

The background of the slide features a close-up, black and white photograph of several pushpins. The pushpins are arranged in a way that their heads and stems create a sense of depth and perspective. Some pushpins are in sharp focus, while others are blurred in the background. The lighting creates soft shadows on the surface they are pinned to.

Emerging tools for email preservation

Preserving Email: Directions and Perspectives - July 2011

Tom Jackson
Information Science Department
www.drthomasjackson.com

Overview

- What are we trying to achieve?
- Email preservation considerations
- Current strategies
- Emerging strategies
- Useful links

What are we trying to achieve?

- Is it just Preservation?
- What about adding extra info?
 - Classification
 - Categorisation
 - Decision Capturing
 - Knowledge
- As well as preservation and archiving?

Considerations

- Wider Context:
 - Beyond converting to a reusable format
 - Volume of Email
 - Employee Time
 - Employee Know-How
 - Structuring Data

Considerations

- Preservation: Volume of Email
 - Space Management
 - Average employee spends 40 minutes a day managing their inbox to alleviate space restrictions
 - Responses range from 0 minutes to > 3 hours!

“...63% reported that space restrictions did not help them manage their inbox more effectively...”

Considerations

- Preservation: Volume of Email
 - 39% of email is information only (read only)
 - 29% copied in unnecessarily (cc,reply-to-all)
 - 17% irrelevant or untargeted (inc. SPAM)
 - 15% action required
- 46% of email an employee receives does not require storing
- What about stress and email management?

Considerations

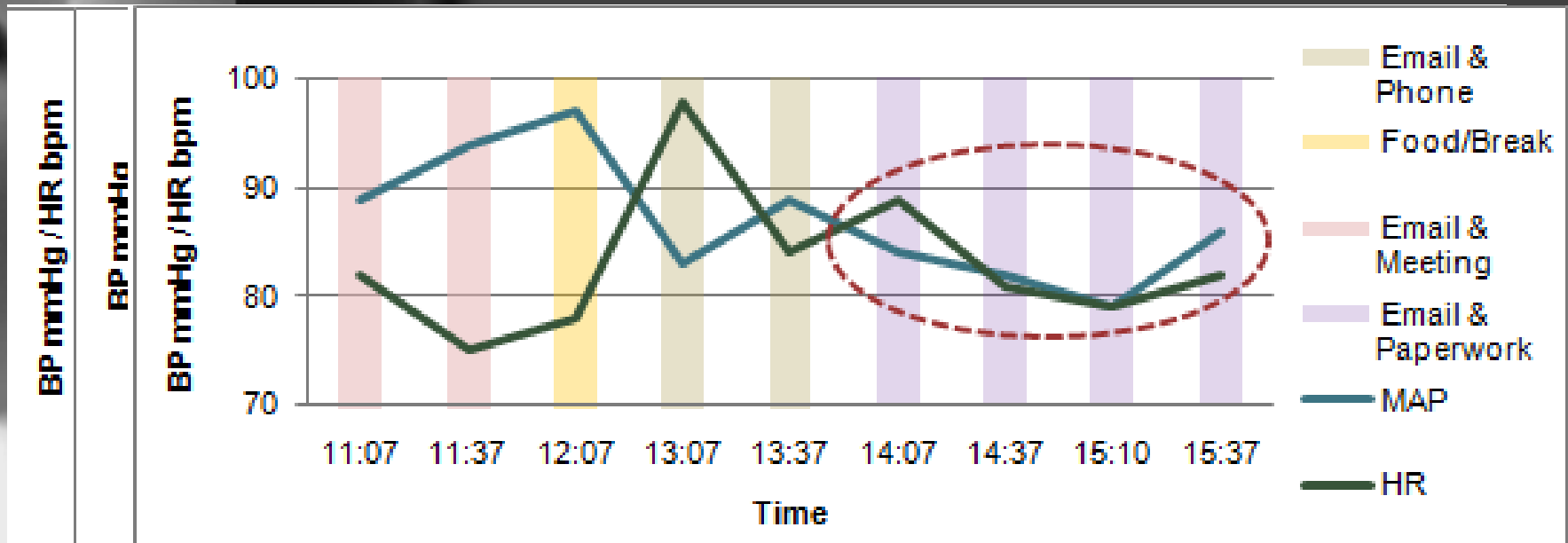


Data Collection

- Thirty invited volunteers
- Two monitoring periods:
(1) Email Use (2) Email Free Time
- Distribute questionnaires
- ABP machine attach/remove
- Saliva samples & fridge stored (within 4 hours) & freezer stored (12 hours)
- Diary used to log events

