A Visual Analytic Framework for Exploring Relationships in Textual Contents of Digital Forensics Evidence

T.J. Jankun-Kelly, D. Wilson, A. Stamps, J. Franck, J. Carver, J. E. Swan II
Mississippi State University & University of Alabama

VizSec 2009

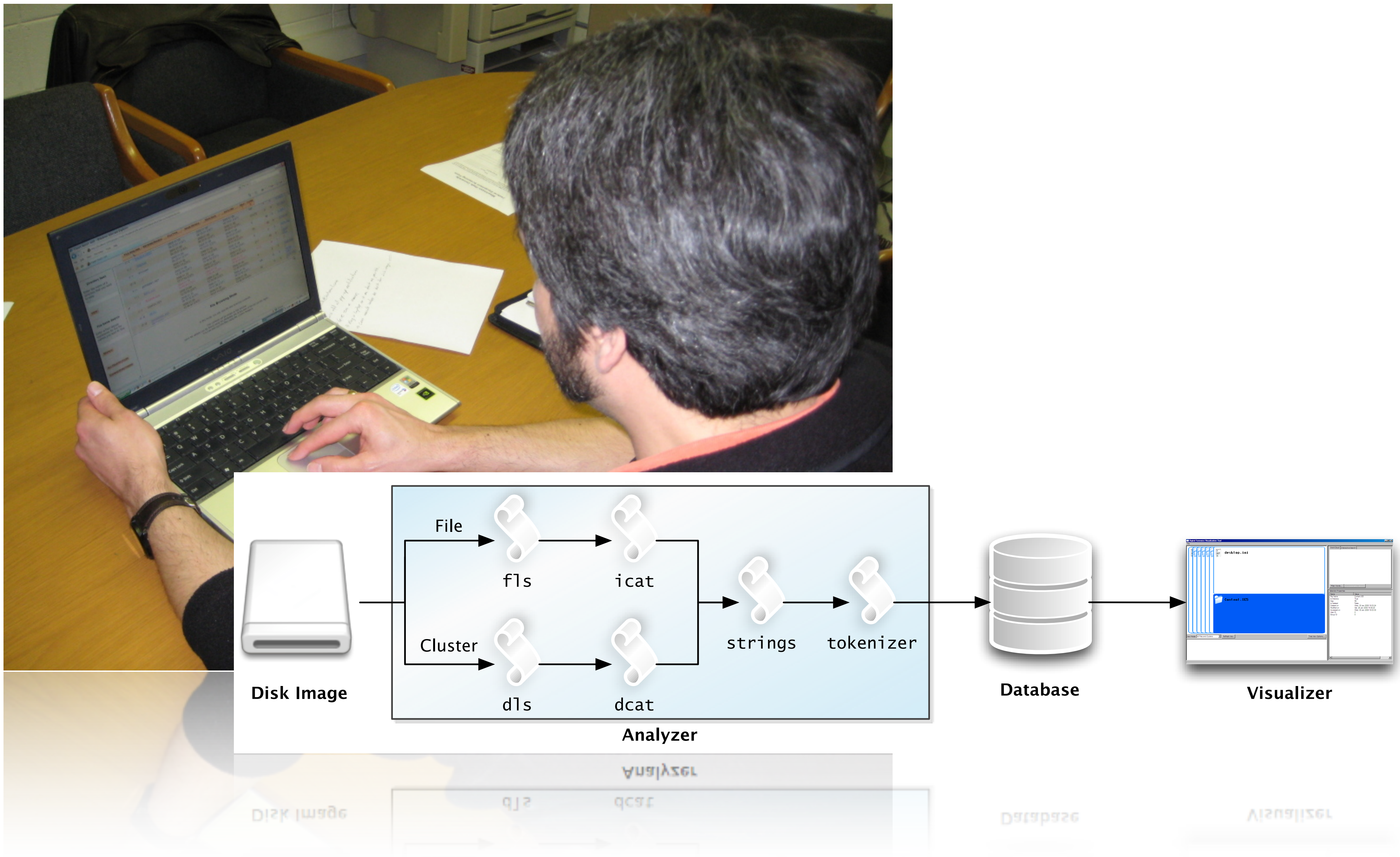
VisWeek 09
VIS • INFOVIS • VAST

MISSISSIPPI STATE UNIVERSITY™

JAMES WORTH BAGLEY COLLEGE OF ENGINEERING MISSISSIPPI STATE UNIVERSITY

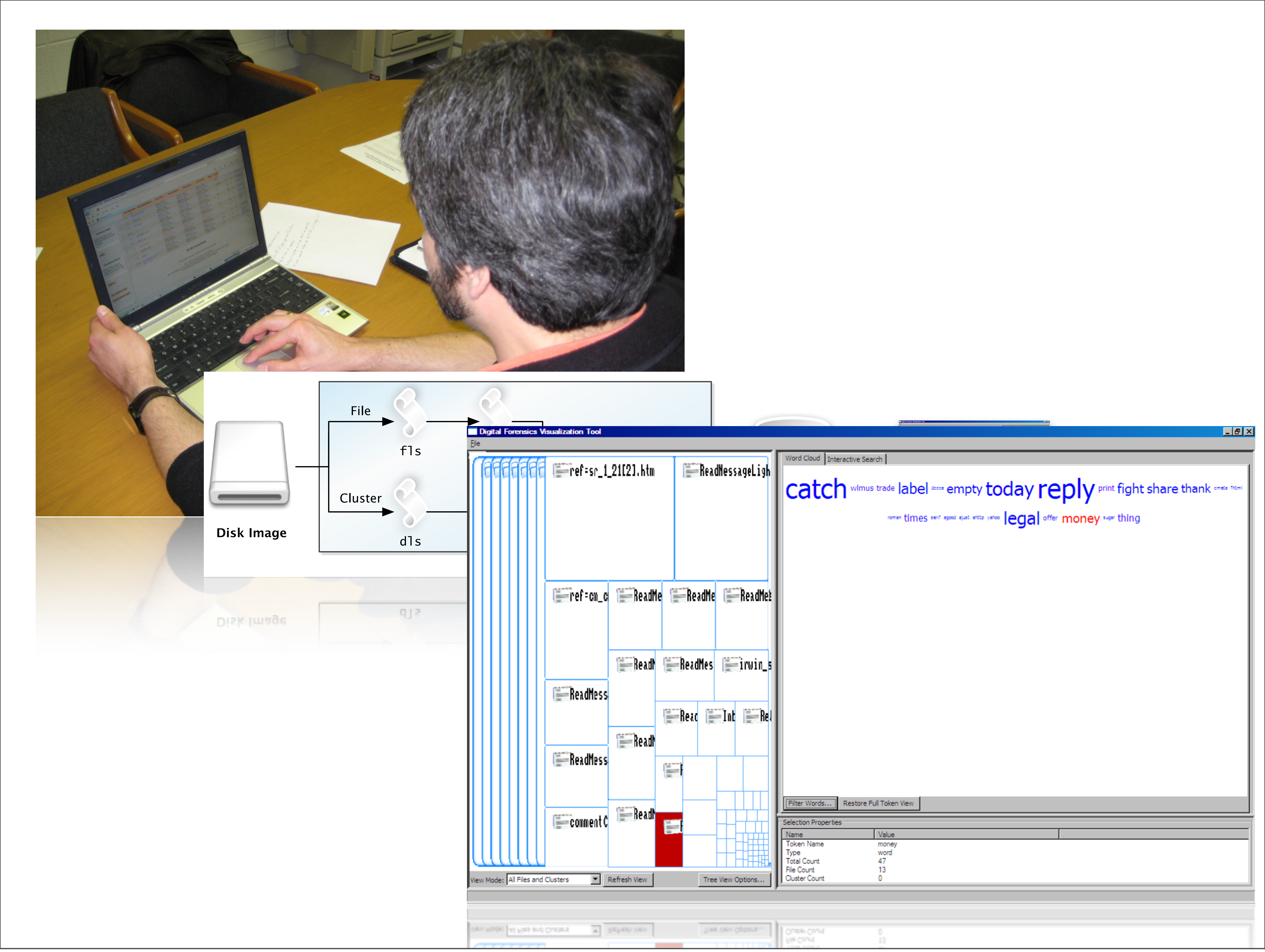
This talk is a bit of a departure from the others here at VizSec: It is about computer forensics (specifically hard disk forensics) as opposed to computer security. Hard drive forensics is an area ripe with opportunities. Current practices are laborious—direct linear search of a very large space is the most common approach. Visualization has the potential for significant impact. HD forensics is a large area; we are focused on text forensics, specifically in email forensics. [Image mollazi http://www.sxc.hu/photo/1153697]

The focus of this talk is our visual analysis framework for performing text relation analysis: I.e., finding terms, related terms that appear in the same documents, and how these words are distributed through (related) files. Such textual relationships are important in many cases. We chose this task and our design based upon a **contextual analysis** that we conducted with forensics officers; I'll discuss these shortly. The study informed our **analysis framework** which processes a hard disk to find terms and their relationships on disk which is then visualized with a **search-sensitive file hierarchy and tag cloud**. Each of these will be discussed in turn.



