

# electronic Knowledge Management (eKM) Today

Automating concept extraction to make it easier to see the big picture and digest the specifics, with examples in the areas of: expertise identification, document categorization, historical archival, and current politics

Jorge H. Román (JHR@lanl.gov)

A. Shelly Spearing (SHELLYS@lanl.gov)

Los Alamos National Laboratory, High Performance Computing Division

DIGIN

Digital Preservation Conference

June 4-6, 2008



Operated by Los Alamos National Security, LLC for DOE/NNSA

UNCLASSIFIED

LA-UR-08-2421



## Knowledge Management

- Knowledge Management ('KM') comprises a range of practices used by organizations to identify, create, represent, and distribute knowledge. *Wikipedia.org*
- Often approached from two distinct angles:
  - Human Resources: training, mentorship, retention, succession planning
  - Information Technology: concept extraction, data mining, information visualization, decision support, other computer-based tools
- Many information systems contain human-generated knowledge (i.e., codified "lessons learned"). We are extending these technologies.



Operated by Los Alamos National Security, LLC for NNSA

UNCLASSIFIED

LA-UR-08-2421

Slide 2

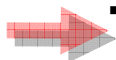


## The Goal

- Quickly access *relevant* information
  - ...without being distracted by *irrelevant* information
- Be able to
  - Synthesize related facts
  - Form hypotheses
  - Draw conclusions
  - Articulate additional questions
  - Make decisions with confidence

## Challenges of the Electronic Information Age

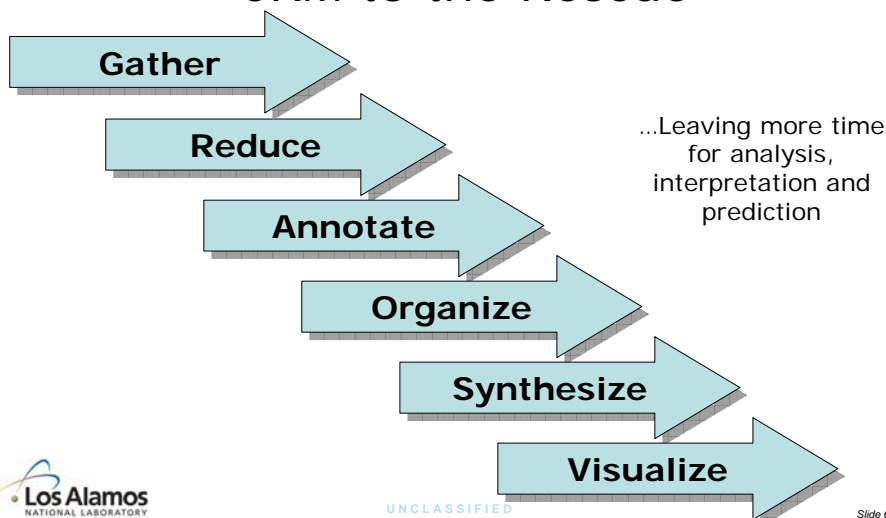
- The ever-increasing volume of data presents challenges for providers and for users
  - Providers
    - Security
    - Computing resources (Processors, Bandwidth, Storage)
  - Users
    - Providing context increases precision, but screens out many good matches
    - General searches yield too many documents and it is impossible to sift through them all
    - A typical “knowledge hunt” necessitates dozens of queries, conducted in series, taking lots of time and yielding results in a somewhat disjointed order



## Solution: Smarter Tools

- Facilitate assimilation of e-content by providing tools that:
  - Allow broad searches but only retrieve highly-relevant content
  - Organize and annotate results, focusing attention on "the good stuff"
  - Extract key concepts to allow a big-picture view at both the collection and document levels
  - Automate linkage among documents where none existed prior to the search

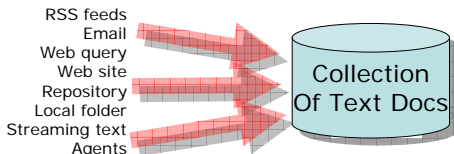
## eKM to the Rescue



## eKM to the Rescue (continued)

### Gather

- Broad and context-rich searches retrieve fewer, but more relevant, documents



### Reduce

- Compression from full text to kSig
- Generate document summary
- Post-processing automatically identifies and extracts main themes in document collection
- Further eliminate documents that don't fall within the collection "core"

### Annotate

- Display goodness-of-fit score vs. targeted concepts
- Color-code main themes and other targeted concepts to focus attention
- Link to an automatically-generated document summary