Stress during Email and Other Activities

- Increased BP & HR during email and phone use
- Increased BP during email and face-to-face meeting
- Decreased BP & HR during email and paperwork.



- Not filing email causes stress – Chris!

Considerations

- Preservation: Volume of Email
- Push and pull storage
 - Deciding if to store on sending
 - Time saving if sent to large number of recipients
 - Recipient knows message has already been saved

Considerations

- Preservation: Employee Time

- Overwhelmed by email

- “...53% of employees receive more email than they can handle...”

- Time for work?

- “...76% of employees feel they don't have sufficient time to do their work....”

- Email Overload (Stephen)?

- “...87% suffer or have suffered email overload...”

Considerations

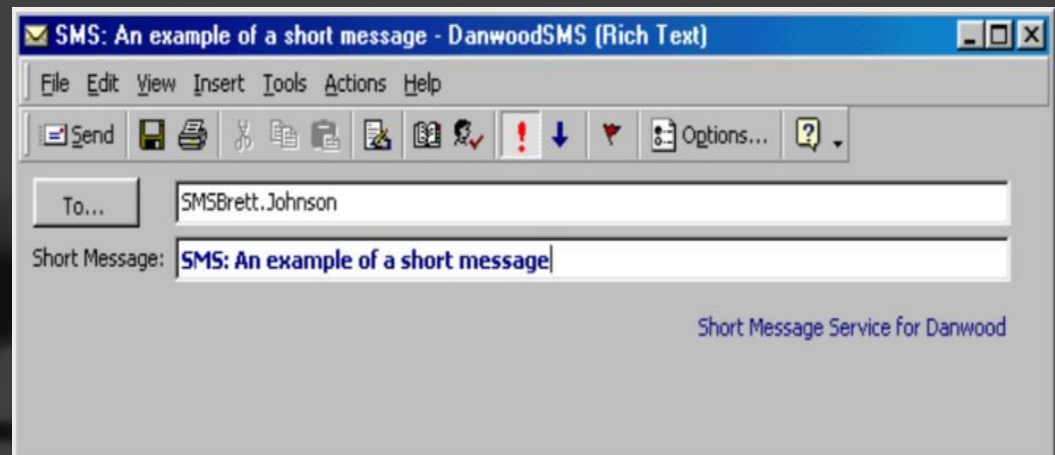
- Preservation: Employee Know-How
 - Do they know what to store?
 - Where it should be stored?
- IM Capabilities of Employees
 - Research at Leicestershire County Council into IM capabilities
 - Employee IM assessment
 - Identify training areas

Considerations

- Preservation: Structuring Data
 - Well known issue for preserving email
 - Number of systems available to convert formats into reusable format e.g. CERP email parser
 - Proactive approach: structure new emails (Outlook 2007 add-in, Outlook 2010 built-in).

Considerations

- Preservation: Structuring Data
 - Need to structure to aid in processing
 - XML (Digital Preservation Testbed in the Netherlands)
 - Predefined forms Not moved much further from the 90s (Lotus Notes - forms)