Sunday, October 11, 2009

The focus of this talk is our visual analysis framework for performing text relation analysis: I.e., finding terms, related terms that appear in the same documents, and how these words are distributed through (related) files. Such textual relationships are important in many cases. We chose this task and our design based upon a **contextual analysis** that we conducted with forensics officers; I'll discuss these shortly. The study informed our **analysis framework** which processes a hard disk to find terms and their relationships on disk which is then visualized with a **search-sensitive file hierarchy and tag cloud**. Each of these will be discussed in turn.



Selection Properties

Name	Value
Token Name	money
Type	word
Total Count	47
File Count	13
Cluster Count	0

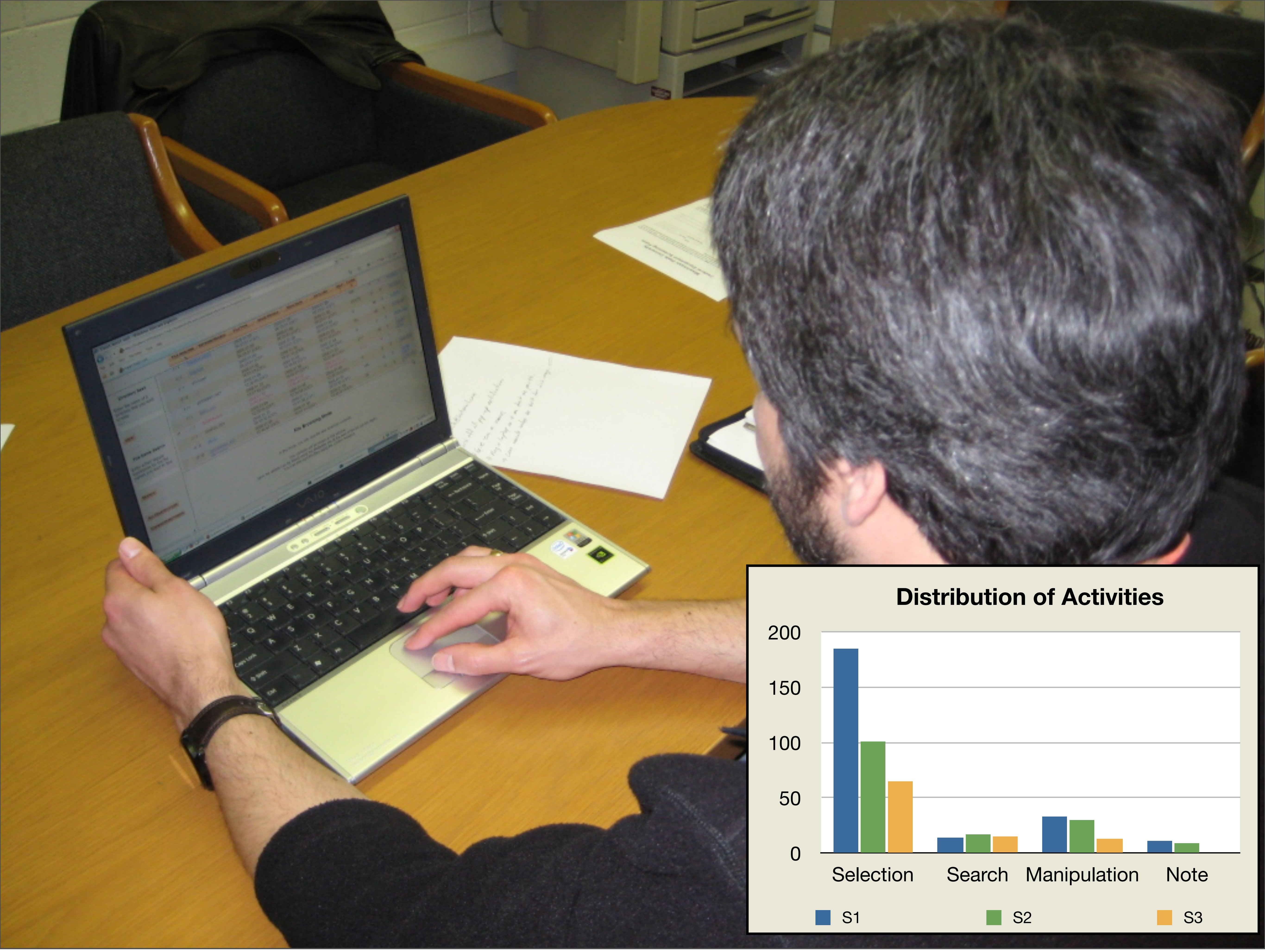
The focus of this talk is our visual analysis framework for performing text relation analysis: I.e., finding terms, related terms that appear in the same documents, and how these words are distributed through (related) files. Such textual relationships are important in many cases. We chose this task and our design based upon a **contextual analysis** that we conducted with forensics officers; I'll discuss these shortly. The study informed our **analysis framework** which processes a hard disk to find terms and their relationships on disk which is then visualized with a **search-sensitive file hierarchy and tag cloud**. Each of these will be discussed in turn.

Pre-Design Study

Sunday, October 11, 2009

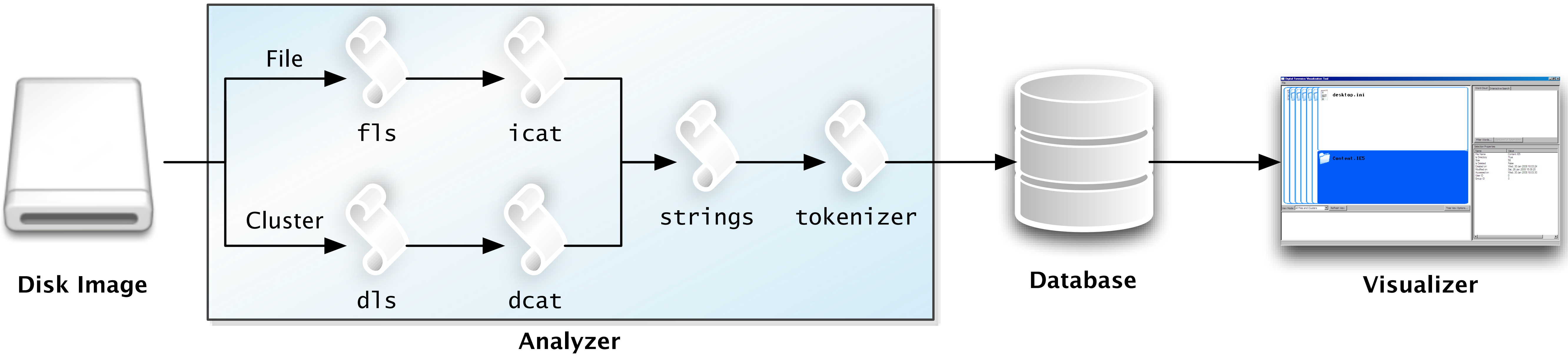
Before we initiated our design we studied forensics officers in order to determine how they currently do analysis. We specifically were interested in finding out where there were inefficiencies with the goal of finding areas that could be benefited by visualization. This was the study we reported upon at last years VizSec.

