



Overview

- Problem: Conventional digital forensic tools were designed for native speakers and languages similar to English.
- Agenda
 - Intro to Digital Forensics
 - Intro to Character Encodings and Unicode
 - Challenges to Digital Investigations
 - *Equivalent Encodings*
 - *Orthographic Variations*
 - Odyssey Digital Forensics Search

Warning

I'm not a linguist and speak only English and programming languages



What is Digital Forensics?

- A process to answer questions about digital events and states.
 - Who broke into the server?
 - Does this computer contain intelligence on X?
 - Who does this computer belong to?
- The process is similar to a physical crime scene investigation:
 - Preserve the crime scene
 - Do a basic search for obvious evidence
 - Do more detailed searches and reconstruct events
- Keyword searching is a common technique used to find evidence / intelligence.



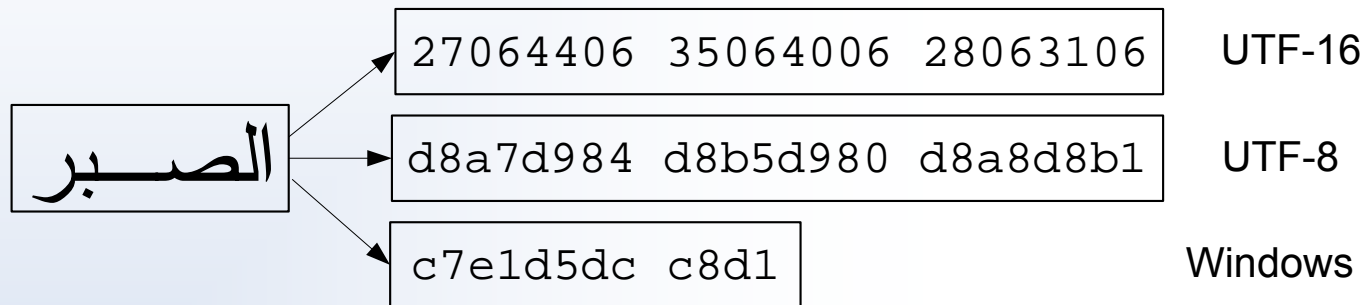
Basic Concept of Character Encoding

- Character encodings play a critical role in keyword searching.
- Computer hardware knows only about numbers.
- Characters are stored by assigning a number to them.
- Many encodings exist:
 - ASCII, Unicode, ISO 8859-X, IBM, Windows, DOS, ISO 2022-X, Shift-JIS, EUC-XX, ...
- Examples:
 - A -> 65 (ASCII / Unicode)
 - B -> 66 (ASCII / Unicode)
 - ш -> 1096 (Unicode)
 - ح -> 1581 (Unicode)
 - ح -> 205 (Windows 1256)



Conventional Keyword Search Process

- Investigator enters the keywords (s)he is looking for
 - May need to select the code pages to consider
- Tool determines the byte sequence for each keyword and each code page.
- Tool searches for files with one of the byte sequences.
- Tool presents the investigator with the list of files that contain one of the byte sequences.





Challenges to Digital Investigations

1. Multiple possible encodings and storage formats
2. Encodings may have equivalent characters
3. Language-specific orthographic variances
4. Numeral systems
5. Extracting text
6. Viewing results



Problem: Multiple Encodings

- The keyword could be in any of multiple encodings
- Investigators may not know all encodings and which apply to each language.
- Unicode values can be stored in five different ways:
 - UTF-8:
 - *Used internally by Unix systems and on Internet and in HTTP*
 - UTF-16:
 - *Used internally by Windows and Office*
 - UTF-32:
 - UTF-16 and UTF-32 can be big or little endian!



Unicode Storage Examples

- Latin Uppercase A: A (U+0041)
 - UTF-8: 0x41
 - UTF-16LE: 0x4100
 - UTF-16BE: 0x0041
 - UTF-32LE: 0x41000000
 - UTF-32BE: 0x00000041
- Arabic Waw with Hamza Above: و (U+0624)
 - UTF-8: 0xD8A4
 - UTF-16LE: 0x2406
 - UTF-16BE: 0x0624
 - UTF-32LE: 0x24060000
 - UTF-32BE: 0x00000624



Problem: Equivalent Encodings

- Many languages have *diacritics*
 - A mark added to a character to alter a word's pronunciation or to distinguish words
- Characters with diacritics can be encoded as:
 - A single “pre-composed” character that includes the diacritic
 - Multiple symbols that include the base character followed by the diacritics

á = a + ´

U+00E1 U+0061 + U+00B4

ؤ = و + ء



Equivalent Encodings (Japanese)

- Japanese, Korean, and Latin characters have both half-width and full-width encodings

Full-width	カタカナ	SHELL
Half-width	かたか	SHELL



Problem: Language Orthographic Variances

- Orthographic variations are different words for the same thing:
 - color versus colour
- Chinese government created a simplified set of symbols in 1950s - called simplified Chinese.
 - Taiwan still uses traditional Chinese.