Current Strategies

Current Strategies

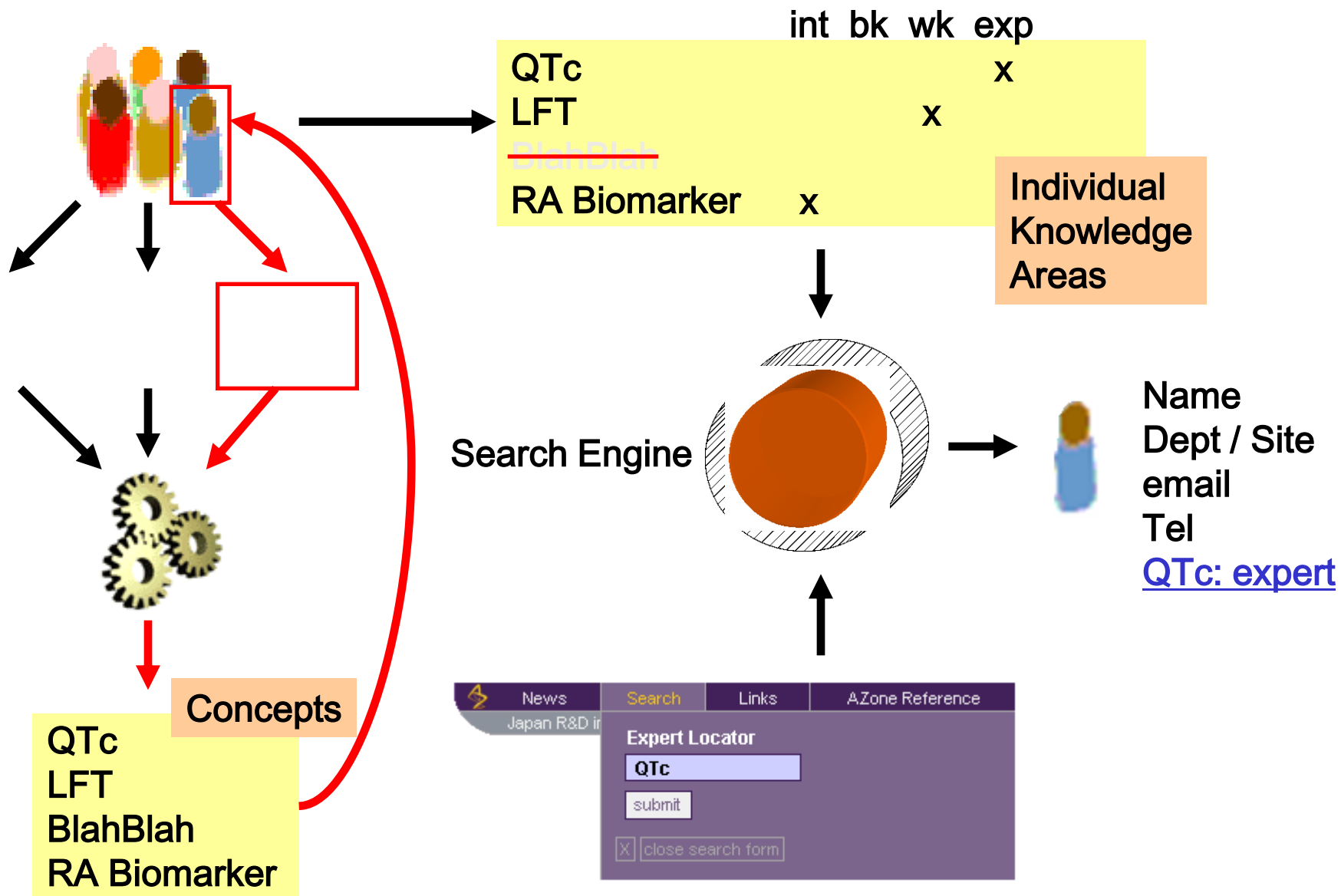
- Not involving Employees
 - Preserving existing Email in vaults
 - No real impact on employees
- Categorising & Classify Email
 - High user involvement
 - C&C when sending and receiving email
 - The WAG – employees not having time or understanding how to do it



Emerging Strategies

Emerging Strategies

- Solutions have to be integrated into business processes to be successful
- Employees too busy....
- Example - Email Knowledge Extraction



A working example of an email sent through the keyphrase extraction system

>>> Obtain email text

Mary & Mike, I spoke to John today who is working on trying to construct a simple version of the email trainer. Mike, it might be worth you mentioning to John the web site that re-writes text so it has a better structure. Thomas

>>> Tokenise the text

<mary>, <&>, <mike>, <,>, <i>, <spoke>, <to>, <john>, <today>, <who>, <is>, <working>, <on>, <trying>, <to>, <construct>, <a>, <simple>, <version>, <of>, <the>, <email>, <trainer>, <.> and so on....