Our testbed is webmail based forensics. We created two datasets with fraudulent behavior via several false webmail accounts and emails back and forth. We mixed these in with more legitimate sources by signing up the main accounts to various mailing lists. In addition, the main account performed various web-browsing behavior, some related to the fraud, others not. Two disc images of these were then used as the base data for observation. As discussed last year, we recorded the interactions with video and key/mouse logging for 3 law officers. [Summarize]



Our testbed is webmail based forensics. We created two datasets with fraudulent behavior via several false webmail accounts and emails back and forth. We mixed these in with more legitimate sources by signing up the main accounts to various mailing lists. In addition, the main account performed various web-browsing behavior, some related to the fraud, others not. Two disc images of these were then used as the base data for observation. As discussed last year, we recorded the interactions with video and key/mouse logging for 3 law officers. [Summarize]

The Framework



Disk Image

File

Cluster

fls

icat

dls

dcat

strings

tokenizer

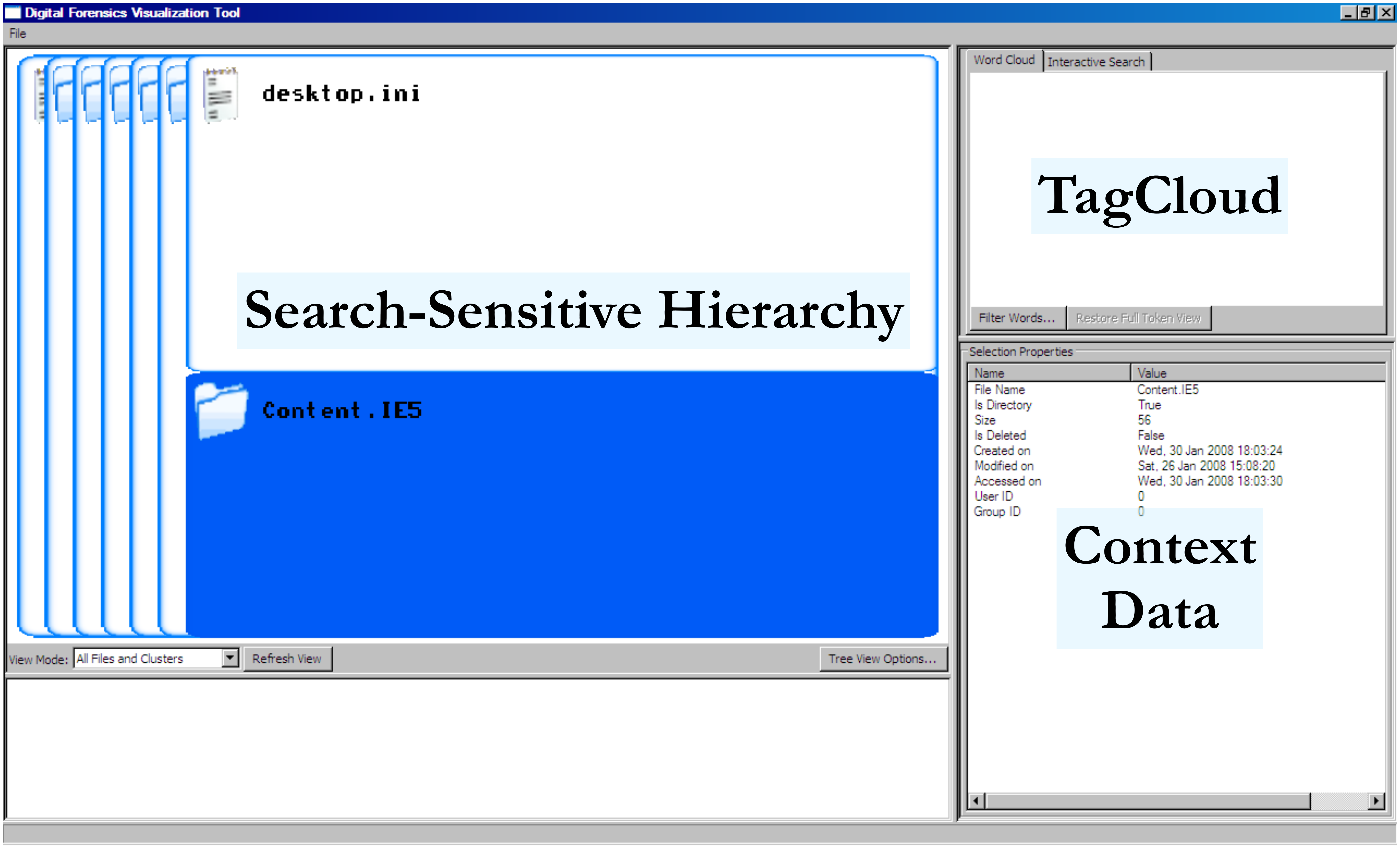
Analyzer

Database

Visualizer

Our analytic framework consists of three major components: The **Analyzer** which processes the disk image, the **Text Relation Database** which stores the extracted relations, and the **Visualizer** which I'll discuss later. To begin the analysis, we use the Sleuth Kit to walk through the directory structure and extract any textual information in the given file; a similar process is done for unallocated (deleted) clusters. These are then split into tokens and categorized into common entities: URLs, email addresses, currency, HTML/XML tags, and so forth. Each token is then entered into the database which records each word and each file/cluster, which files each word appeared in, what words a file contains, and before/after relationships between words; frequency of occurrence within a file are also tabulated for later use. This processing is slow (~20min for 4.5MB, ~20hr for 4.5GB), but is similar to normal processing and is only done once: All our visualizations use the textual relationship database directly.

Visualizations



Search-Sensitive Hierarchy

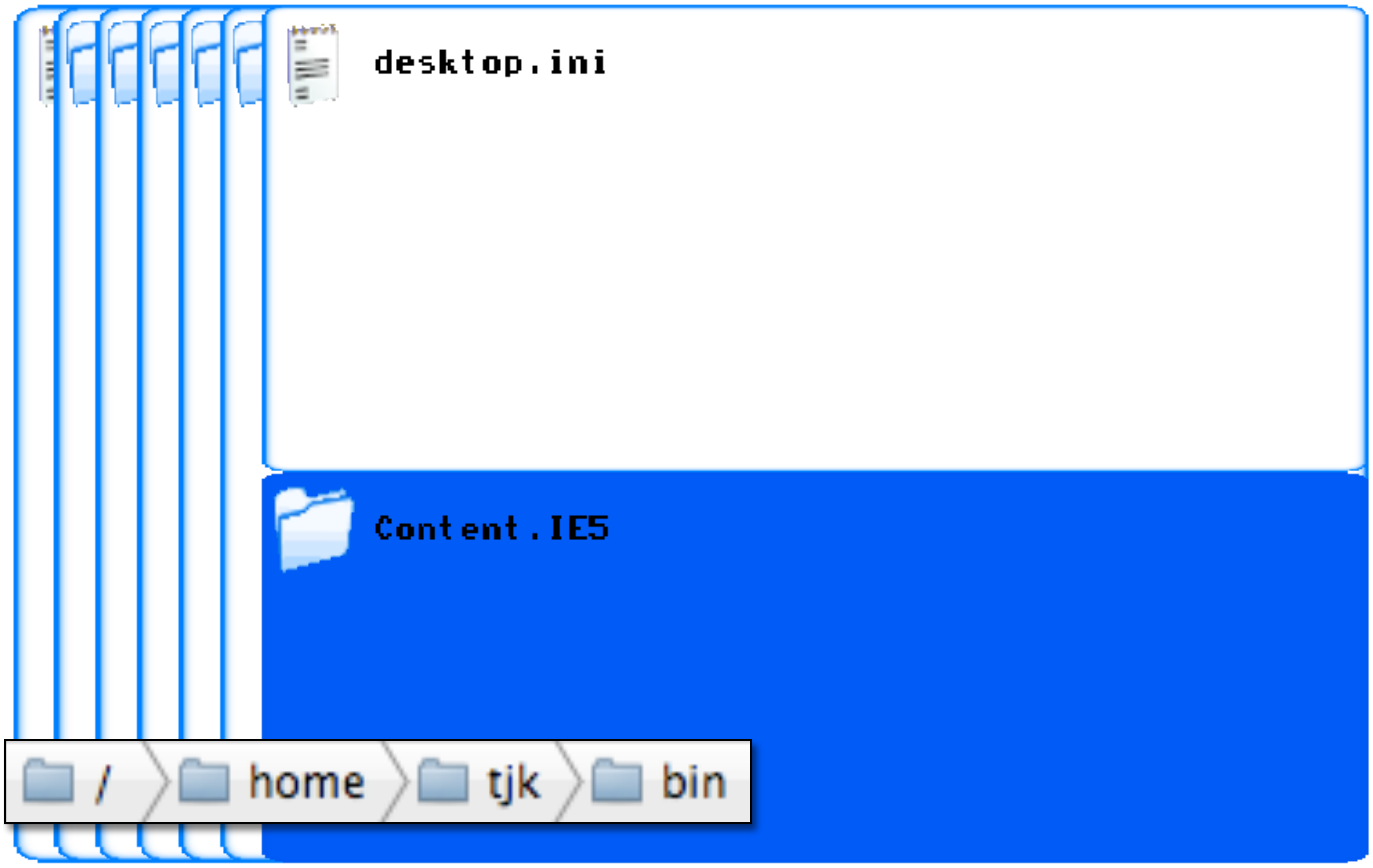
TagCloud

Context Data

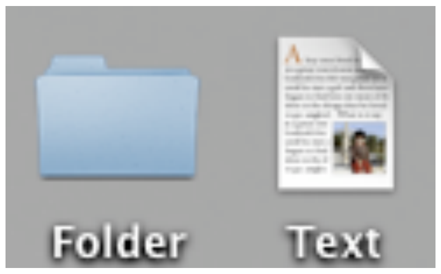
Name	Value
File Name	Content.IE5
Is Directory	True
Size	56
Is Deleted	False
Created on	Wed, 30 Jan 2008 18:03:24
Modified on	Sat, 26 Jan 2008 15:08:20
Accessed on	Wed, 30 Jan 2008 18:03:30
User ID	0
Group ID	0

Sunday, October 11, 2009

The visualizer allows for browsing of the disk hierarchy, searching for terms, and analyzing their relationships. It is divided into three regions: The Search-Sensitive Hierarchy, a interactive TagCloud, and a window for showing file meta-data and for conjunctive search.



