XKEYSCORE

Tech Strings in Documents

(aka Tech Extractor)

December 2009

Derived From: NSA/CSSM 1-52
Dated: 20070108
Declassify On: 20291123

DERIVED FROM: NSA/CSSM 1-52

08

What is the Tech. Extractor?

- The "Tech Extractor" is a way of finding valuable intelligence based on keywords in the content of DNI sessions but it is a departure from traditional "soft selection" which tends to bring back a lot of junk.

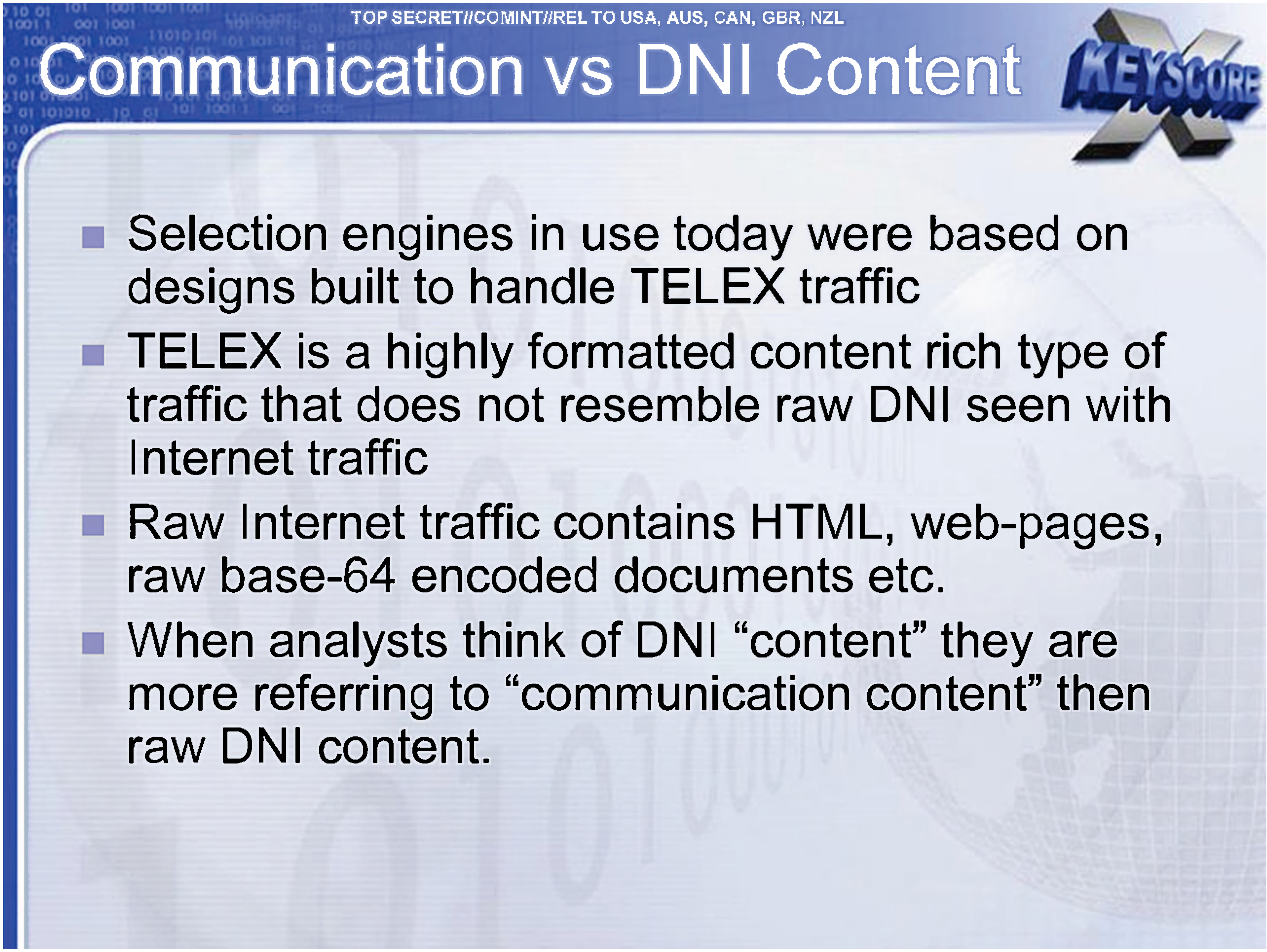
What is soft selection?