>>> Apply POS Tagger

<mary/NN>, <&/cc-tl>, <mike/NN>, <./,>, <i/nn>, <spoke/vbd>, <to/to>, <john/vb>, <today/nr>, <who/wps>, <is/bez>, <working/vbg>, <on/in>, <trying/vbg>, <to/to>, <construct/vb>, <a/at>, <simple/jj>, <version/nn>, <of/in>, <the/at>, <email/NN>, <trainer/NN>, <./.> and so on....

>>> Pick Keyphrases from within each candidate phrase

S: <mary/NN> <&/cc-tl> <mike/NN> <./,> <i/nn> <spoke/vbd> <to/to> <john/vb> <today/nr> <who/wps> <is/bez> <working/vbg> <on/in> <trying/vbg> <to/to> <construct/vb> <a/at> <simple/jj> <version/nn> <of/in> (Key phrase: <the/at> <email/NN> <trainer/NN>) <./.> and so on....

>>> Apply linguistic filters (initially when wordNet was not used)

(<email/NN> <trainer/NN>), (<site/nn>)

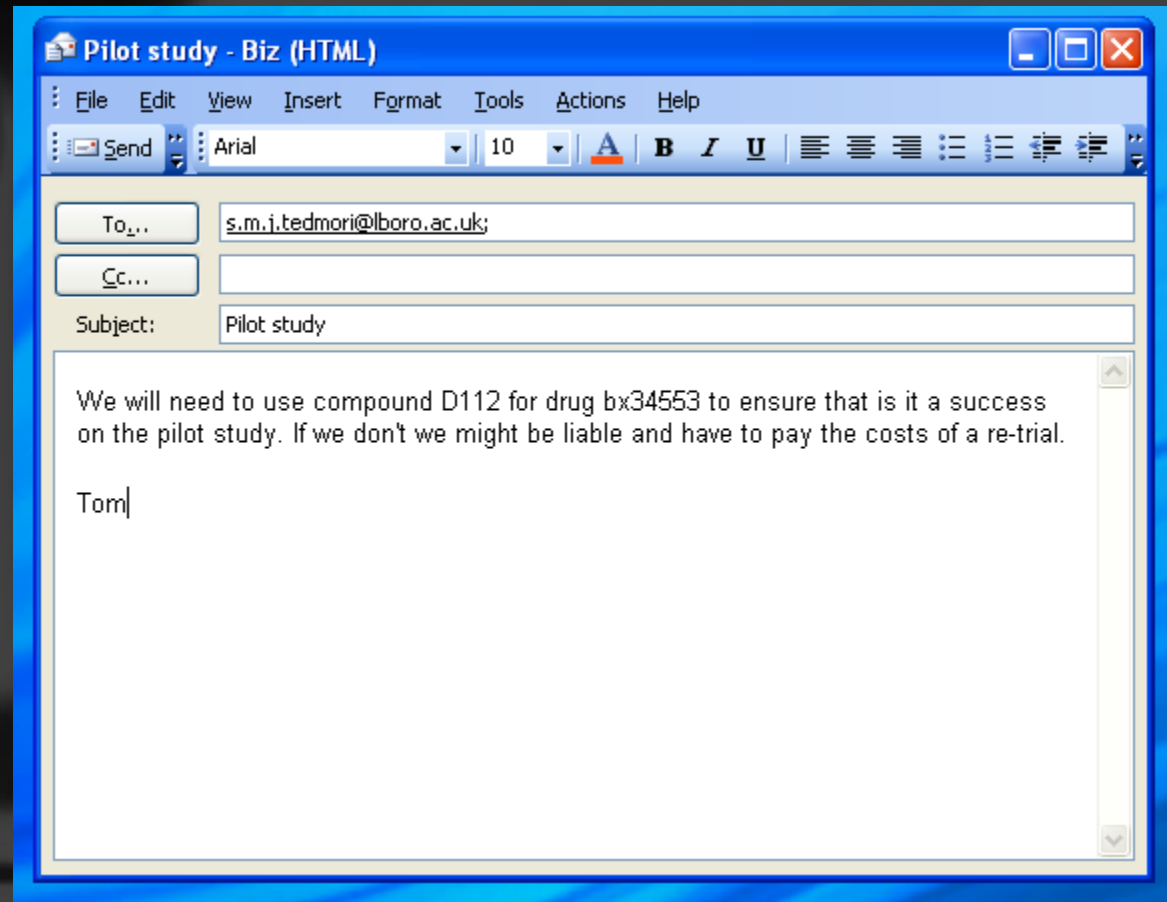
>>> Apply linguistic filters (with wordNet used)