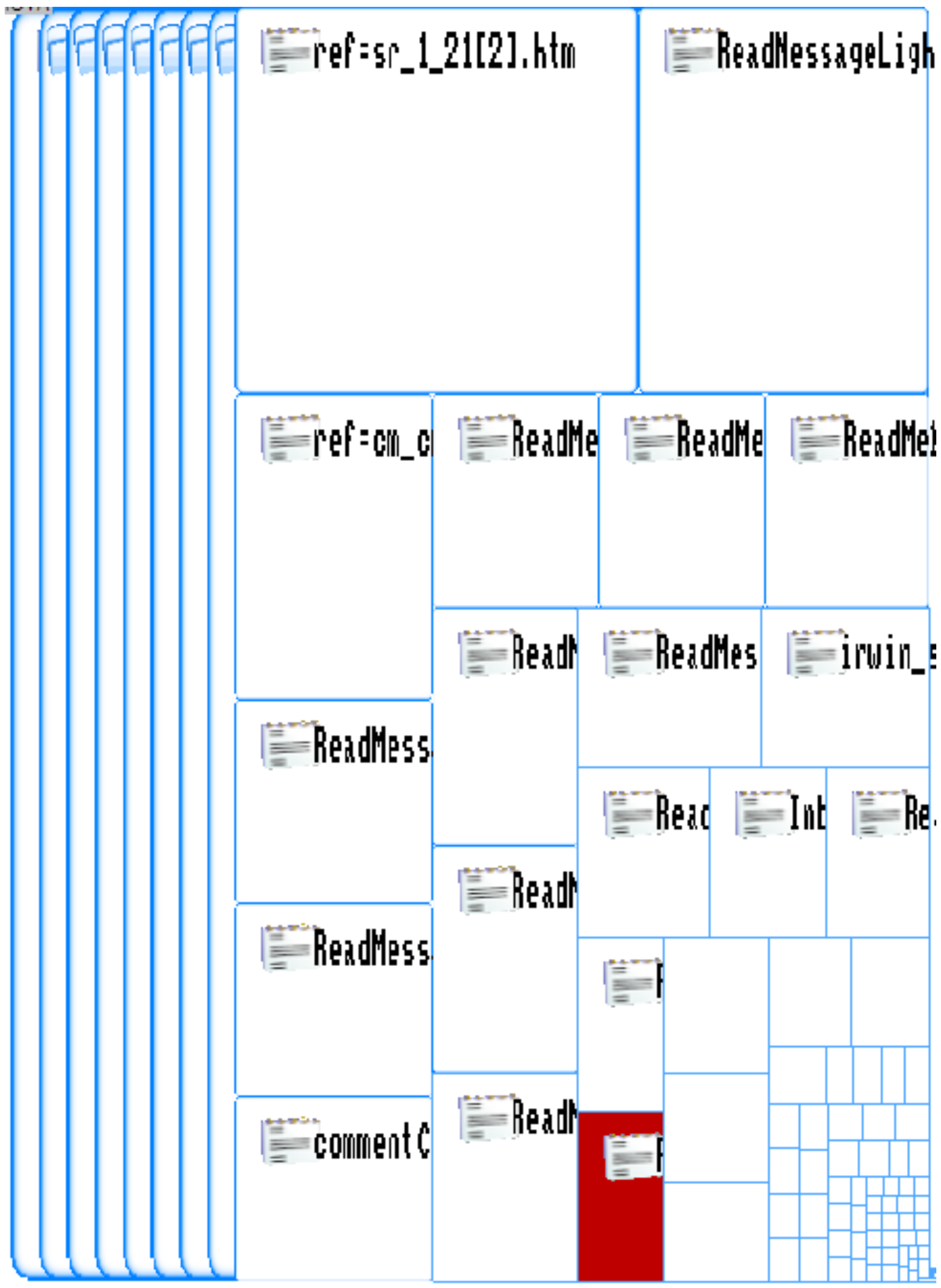
Spring-Loaded



Familiar Icons

Search-Sensitive Hierarchy

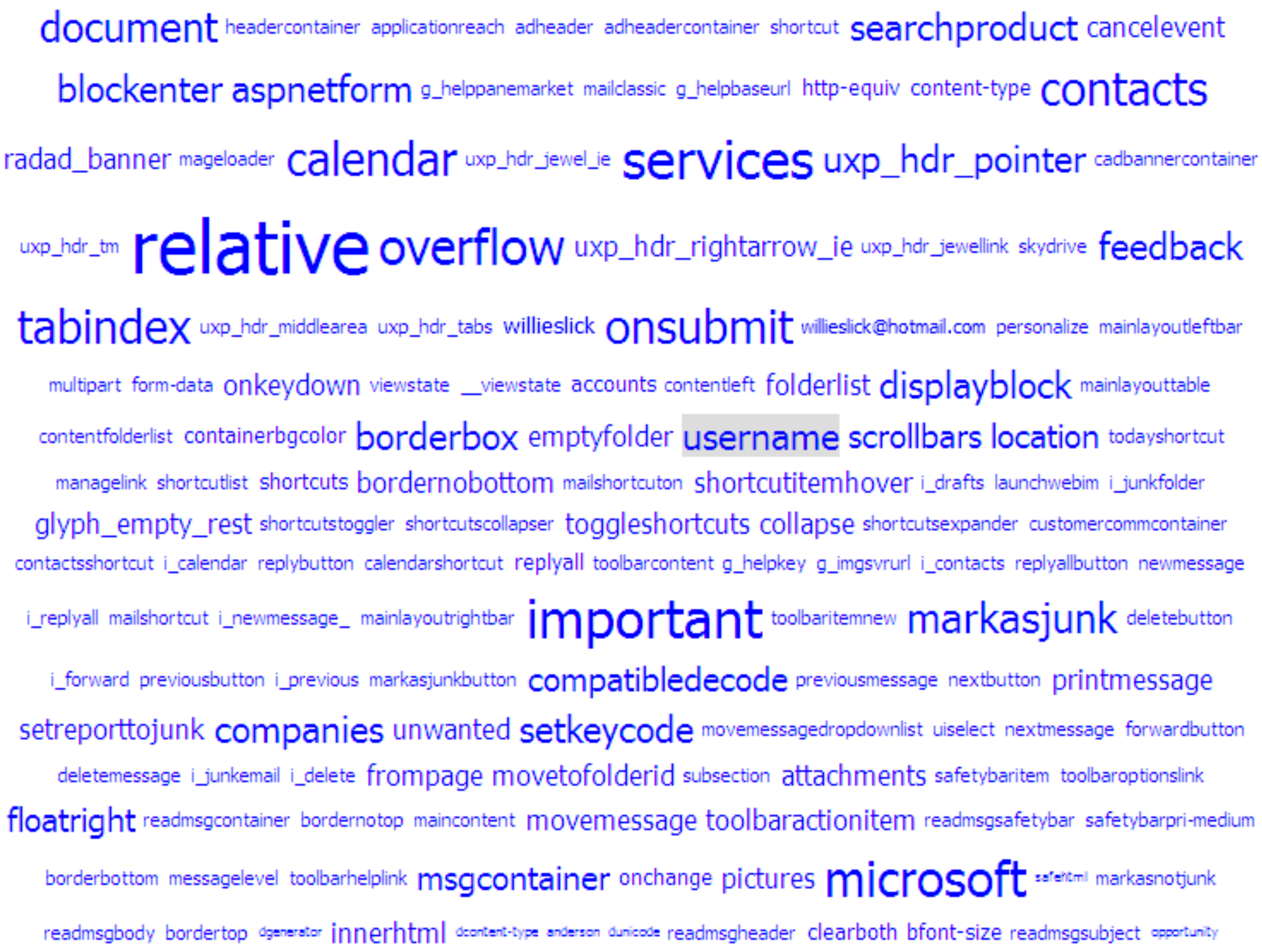
The hierarchy is a modified squarified treemap that only displays files with textual data; the size is determined by the # of text elements. It is modified to make it more familiar to officers that are not typical visualization system users. It uses a “spring-loaded” display of the hieararchy (showing the root->child from left->right) similar to other OSes (also has advantage of saving space); in addition, it labels files and directories with familiar “Explorer”- or “Finder”-like folder and file icons. The reason we call the treemap “search-sensitive” is that any words selected in the tagcloud or other views cause files with related words to light up.



