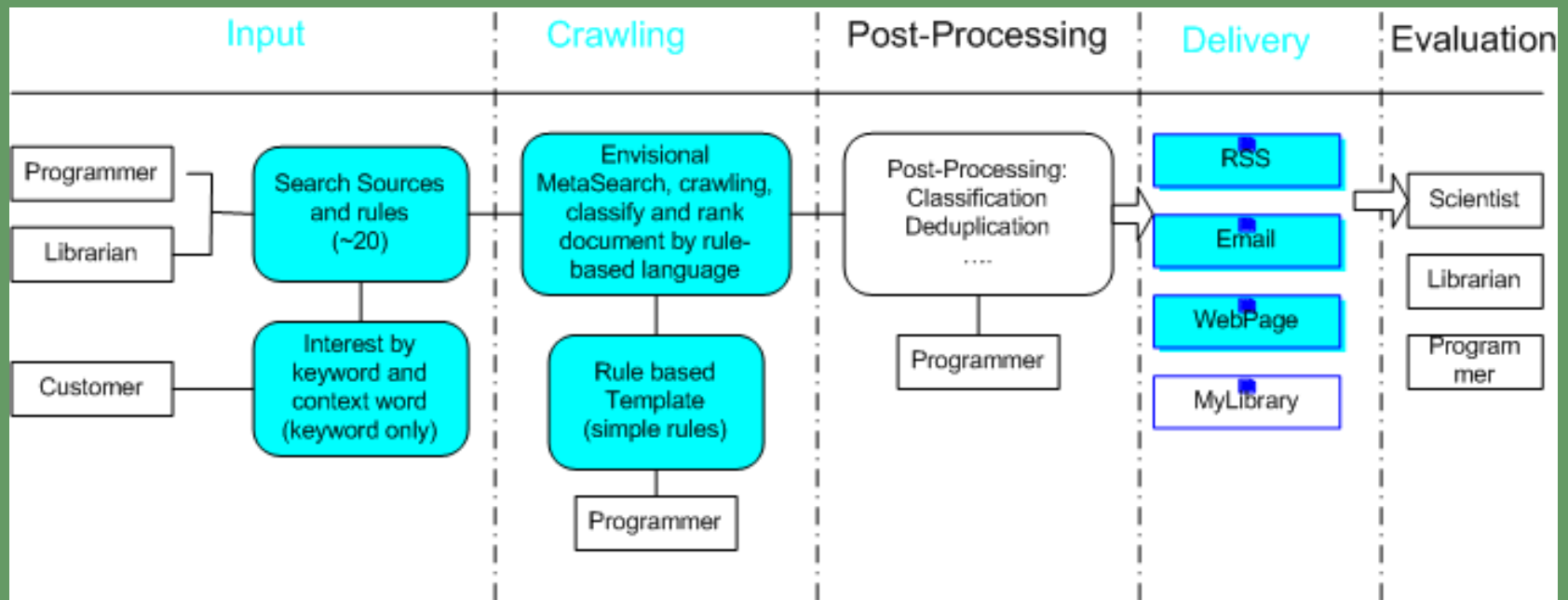
# Why Another Awareness Service

- Coverage.
  - Internal and important external resources.
  - Delivered in unified way.
- Precision.
  - Better than keyword based system.
- Comparison with other systems.
  - GoogleAlert.
  - Scisearch alert.
  - Federated searching.

# Architecture



# Current Status



# Demonstration



## My Weather

☀ **Santa Fe:**

Clear sky 23.0 C



## LANL-RL Alerts

- [Context-Sensitive Linking, Serials Review, Vol: 28, Issue: 4](#)
- [An Interview with Steve Shadle, Serials Review, Vol: 28, Issue: 4](#)
- [The OpenURL, Serials Review, Vol: 28, Issue: 1](#)
- [Useful or Useless Use Statistics? A Summary of Conference Presentations on Usage Data from the 22nd Annual Charleston Conference, Issues in Book and Serial Acquisition, Serials Review, Vol: 29, Issue](#)
- [Special section: metadata, Library Collections, Acquisitions, and Technical Services, Vol: 26, Issue: 3](#)
- [The Role of the ISSN in the Electronic Linking Environment, Serials Review, Vol: 29, Issue: 2](#)



## Mail summary

<a href="#">Outbox</a>	0/0
<a href="#">Inbox</a>	53/99



## Appointments

 [10:30 02 October, KDD Forum](#)



## Tasks

# Key Issues

- Identify scientific search engines and add to Envisional.
- Template of rules.
- Scalability.
- Post-Processing.
- Integration with local system
  - Email
  - RSS
  - HTML
  - MyLibrary?

# Identify Scientific Resources

Table 1: Status Report

Name	Status	Notes
ACM	Pass	
ACS	Fail	URL encoding
AIP	Pass	
APS	Fail	Session Problem
arXiv.ph	Pass	
arXiv.cs	Pass	
arXiv.math	Pass	
arXiv.nlin	Pass	
biomedical	Pass	
biosis.lanl.gov	Pass	
CERN	Pass	
DOEEnergy.lanl.gov	Pass	
EngIndex.lanl.gov	Pass	
Google	Pass	
GoogleNews	Pass	
Highwire	Pass	
IEEE Computer	Pass	
IEEE Explorer	Fail	URL encoding problem
inspec.lanl.gov	Pass	
IOP	Pass	
Jane	Fail	URL redirect
JSTORE	Fail	Search Result not shown directly
Nature	Pass	
pubMed	Pass	
sciencserver.lanl.gov	Pass	
scienceDirect	fail	no subscription
Scirus	pass	
scisearch.lanl.gov	Fail	URL Encoding
siam	Fail	Donot accept Get method
spicdl.aip.org	Pass	
Springer	Fail	Session

# Scalability

- 2000 Awareness Profiles
- 20 search engines
- weekly Awareness Feed
- each search engine returns 50 hits with 5 new.

<b>Weekly</b>	<b>Queries</b>	<b>Downloads</b>
<b>Search Engine</b>	2,000	10,000
<b>Awareness System</b>	40,000	200,000

# Open Issues

- Identify important search sources
- Scalability
- Exploit advanced search interface
- Precision/recall
- Evaluation

# What's Next

- Evaluation of prototype.
  - <http://prawn:8080/alert/index.html>
- New version of Envisional?
- Advanced Envisional template.
- Post-processing.
- Vertical portal.

Questions?