(<email/NN> <trainer/NN>)

For the complete set of tags used in the Brown corpus please refer to <http://www.comp.leeds.ac.uk/amalgam/tagsets/brown.html>

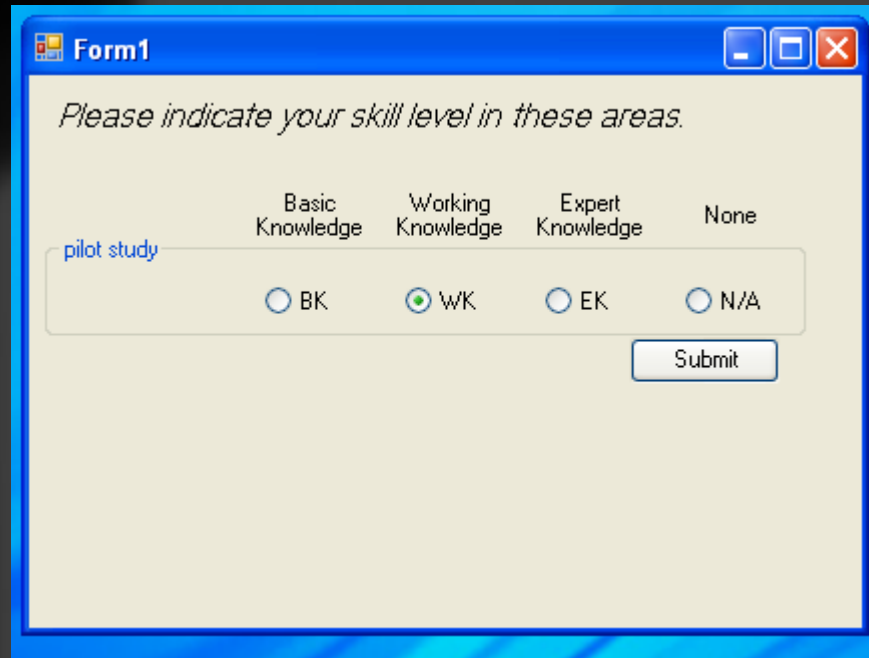
Emerging Strategies

- Standard interface



Emerging Strategies

- User asked to rank immediately



Form1

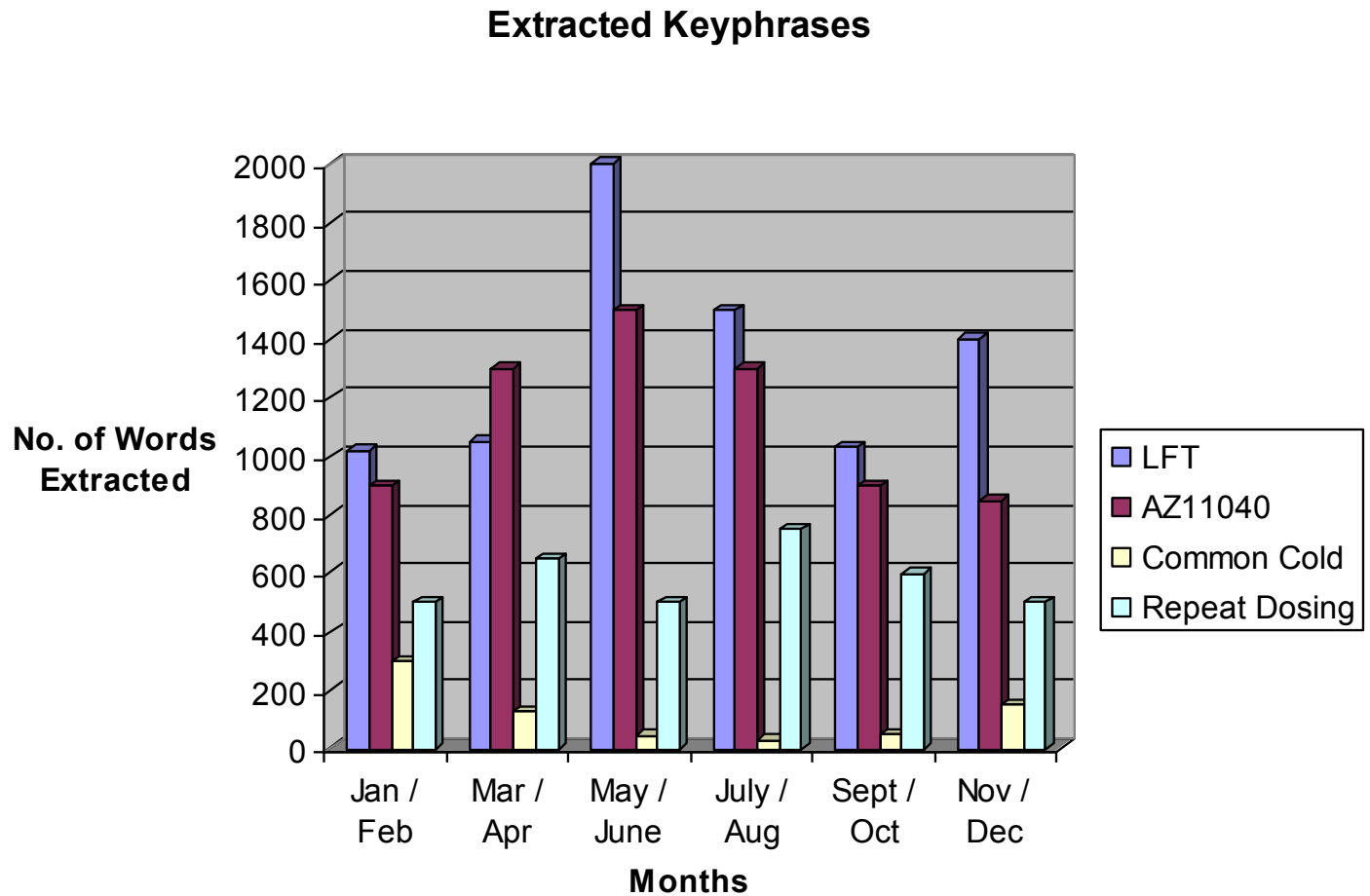
Please indicate your skill level in these areas.

	Basic Knowledge	Working Knowledge	Expert Knowledge	None
pilot study	<input type="radio"/> BK	<input checked="" type="radio"/> WK	<input type="radio"/> EK	<input type="radio"/> N/A

Submit

- System learns from user
- User will not see keyphrase again
- Best f-measure in the world

Organisational Knowledge



Emerging Strategies

- Wish list requirements:
 - Convert to XML on sending
 - Categorise Email
 - Classify Email
 - To avoid distributing the end user too much
 - Remove duplication
 - Capture decisions
 - New Research at TNA – Email Classification & Categorisation System