	Hair	Computer
Simplified	头发	计算机
Traditional	頭髮	電腦



Orthographic Variances (Japanese)

- Japanese text can use four scripts (at the same time):
 - Hiragana
 - Katakana
 - Kanji (Chinese ideographs)
 - Latin (Romanji)
- All words can be written in Hiragana, Katakana, or Romanji.
 - Many (but not all) can be written in Kanji.
 - Typically, each script is used for a different type of word.
 - Computer (should be in Katakana):

Hiragana	こんぴゅうた
Katakana	コンピュータ
Kanji	電子計算機
Romanji	Konpyuuta



Orthographic Variances (Arabic)

- Arabic vocalization marks may not be present
 - لَغْوِيَّة
 - لُغْوِيَّة
 - لَغْوِيَّة
- Alif maqsura, yeh, and farsi yeh are sometimes used interchangeably:
 - العَرَبِي Alif maqsura (U+0649)
 - العَرَبِي Farsi yeh (U+06CC)
 - العَرَبِي Yeh (U+064A)



Orthographic Variances (Arabic)

- Arabic Kashida (U+0640) is used to connect characters - number used is based on page width or author's style
- Examples of the same word (found on the Internet):
 - الصبر (0627 0644 0635 0640 0628 0631)
 - الصبر (0627 0644 0635 0640 0640 0628 0631)
 - الصبر (0627 0644 0635 0640 0640 0640 0628 0631)
 - الصبر (0627 0644 0635 0640 0640 0640 0640 0640 0628 0640 0640 0640 0631)
 - الصبر (0627 0644 0640 0640 0635 0640 0640 0640 0640 0628 0640 0640 0631)



Orthographic Variances (Arabic)

More Arabic examples:

Arabic, Farsi and Urdu Text Normalization for Natural Language Processing

By: Zina Saadi at 4:30PM

2006 Conference Presentation:

Orthographic Variations in Arabic Corpora

By: Bushra Zawaydeh and Zina Saadi

<http://www.basistech.com/knowledge-center/>



Problem: Equivalent Numerals

- Different languages have different numbering systems.