### Organize

- Logical organization by time, author, collection, theme frequency
- Timeline representation
- Generate full-text index

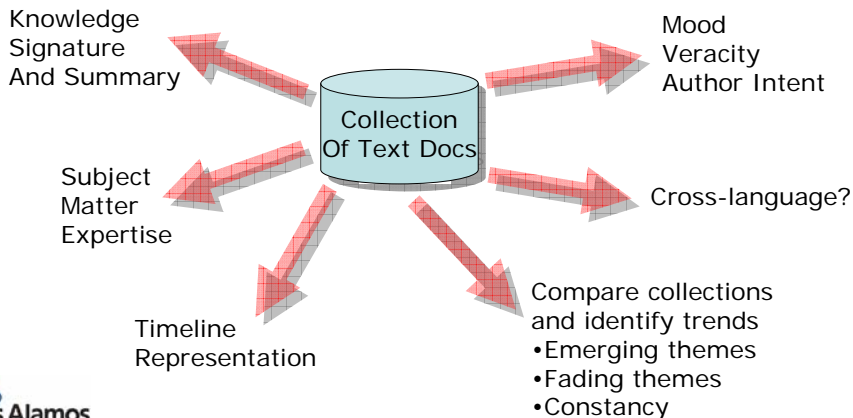
### Synthesize

- Collection-level "Big Picture"
- Compute trends
- Identify core concepts
- Compute cross-linkage to relevant previously-unrelated documents

### Visualize

- Reveal trends
- Reveal outliers and anomalies

## Choose Outputs Based on Need





## Examples

- 1000 documents → 1 page of trends
  - Emerging themes, fading themes, concept consistency
- 1000+ email messages → Subject Matter Expertise
  - Identification of SME in various areas based upon recurring themes in correspondence
- 172,000 documents → Manageability
  - Distilled to 22,000 somewhat-topically-related documents (for further drill-down and analysis)



Slide 9



## Live Demo

Explaining the next few slides



Slide 10

Sample Knowledge Signature (kSig) for historical record

**RADIATION PROTECTION DURING THE EARLY STAGES OF SITE DECOMMISSIONING AT THE UKAEA'S DOUNREAY SITE**

P.J. Thompson, W. Sinclair, D. Mowat, S. White, R. Kerr, T. Chalmers and S.M. Calder  
 The United Kingdom Atomic Energy Authority, Dounreay, Thurso, Caithness, KW14 7TZ.  
 RWE NUKEM Ltd, Dounreay, Thurso, Caithness, KW14 7TZ.

In 1998 the United Kingdom Government announced that the United Kingdom Atomic Energy Authority (UKAEA) site at Dounreay in northern Scotland would no longer be seeking further commercial reprocessing contracts. This decision laid down the foundations for the UKAEA to focus firmly on the task of decommissioning the UKAEA's site at Dounreay. Fifty to sixty years has been identified as the period in which to decommission the site and restore its environment.

The business of decommissioning at Dounreay presents a number of interesting challenges that need to be addressed. There are a number of complex and unique projects that must be undertaken including the decommissioning of the early fast reactors, a materials test reactor, metallurgical laboratories (housing fume cupboards, glove boxes and shielded cells) and novel fuel reprocessing plants. This paper discusses the experience gained during the various stages of decommissioning in the fast reactors and nuclear fuel cycle reprocessing plants, focusing on the Dounreay Fast Reactor, the Prototype Fast Reactor, a critically test facility, fuel reprocessing plants, laboratories and associated environment.

The UKAEA at Dounreay is dedicated to restoring the environment both safely and cost-effectively. This paper discusses the practical radiation protection issues that have been met during the early stages of a number of decommissioning projects on the site. The conference presentation will give an update on our experience and discuss lessons learnt.

**INTRODUCTION**  
 The United Kingdom Atomic Energy Authority (UKAEA) site at Dounreay was opened in 1955, and was built on a former Admiralty airfield and adjacent farmland. Dounreay was instrumental in developing the United Kingdom's knowledge of fast reactors. It is home to both the Dounreay Fast Reactor (DFR) and the Prototype Fast Reactor (PFR). The DFR went critical in November 1959, supplying electricity power for commercial use from October 1962 and the Prototype Fast Reactor (PFR) went critical in 1974. PFR was the postulated forerunner of large-output commercial fast reactors and an important facility within the European collaborative programme.