Emerging Strategies

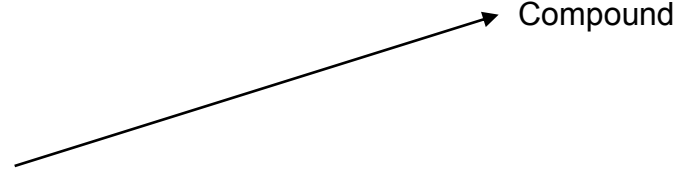
- Real-time vs Stored Email
- Real-time snap shot view?
- Stored email
 - News articles, other emails, organisational structure, retention policy, decision making
- Other Issues
 - Inconsistencies
 - Taxonomy, Ontology - OntoFarm

The Right Context.....(inconsistencies)

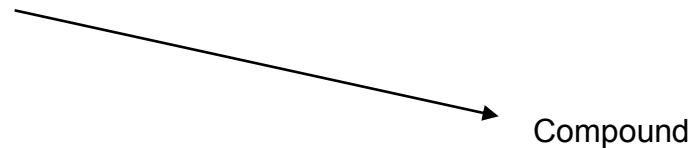
Suzy



Hi Tony,
Use the **compound** approach
Suzy



Compound



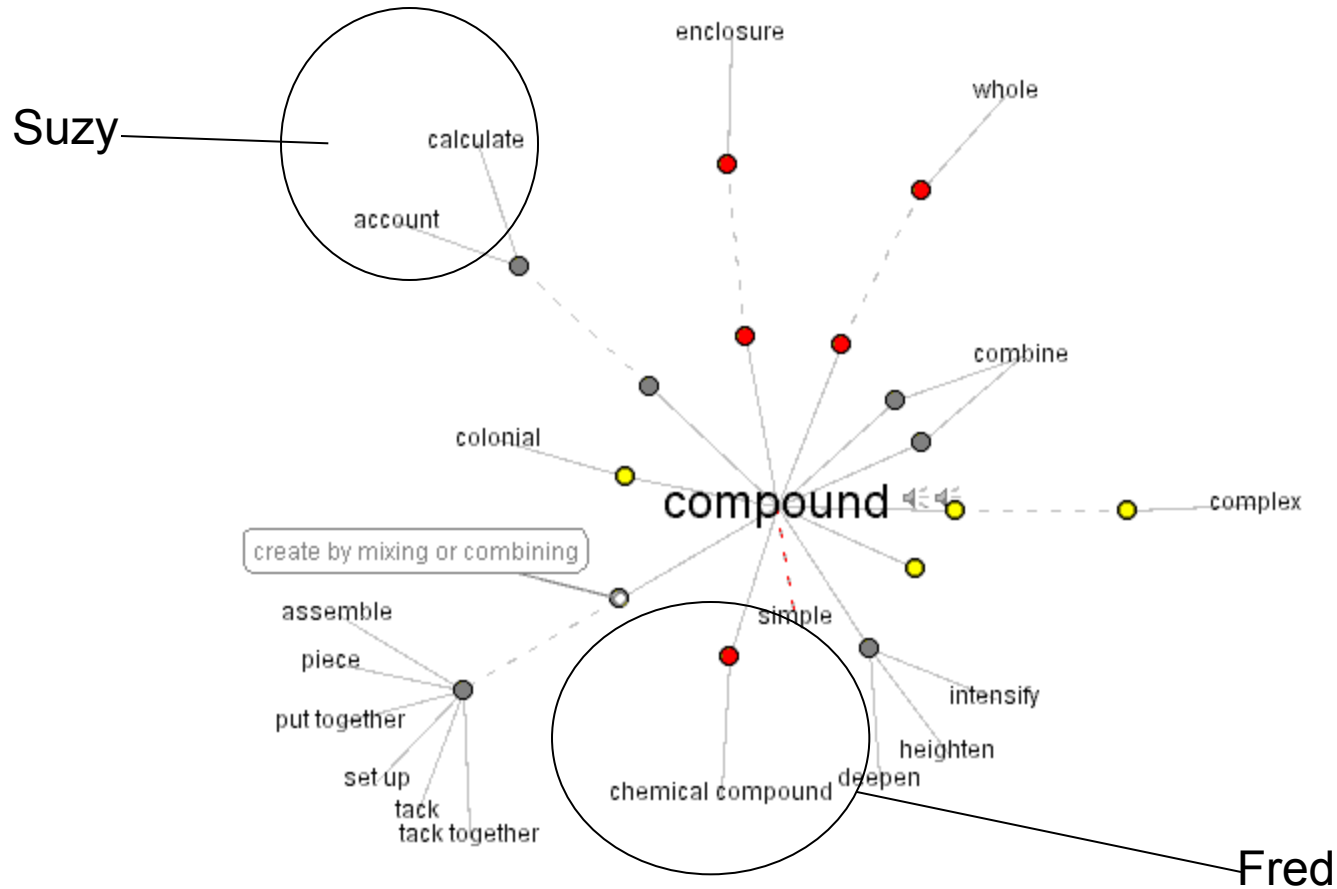
Compound

Fred



Mary,
The outcome depends on the **compound** you use
Fred

Technical Issues



Emerging Strategies

- Semi-automated Ontology Creator
 - OntoFarm

Concept Label and Description

test:Ruby_On_Rails

Ruby on Rails is an open source web application framework for the Ruby programming language. It is often referred to as "Rails" or "RoR".

[Browse Concepts](#) [Destroy](#) [Export Namespace as Text](#) [Export Namespace as XML](#) [Export Namespace as OWL](#)

Harvester

Ruby on rails spider? Google Wikipedia SDN

location: <http://en.wikipedia.org/wiki/Ruby%20on%20rails>

Help us provide free content to the world by [donating today!](#) [Log in / create account](#)

[article](#) [discussion](#) [edit this page](#) [history](#)

Ruby on Rails

From Wikipedia, the free encyclopedia
(Redirected from [Ruby on rails](#))

Ruby on Rails is an open source web application framework for the [Ruby programming language](#). It is often referred to as "Rails" or "RoR". It is intended to be used with an [Agile development methodology](#), which is often utilized by web developers for its suitability for short, client-driven projects.

Contents [hide]

- 1 History
- 2 Technical overview
- 3 Framework structure
- 4 Philosophy and Design
- 5 See also