Dynamic Updates

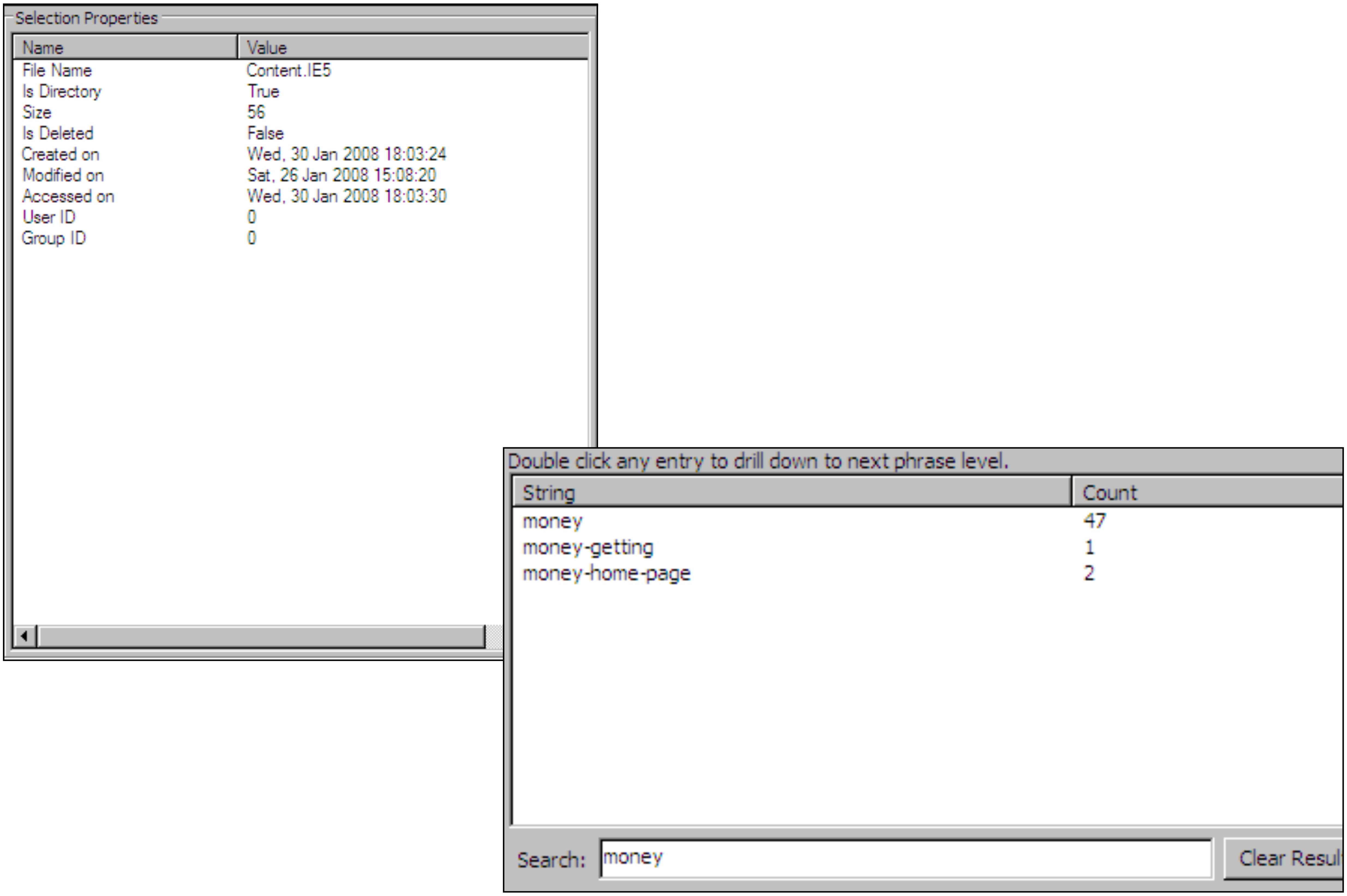
Search-Sensitive Hierarchy

The hierarchy is a modified squarified treemap that only displays files with textual data; the size is determined by the # of text elements. It is modified to make it more familiar to officers that are not typical visualization system users. It uses a “spring-loaded” display of the hieararchy (showing the root->child from left->right) similar to other OSes (also has advantage of saving space); in addition, it labels files and directories with familiar “Explorer”- or “Finder”-like folder and file icons. The reason we call the treemap “search-sensitive” is that any words selected in the tagcloud or other views cause files with related words to light up.


document headercontainer applicationreach adheader adheadercontainer shortcut searchproduct cancelevent
blockenter aspnetform g_helppanemarket mailclassic g_helpbaseurl http-equiv content-type contacts
radad_banner mageloader calendar uxp_hdr_jewel_ie services uxp_hdr_pointer cadbannercontainer
uxp_hdr_tm relative overflow uxp_hdr_rightarrow_ie uxp_hdr_jewellink skydrive feedback
tabindex uxp_hdr_middlearea uxp_hdr_tabs willieslick onsubmit willieslick@hotmail.com personalize mainlayoutleftbar
multipart form-data onkeydown viewstate __viewstate accounts contentleft folderlist displayblock mainlayouttable
contentfolderlist containerbgcolor borderbox emptyfolder username scrollbars location todayshortcut
managelink shortcutlist shortcuts bordernobottom mailshortcuton shortcutitemhover i_drafts launchwebim i_junkfolder
glyph_empty_rest shortcutstoggler shortcutscollapser toggleshortcuts collapse shortcutsexpander customercommcontainer
contactsshortcut i_calendar replybutton calendarshortcut replyall toolbarcontent g_helpkey g_imgsvrurl i_contacts replyallbutton newmessage
i_replyall mailshortcut i_newmessage_ mainlayoutrightbar important toolbaritemnew markasjunk deletebutton
i_forward previousbutton i_previous markasjunkbutton compatibledecode previousmessage nextbutton printmessage
setreporttojunk companies unwanted setkeycode movemessagedropdownlist uiselect nextmessage forwardbutton
deletemessage i_junkemail i_delete frompage movetofolderid subsection attachments safetybaritem toolbaroptionslink
floatright readmsgcontainer bordernotop maincontent movemessage toolbaractionitem readmsgsafetybar safetybarpri-medium
borderbottom messagelevel toolbarhelplink msgcontainer onchange pictures microsoft safehtml markasnotjunk
readmsgbody bordertop dgenerator innerhtml dcontent-type anderson dunicode readmsgheader clearboth bfont-size readmsgsubject opportunity

Tag Cloud

The tag cloud is a standard one: It highlights word frequencies within a document or a directory (a sum over documents). Words can be selected as needed by clicking (highlighted in grey); in addition, search terms from the search view are highlighted in red. There are several interactive filters that can be applied which assist in analyzing the tag cloud/text data.



Selection Properties

Name	Value
File Name	Content.IE5
Is Directory	True
Size	56
Is Deleted	False
Created on	Wed, 30 Jan 2008 18:03:24
Modified on	Sat, 26 Jan 2008 15:08:20
Accessed on	Wed, 30 Jan 2008 18:03:30
User ID	0
Group ID	0

Double click any entry to drill down to next phrase level.