- Soft selection, aka content based selection, is an approach at targeting traffic by looking for keywords or phrases rather than specific E-mail accounts
- Content based selection has suffered because of the poor design of content based selection engines

Soft Selection vs Surgical Selection

- Existing selection techniques are blunt instruments
- XKEYSCORE contextual dictionaries provide an extremely sharp knife to make accurate selection decisions

"That's not a knife.....*THAT's* a knife!"



Communication vs DNI Content

- Selection engines in use today were based on designs built to handle TELEX traffic
- TELEX is a highly formatted content rich type of traffic that does not resemble raw DNI seen with Internet traffic
- Raw Internet traffic contains HTML, web-pages, raw base-64 encoded documents etc.
- When analysts think of DNI “content” they are more referring to “communication content” then raw DNI content.

Communication vs DNI Content

- If an analyst tasks a Boolean equation “bomb” and “chemical” they likely want to see all communication that mentions ‘bomb’ and ‘chemical’ and not all web pages, news stories, blog posts etc. where those two words appear
- What we need is a context-aware scanning engine that knows where it is inside of the raw DNI in order to properly apply analyst tasking

What is the Tech Extractor

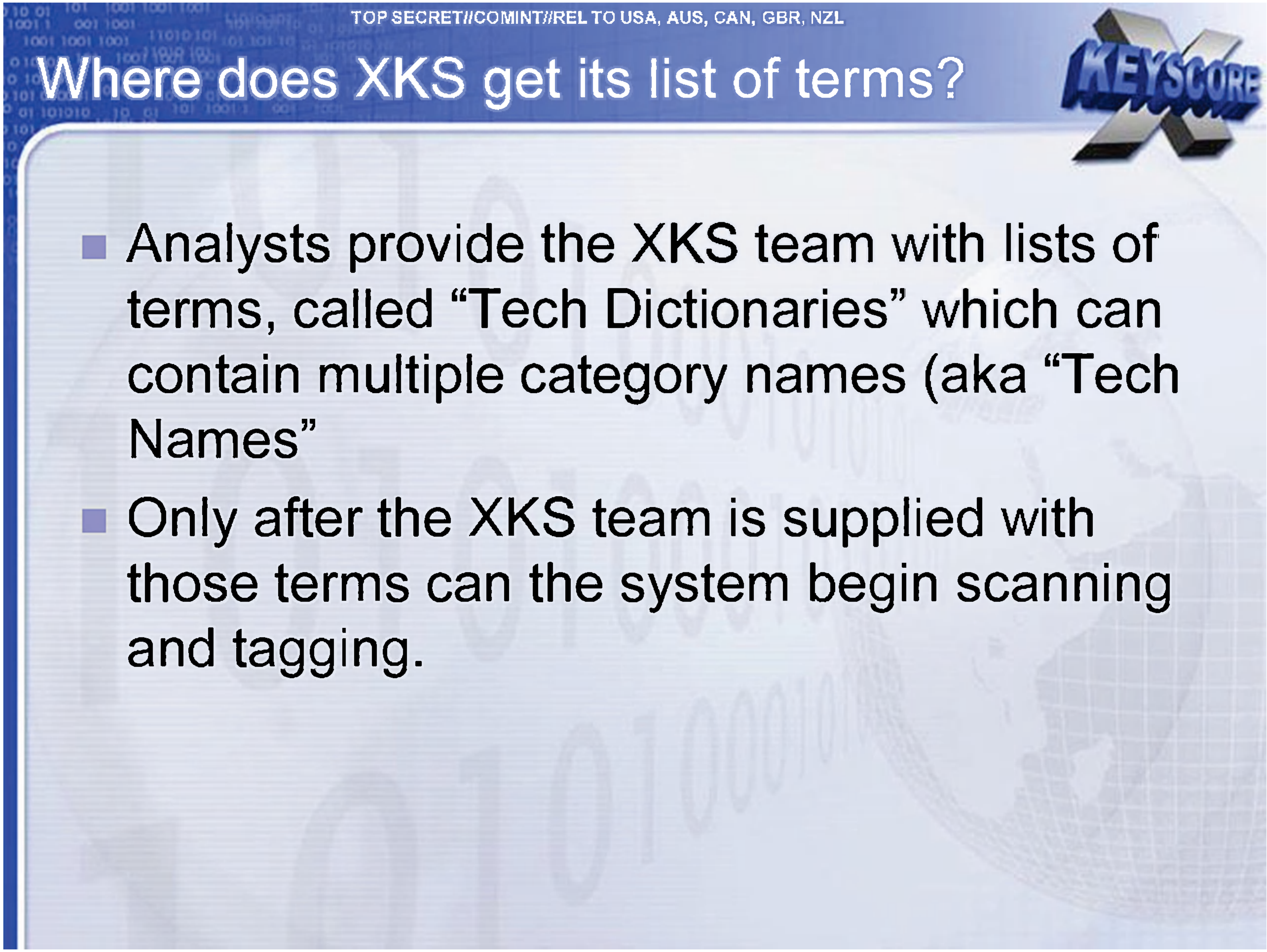
- The Tech Extractor was X-KEYSCORE’s first stab at context-aware scanning and it only focuses on three contexts:
  - E-mail Bodies
  - Chat Bodies
  - Document Bodies:
    - Microsoft Word, Excel, PowerPoint, Project, Visio
    - Adobe PDF, Postscript
    - Rich Text Format (RTF)