ajax application
dauid development
edit framework
frameworks free
hansson heinemeier
links page php
programming rails
retrieved ruby
server toolkit web

Order results by: [frequency](#), [occurance](#), [name](#)

Developed by [Lex](#) [Relation](#)
[Rep](#)

Rails Core Team [Lex](#) [Relation](#)
[Rep](#)

Latest release [Lex](#) [Relation](#)
[Rep](#)

Written in [Lex](#) [Relation](#)
[Rep](#)

Properties

Subject of Relations

Ruby_On_Rails relatedTo test: [Create](#)

Subject	Predicate	Object	Delete
---------	-----------	--------	--------

Object of Relations

test: relatedTo Ruby_On_Rails [Create](#)

Subject	Predicate	Object	Delete
test:Ruby_Programming_Language	coeprop.relatedTo	Ruby_On_Rails	Delete

Lexical Representations

[Create](#)

Lexical Representation	Delete
Ruby on rails	Delete

Lexical Representations

Harvest View

Conclusion

- How will we measure success?
- What is the business case?
 - Legal requirements
 - Value added – extracting knowledge
 - Value added – ontology
 - Time required to store and retrieve
- How far do we go – big brother?
 - Interpretation – risk?

Useful Links

- **Archivematica (open source software) -**
http://archivematica.org/wiki/index.php?title=Main_Page
- **Email Preservation Project**
 - Blogs and discussion
 - Recommendations
 - Resources
- **Preservation of Electronic Mail Collaboration Initiative -**
http://www.records.ncdcr.gov/emailpreservation/technical_resources.htm
- **Digital Preservation Testbed in the Netherlands: "XML for Digital Preservation - XML Implementation Options for E-mails"**
- **Collaborative Electronic Records Project (CERP)**
- **Preserving Access to Digital Information (PADI) Email**
- **CERP Email Parser (open source) -** <http://siarchives.si.edu/cerp/index.htm>