String	Count
money	47
money-getting	1
money-home-page	2

Search: money

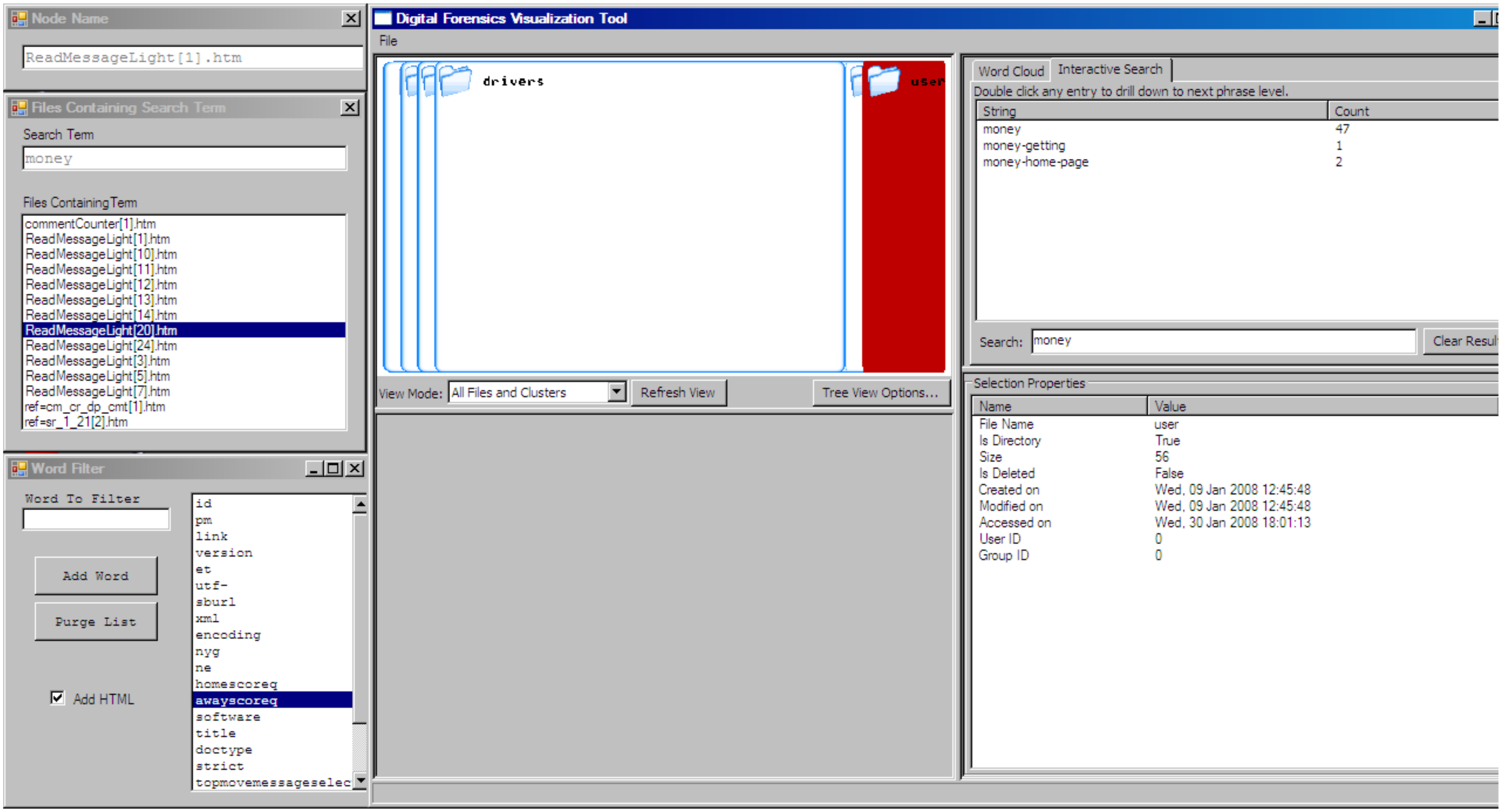
Context Metadata & Search

Context area shows file metadata (standard). The search box allows direct searching of terms (either in a file or directory, depending on selection), gives exact size, and shows where following or leading terms are used. This is a very naive right now (the building up of these); improvements are future work.

Usage Study

Sunday, October 11, 2009

Here we walk through a scenario of searching for investment fraud online. We don’t know the details other than the user of the machine (“William Slick”) is suspected of fraud.



Node Name: ReadMessageLight[1].htm

Search Term: money

Files Containing Term:

- commentCounter[1].htm
- ReadMessageLight[1].htm
- ReadMessageLight[10].htm
- ReadMessageLight[11].htm
- ReadMessageLight[12].htm
- ReadMessageLight[13].htm
- ReadMessageLight[14].htm
- ReadMessageLight[20].htm
- ReadMessageLight[24].htm
- ReadMessageLight[3].htm
- ReadMessageLight[5].htm
- ReadMessageLight[7].htm
- ref=cm_cr_dp_cmt[1].htm
- ref=sr_1_21[2].htm

Word To Filter:

☑ Add HTML

```
id
pm
link
version
et
utf-
sburl
xml
encoding
nyg
ne
homescoreq
awayscoreq
software
title
doctype
strict
topmovemessageselec
```

View Mode: All Files and Clusters

Double click any entry to drill down to next phrase level.

String	Count
money	47
money-getting	1
money-home-page	2

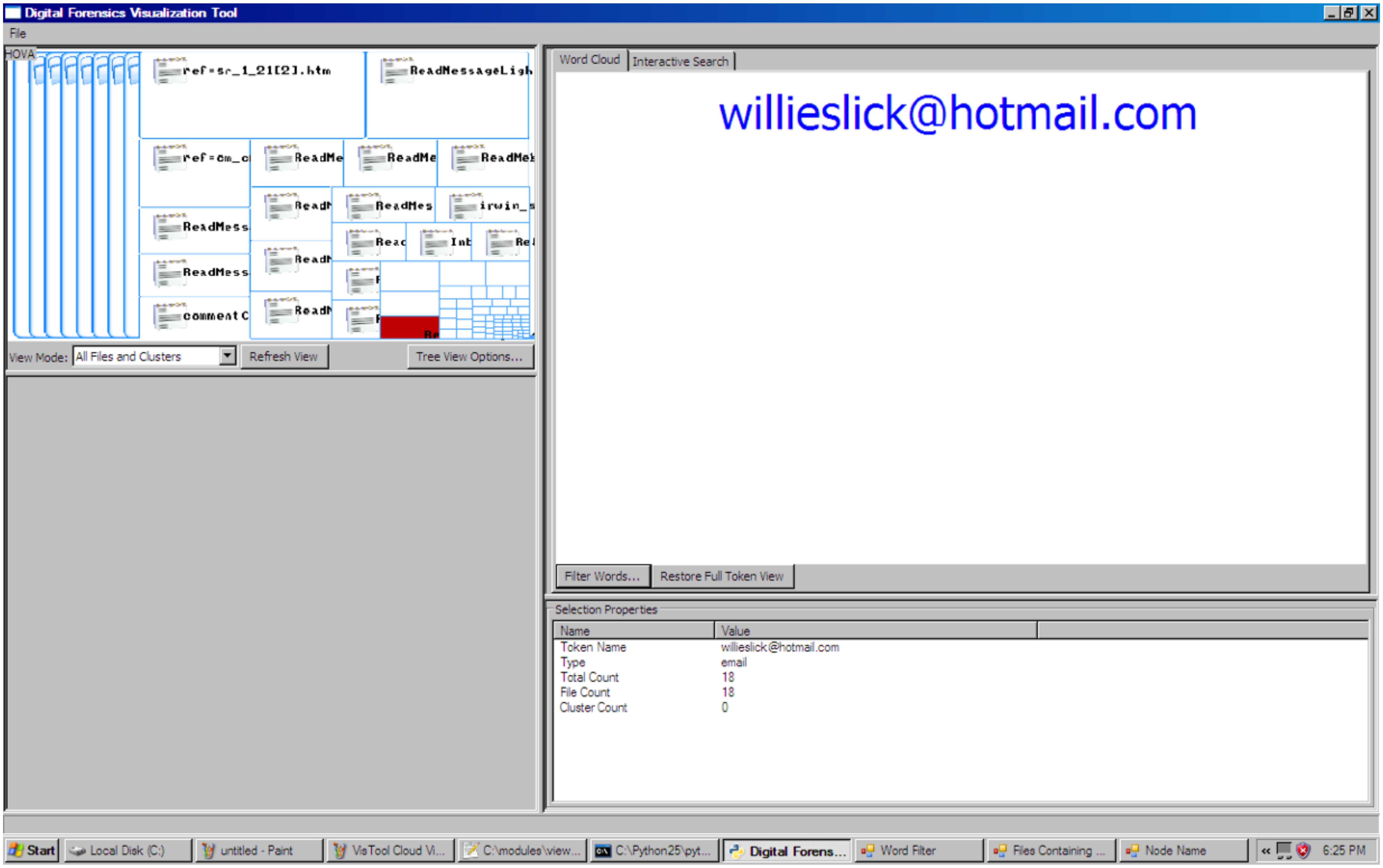
Search: money

Selection Properties

Name	Value
File Name	user
Is Directory	True
Size	56
Is Deleted	False
Created on	Wed, 09 Jan 2008 12:45:48
Modified on	Wed, 09 Jan 2008 12:45:48
Accessed on	Wed, 30 Jan 2008 18:01:13
User ID	0
Group ID	0

Phase 1: Term Search

Not knowing the particulars, we can search for “money” or “investments” across the disk. There are 47 occurrences of this term in one directory, the web-mail directory. In addition, the “money getting” co-occurrence looks promising, so lets explore the file with that.