How does the Tech Extractor work?

- The Tech Extractor works by scanning a list of keywords against those three contexts and then tagging the results.
- It's important to note that this is not "filtering and selection" and we're not forwarding any data home
- XKS is simply tagging sessions with meta-data, much like we do with appids+fingerprints

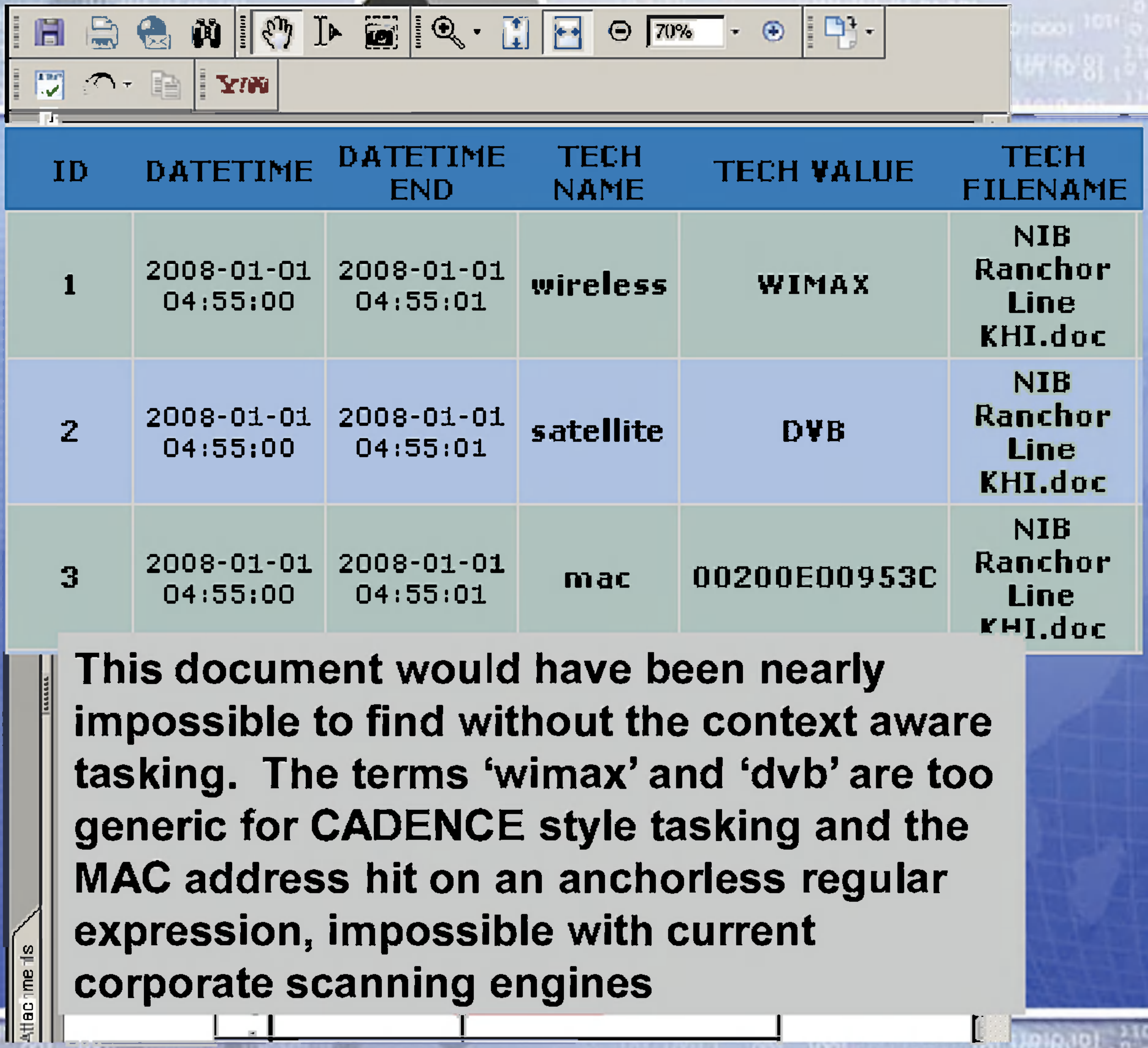
How does the Tech Extractor work?

- After the meta-data tag is applied, analysts can then use that meta-data tag as part of a *USSID-18* compliant query for traffic
- It's important to note, just like AppIDs+Fingerprints, Tech Extractor tags aren't necessarily USSID-18 compliant by themselves. You may need to add a valid foreign IP address, MAC address or country code before you query!



Where does XKS get its list of terms?

- Analysts provide the XKS team with lists of terms, called “Tech Dictionaries” which can contain multiple category names (aka “Tech Names”
- Only after the XKS team is supplied with those terms can the system begin scanning and tagging.