Original content with Color-coded themes



Operated by Los Alamos National Security, LLC for NNSA

LA-UR-08-2421

Slide 11



Timeline of President's Bush speeches related to Iraq

Theme analysis of text from the White House Press Archive on Iraq. The timeline shows the top-level themes for the 475 documents.

	02	03	04	05	06	07
ATTACK	ATTACKS	ATTACKS			ATTACKS	ATTACKS
COMMENTS	COMMENTS	COMMENTS				
COUNCIL			COUNCIL		CONGRESS	CONGRESS
COUNTRY			COUNTRY		COUNTRY	COUNTRY
ECONOMY	ECONOMY	ECONOMY		ECONOMY	ECONOMY	ECONOMY
GOVERNMENT	GOVERNMENT	GOVERNMENT	GOVERNMENT	GOVERNMENT	GOVERNMENT	GOVERNMENT
HISTORY	HISTORY	HISTORY	HISTORY	HISTORY	HISTORY	HISTORY
HOMELAND SECURITY	HOMELAND SECURITY	HOMELAND SECURITY				
IMMEDIATE	IMMEDIATE	IMMEDIATE				
INSPECTORS	INSPECTORS	INSPECTORS				
IRAQ	IRAQ	IRAQ	IRAQ	IRAQ	IRAQ	IRAQ
IRAQI	IRAQI	IRAQI	IRAQI	IRAQI	IRAQI	IRAQI
MIDDLE EAST	MIDDLE EAST	MIDDLE EAST	MIDDLE EAST	MIDDLE EAST	MIDDLE EAST	MIDDLE EAST
NATION	NATION	NATION	NATION	NATION	NATION	NATION
NATIONAL SECURITY	NATIONAL SECURITY	NATIONAL SECURITY	NATIONAL SECURITY	NATIONAL SECURITY	NATIONAL SECURITY	NATIONAL SECURITY
NEWS	NEWS	NEWS	NEWS	NEWS	NEWS	NEWS
NUCLEAR WEAPONS	NUCLEAR WEAPONS	NUCLEAR WEAPONS				
PEACE	PEACE	PEACE				
PRESIDENT	PRESIDENT	PRESIDENT	PRESIDENT	PRESIDENT	PRESIDENT	PRESIDENT
PRESIDENT BUSH	PRESIDENT BUSH	PRESIDENT BUSH	PRESIDENT BUSH	PRESIDENT BUSH	PRESIDENT BUSH	PRESIDENT BUSH
PRIME MINISTER	PRIME MINISTER	PRIME MINISTER	PRIME MINISTER	PRIME MINISTER	PRIME MINISTER	PRIME MINISTER
REGIME	REGIME	REGIME	REGIMES			
REPORT	REPORT	REPORT				
RESOLUTION	RESOLUTION	RESOLUTION				
SADDAM	SADDAM	SADDAM				
SADDAM HUSSEIN	SADDAM HUSSEIN	SADDAM HUSSEIN				
SECURITY	SECURITY	SECURITY	SECURITY	SECURITY	SECURITY	SECURITY
SECURITY COUNCIL	SECURITY COUNCIL	SECURITY COUNCIL				
SECURITY COUNCIL RES	SECURITY COUNCIL RES	SECURITY COUNCIL RES				
TERROR	TERROR	TERROR	TERROR	TERROR	TERROR	TERROR
TERRORISTS	TERRORISTS	TERRORISTS	TERRORISTS	TERRORISTS	TERRORISTS	TERRORISTS
THREAT	THREAT	THREAT				
UNITED NATIONS	UNITED NATIONS	UNITED NATIONS				
UNITED STATES	UNITED STATES	UNITED STATES	UNITED STATES	UNITED STATES	UNITED STATES	UNITED STATES
WEAPONS	WEAPONS	WEAPONS				
AFGHANISTAN	AFGHANISTAN	AFGHANISTAN				
CITIZENS	CITIZENS	CITIZENS	CITIZENS	CITIZENS	CITIZENS	CITIZENS
COALITION	COALITION	COALITION	COALITION	COALITION	COALITION	COALITION
EXECUTION	EXECUTION	EXECUTION				
FREEDOM	FREEDOM	FREEDOM	FREEDOM	FREEDOM	FREEDOM	FREEDOM
GOVERNING COUNCIL	GOVERNING COUNCIL	GOVERNING COUNCIL				
HEALTH CARE	HEALTH CARE	HEALTH CARE				
MILITANTS	MILITANTS	MILITANTS				
NATIONAL EMERGENCY	NATIONAL EMERGENCY	NATIONAL EMERGENCY				
PATRIOT ACT	PATRIOT ACT	PATRIOT ACT				
SUPPORT	SUPPORT	SUPPORT	SUPPORT	SUPPORT	SUPPORT	SUPPORT