Name	Value
Token Name	willieslick@hotmail.com
Type	email
Total Count	18
File Count	18
Cluster Count	0

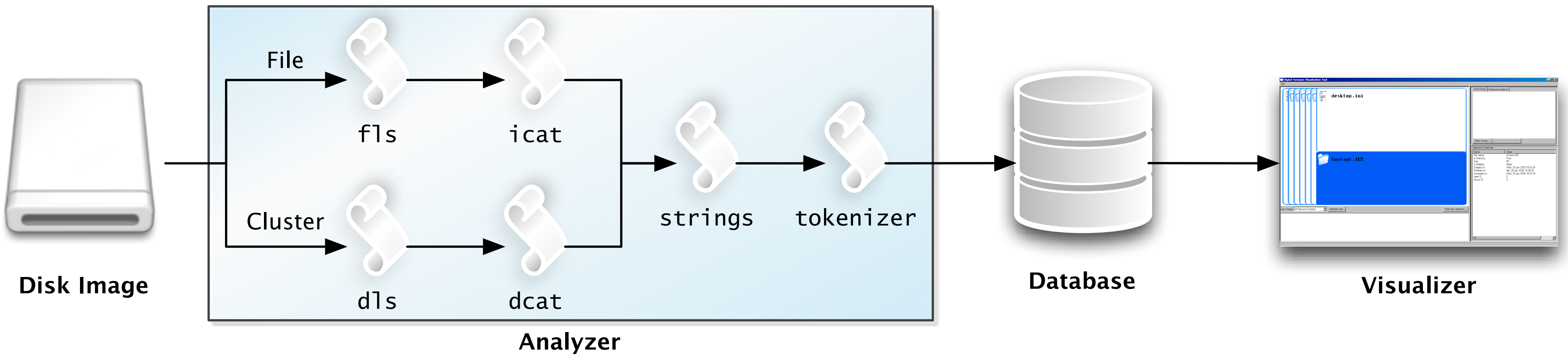
Phase 3: Investigation

Sunday, October 11, 2009

So, lets see if we can tie our suspect to the fraud. We can now turn off everything but email addresses, and one pops up: That of William Slick. Good job! :)