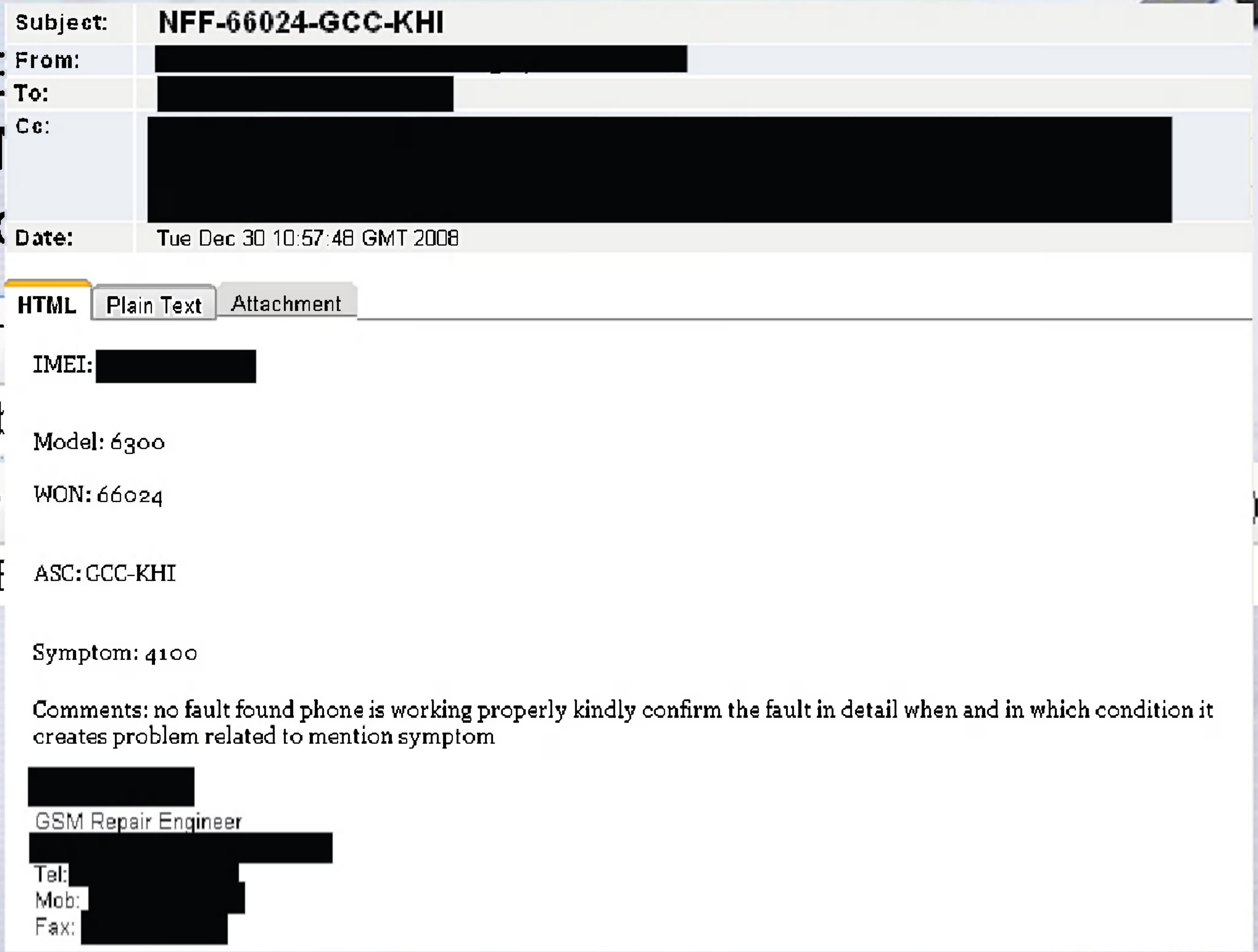
ID	DATETIME	DATETIME END	TECH NAME	TECH VALUE	TECH FILENAME
1	2008-01-01 04:55:00	2008-01-01 04:55:01	wireless	WIMAX	NIB Ranchor Line KHI.doc
2	2008-01-01 04:55:00	2008-01-01 04:55:01	satellite	DVB	NIB Ranchor Line KHI.doc
3	2008-01-01 04:55:00	2008-01-01 04:55:01	mac	00200E00953C	NIB Ranchor Line KHI.doc

This document would have been nearly impossible to find without the context aware tasking. The terms 'wimax' and 'dvb' are too generic for CADENCE style tasking and the MAC address hit on an anchorless regular expression, impossible with current corporate scanning engines

Context-Aware Tagging

- E
  II
  t(

Event T
email_t
Fm City
KLOSTE



Subject: NFF-66024-GCC-KHI

From: [redacted]

To: [redacted]

Cc: [redacted]

Date: Tue Dec 30 10:57:48 GMT 2008

IMEI: [redacted]