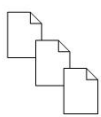

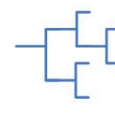

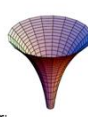



Operated by Los Alamos National Security, LLC for NNSA

LA-UR-08-2421

Slide 12



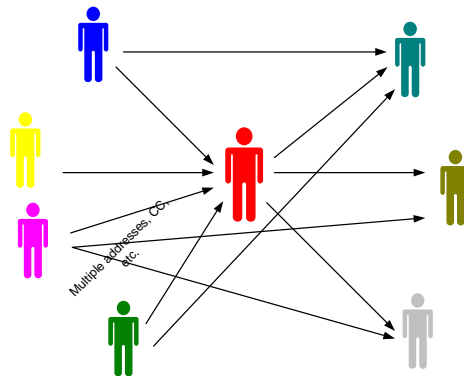
1	2	3	4	5	eKM Order
Automated summarization of conceptual content	Management and definition of collections	Efficient creation of large knowledge sets (hierarchical concepts)	Efficient comparison of large knowledge sets	Presentation of inferred knowledge to aid assimilation	Capability
<p><b>kSig Computation</b></p>  <ul style="list-style-type: none"> <li>- Extract text</li> <li>- Infer Ksig (XML)</li> <li>- Paragraph annotation</li> <li>- Create kSig User Interface (UI)</li> </ul>	<p><b>Collection Management</b></p>  <p>Record into repository:</p> <ul style="list-style-type: none"> <li>- kSig</li> <li>- meta-data</li> </ul> <p>Collection book-keeping</p>	<p><b>Taxonomy</b></p>  <p>Sort by:</p> <ul style="list-style-type: none"> <li>- Frequency</li> <li>- Alphabetically</li> </ul> <p>Create taxonomy UI</p>	<p><b>Comparison</b> (Knowledge Nuggets)</p> <ul style="list-style-type: none"> <li>- kSig</li> <li>- Taxonomy of coll(s)</li> </ul>  <p>kNugget concept compare:</p> <ul style="list-style-type: none"> <li>- Top level only</li> <li>- Vocabulary</li> <li>- Partial hierarchy</li> <li>- Full hierarchy</li> </ul>	<p><b>Reduction</b></p>  <p>Infer:</p> <ul style="list-style-type: none"> <li>- Core/Outlier concepts</li> <li>- Emerging Trends</li> <li>- Subject Matter Expertise</li> <li>- email exchange quantification</li> </ul>	Computation
<ul style="list-style-type: none"> <li>- kSig creation (index into n-space)</li> <li>- kSig navigation</li> <li>- Synopsis generation</li> <li>- Summary generation</li> </ul>  <p>Operated by Los Alamos National Security, LLC for NNSA</p>	<p>Collection: (logical grouping of documents)</p> <ul style="list-style-type: none"> <li>- Creation</li> <li>- Use</li> <li>- Modification</li> <li>- Association</li> </ul>	<ul style="list-style-type: none"> <li>- Taxonomy Navigation</li> <li>- Conceptual linkage among docs. (indexing in n-space)</li> </ul>	<p>Use kNuggets to:</p> <ul style="list-style-type: none"> <li>- Find similar</li> <li>- ID duplicates</li> <li>- Display overlap</li> <li>- Display differences</li> </ul>	<p>Reduce large set of documents to a manageable set of knowledge inferred from the content and the collections</p>	eKM Concepts

UNCLASSIFIED

Slide 13



# Subject Matter Expertise Identification and eMail flow analysis



UNCLASSIFIED

Slide 14



## Conclusion

- Proof of principle capability today
- Systematic approach yields reproducibility
- Can work on small sets (10-20 documents) as well as large sets (100000+)
- Software requires no training set, but can be augmented by training
- Various components can be (and have been) applied to problems in many domains

## Technologies Integrated

- Homegrown code
- CIRILab theme parser
- Porter stemmer
- Lucene full-text indexer
- Apache Derby database
- MIT SIMILE timeline plotter
- Annie feature extractor
- *Future:* Kayvium theme extractor and ontology engine with FreeMind mind-mapping; graphic analysis; other visualization tools