Summary & Future

Study



Disk Image

File

fls

icat

Cluster

dls

dcat

strings

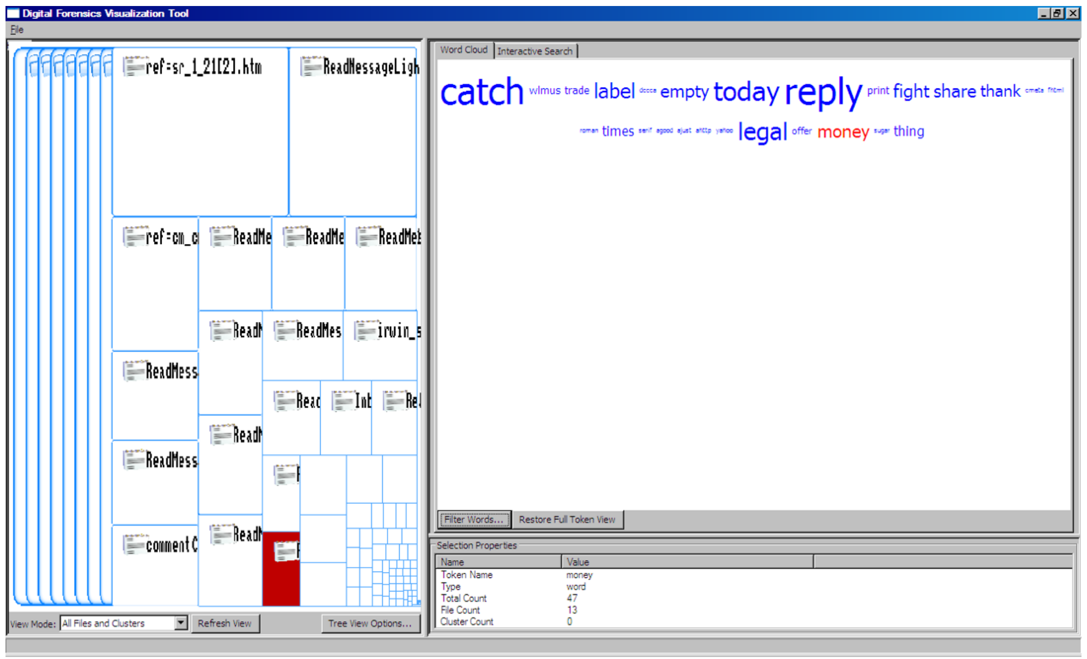
tokenizer

Analyzer

Database

Visualizer

Framework

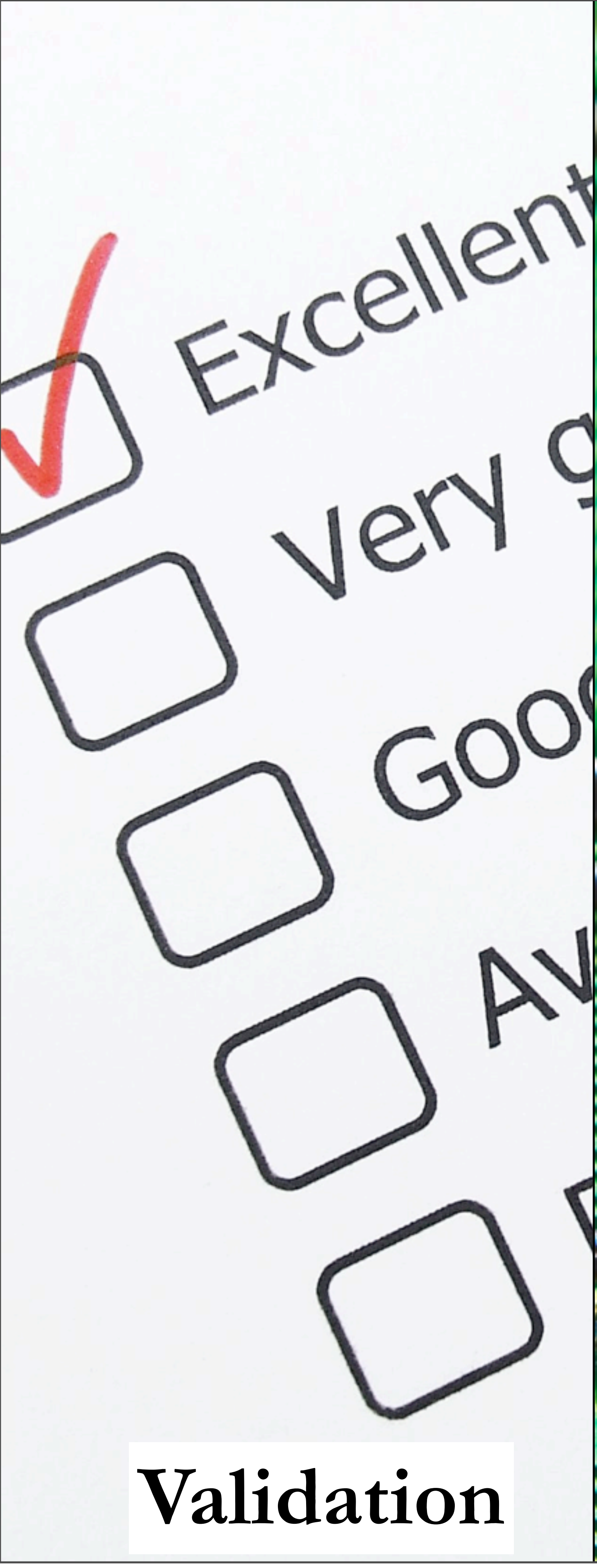


Digital Forensics Visualization Tool

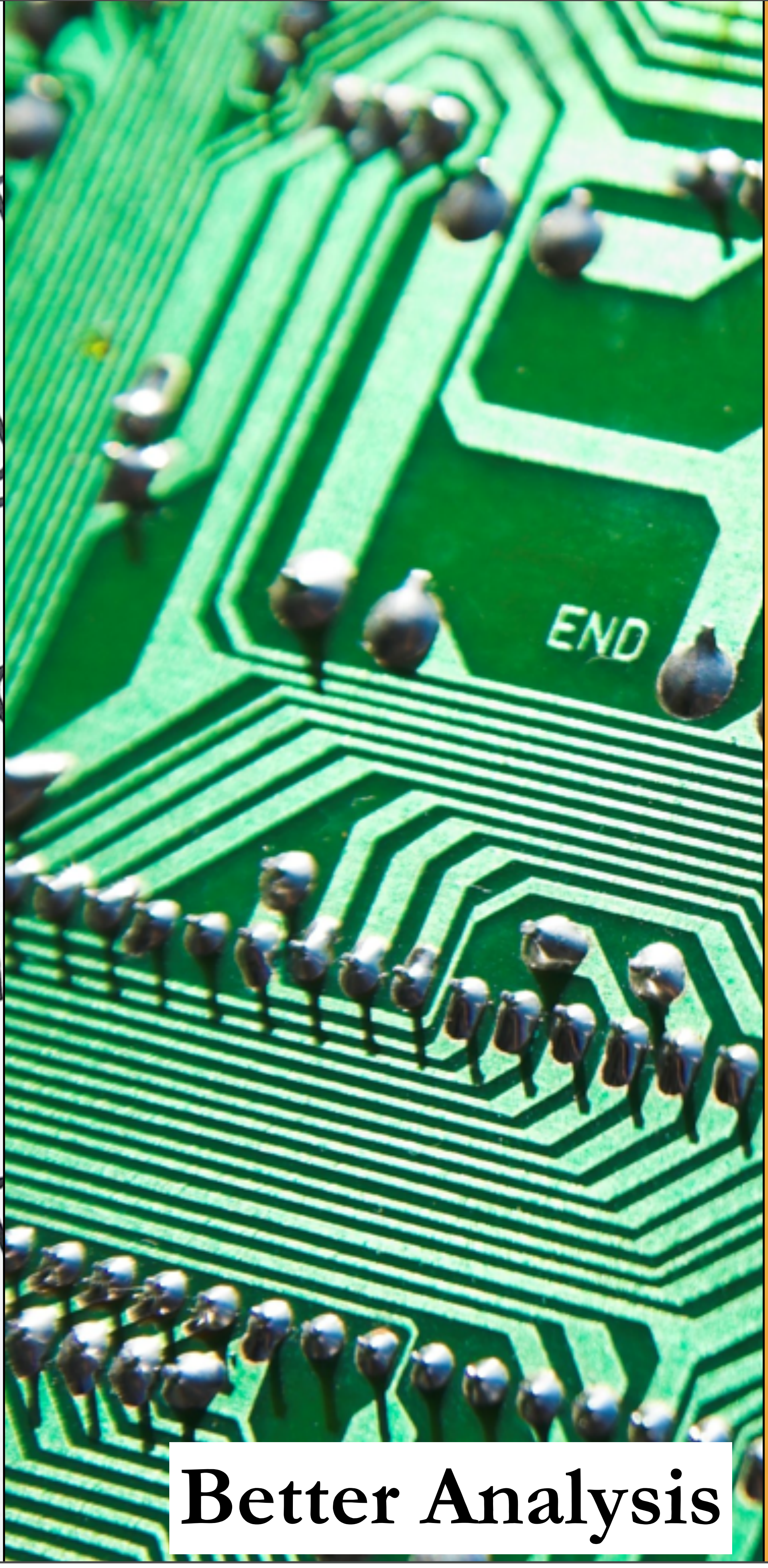
Selection Properties

Name	Value
Token Name	money
Type	word
Total Count	47
File Count	13
Cluster Count	0

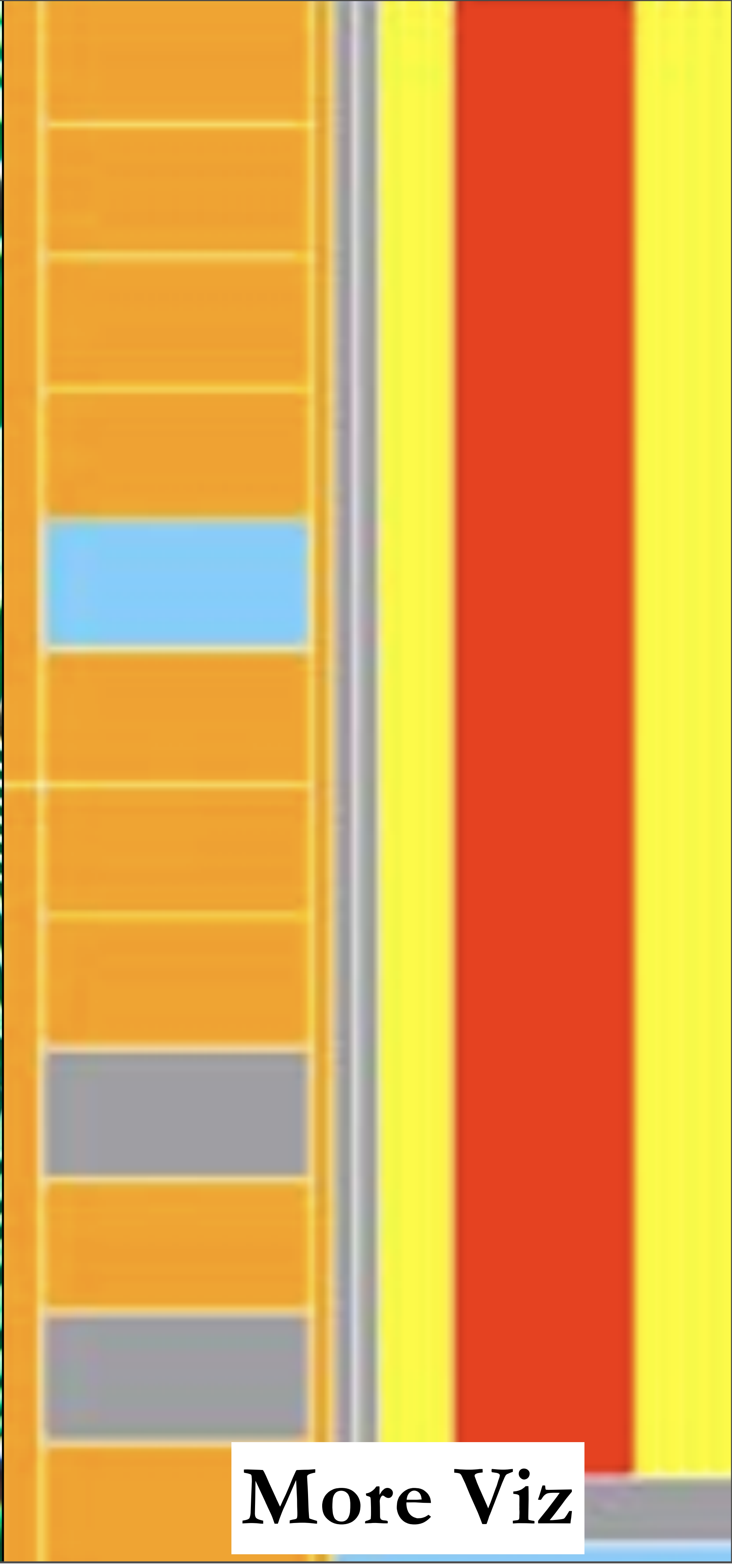
Visualizations



Validation



Better Analysis



More Viz

Future work: We seek to validate the effectiveness of our methods with follow up studies with forensic officers to close the loop. In addition, there are additional methods we can use to improve things: We can use better means to perform the initial analysis (such as FPGAs to speed it up or text-analysis tools to build better text models) and we can think of additional visualization characteristics (such as adding metadata visualization ala Teerlink and Erbacher. [Images courtesy http://www.sxc.hu/photo/866529, http://www.sxc.hu/photo/1135097, Teerlink & Erbacher]

A Visual Analytic Framework for Exploring Relationships in Textual Contents of Digital Forensics Evidence

T.J. Jankun-Kelly, D. Wilson, A. Stamps, J. Franck, J. Carver, J. E. Swan II
Mississippi State University & University of Alabama

VizSec 2009

VisWeek 09
VIS • INFOVIS • VAST

Funded by NSF CyberTrust (#CNS-0627407)

MISSISSIPPI STATE UNIVERSITY™

JAMES WORTH BAGLEY COLLEGE OF ENGINEERING
MISSISSIPPI STATE UNIVERSITY