Model: 6300

WON: 66024

ASC: GCC-KHI

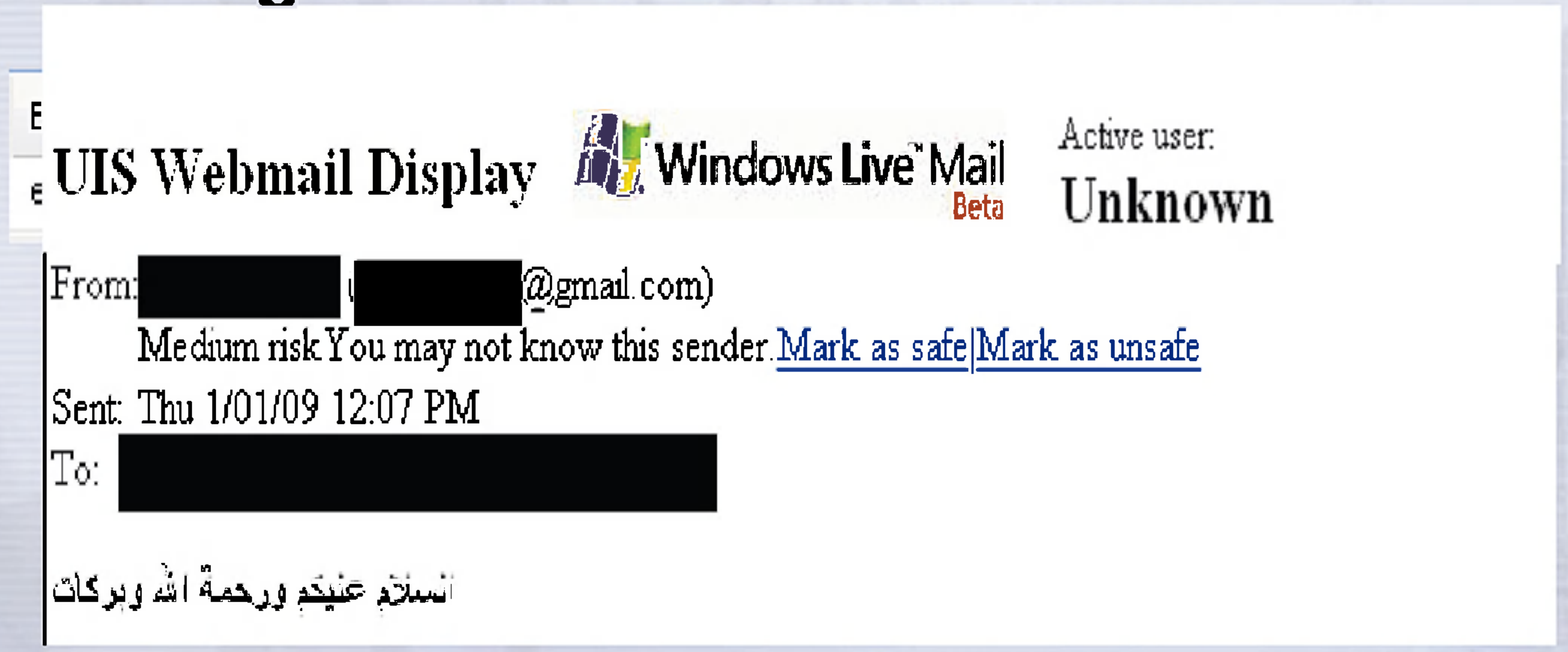
Symptom: 4100

Comments: no fault found phone is working properly kindly confirm the fault in detail when and in which condition it creates problem related to mention symptom

[redacted]
GSM Repair Engineer
[redacted]
Tel: [redacted]
Mob: [redacted]
Fax: [redacted]

Full Foreign Language Support

- Supports full foreign language tagging and querying
- Ex look for common Arabic expressions in E-mails coming from the Pakistan tribal regions:



UIS Webmail Display

Active user: Unknown

From: [redacted] ([redacted] @gmail.com)

Medium risk You may not know this sender. Mark as safe | Mark as unsafe

Sent: Thu 1/01/09 12:07 PM

To: [redacted]

السلام عليكم ورحمة الله وبركات

Live Demo