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

High Performance Computing
HPC-1 Scientific Software Engineering

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.

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

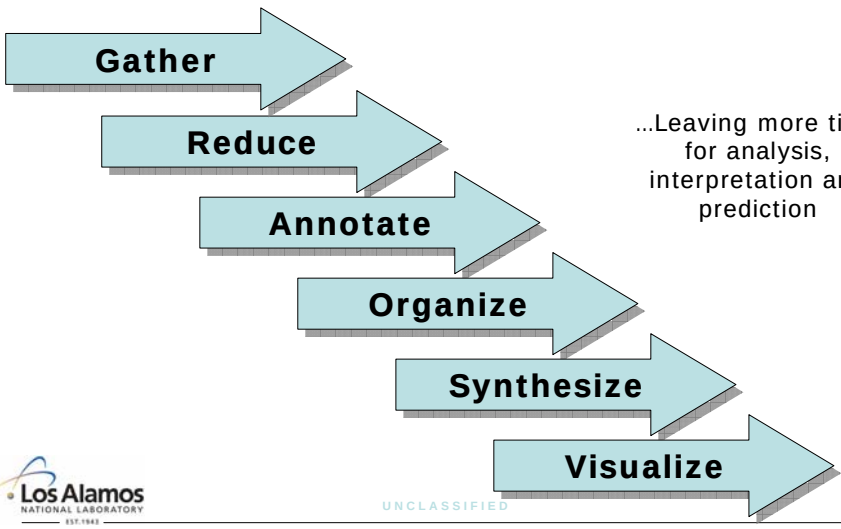
Challenges of the Electronic Information Age

- 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

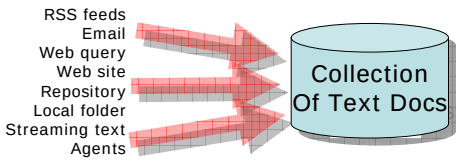
eKM to the Rescue



...Leaving more time for analysis, interpretation and prediction

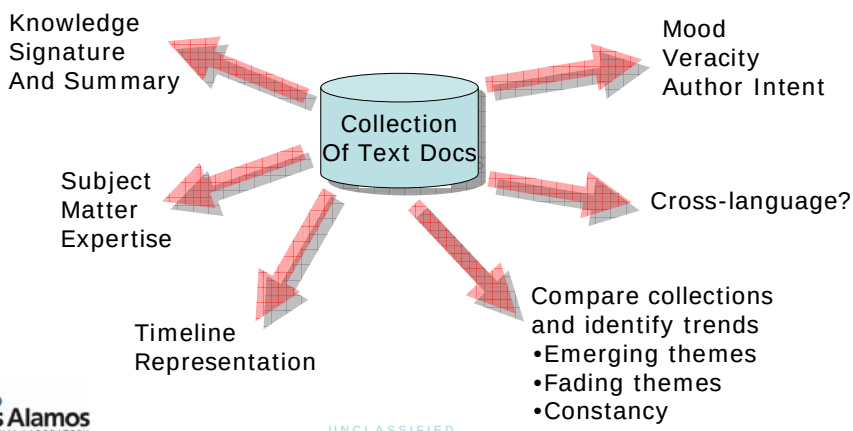
eKM to the Rescue (continued)

- 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

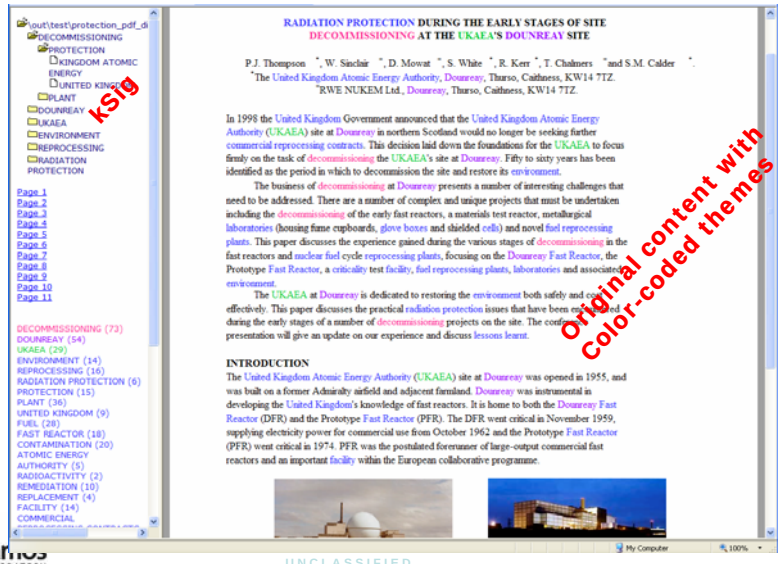
Choose Outputs Based on Need




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

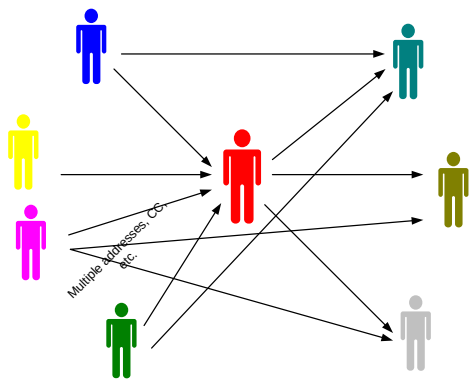


Timeline of President's Bush speeches related to Iraq

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

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
kSig Computation - Extract text - Infer Ksig (XML) - Paragraph annotation - Create kSig User Interface (UI)	**Collection Management** Record into repository: - kSig - meta-data Collection book-keeping	**Taxonomy** Sort by: - Frequency - Alphabetically Create taxonomy UI	**Comparison** (Knowledge Nuggets) - kSig - Taxonomy of coll(s) kNugget concept compare: - Top level only - Vocabulary - Partial hierarchy - Full hierarchy	**Reduction** Infer: - Core/Outlier concepts - Emerging Trends - Subject Matter Expertise - email exchange quantification	Computation
- kSig creation (index into n-space) *- kSig navigation* *- Synopsis generation* *- Summary generation*	*Collection: (logical grouping of documents)* *- Creation* *- Use* *- Modification* *- Association*	*- Taxonomy Navigation* *- Conceptual linkage among docs. (indexing in n-space)*	*Use kNuggets to:* *- Find similar* *- ID duplicates* *- Display overlap* *- Display differences*	*Reduce large set of documents to a manageable set of knowledge inferred from the content and the collections*	eKM Concepts

Subject Matter Expertise Identification and eMail flow analysis



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

High Performance Computing
HPC-1 Scientific Software Engineering

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

Cirilab

Lucene

The Apache DB Project
http://db.apache.org/

Simile

ANNIE

KAYVIUM
Refining Knowledge

FreeMind

Los Alamos NATIONAL LABORATORY EST. 1943
Operated by Los Alamos National Security, LLC for NNSA
UNCLASSIFIED
LA-UR-08-2421
NNSA