Arabic Numerals



Eastern Arabic-Indic Numerals

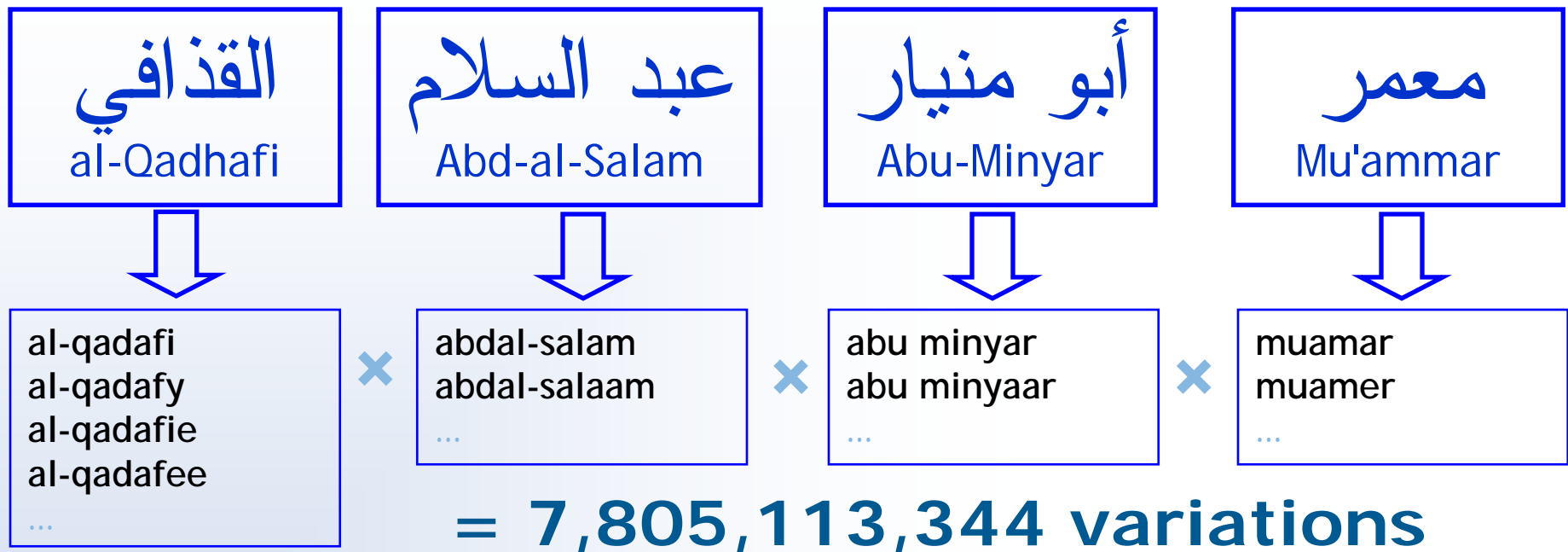
$$١٥٤٦ == ١٥٤٦ == 1546$$

- There are additional Asian numerals as well...



Problem: Name Variations

- May know only the transliterated form of name
- May need to search for the transliterated form of name



Different Script, Same Name: Tools for Matching & Translation.
By: David Murgatroyd (10:30)



Problem: Romanized Alphabets

- The native script is not always available for chat and IM.
- A “Romanized alphabet” is used instead
- Possible Romanized variants of قاسم
 - Qasim / Qasem
 - Gasim / Gasem
 - 8asim / 8asem
 - Kasim / Kasem
 - Asim / Asem
 - 2asim / 2asem
 - 'asim / 'asem
- Decoding Arabic Chat, Bushra Zawaydeh (10:30 AM)



Result: Too Many Combinations!!

- Orthographic variations
 - Script variations
 - Name variations
 - Encoding variations
 - Equivalent encodings
-
- All of these steps require extensive language knowledge!



Result: More Combinations = More False Positives

- A byte sequence may mean something in every encoding
- Consider the byte sequence: 8E E6 8F C1

Encoding	Characters
CP932	取消
CP936 / GB18030	産徠
CP949 / EUC-KR	룸 룰
CP1251	ТжЦБ
HKSCS	舉袜



A New Approach to Keyword Searching

1. Extract / carve files from disk
2. Identify file type and extract text
3. Normalize Text:
 - Identify code page & language
 - Convert to Unicode
 - Normalize Unicode
 - Normalize orthographic variances
 - Normalize numbers
 - Perform case folding (maybe)
 - Perform stemming (maybe)
4. Save normalized Unicode text to index
5. Normalize search query (step 3) and search index.



Results of New Approach

- Analyst does not need to know about:
 - Language-specific code pages
 - Language-specific orthographic variations
- Only text is searched
 - Reduces false positives from arbitrary byte values in data
- Increase of true positives
- Decrease of false positives



Odyssey Digital Forensics™ Search

- Basis Technology tool that implements the new approach:
 - Sleuth Kit is used to extract files
 - RLP is used to normalize the text
 - Normalized text is saved to an index
- English to Arabic keyword translation.
- Ignores “known files” using hash databases.
- Can load files of predefined keywords.
- Reads disk images or folder of files.
- Can sort files by language.



Viewing the Documents

- How do non-speakers interpret the non-English results?
- Automated translation would be ideal, but needs more work.
 - Context sensitive: When is “baker” a last name versus a profession?
- Intermediate approach:
 - Translate names of people, places, and dates.
 - Name translation is not based on context.
- Process:
 - Use named entity extraction technology to analyze text and identify names of people, places, organizations, etc.
 - Transliterate names to Latin alphabet.

Abdul Aziz Alomari

في حين أن السرية الثالثة بقيادة هاني حنجر ضربت مبنى وزارة الدفاع الأمريكية (البنتاغون) في العاصمة واشنطن، ولم يحالف النجاح السرية الرابعة بقيادة زياد الجراح في ضرب هدفها حيث تحطمت في بنسلفانيا بمن فيها من الركاب، ولم يحدد الشريط هدفها قبل تحطمها. ويحتوي الشريط على وصية أحد منفذي الهجمات وهو عبد العزيز العمري الملقب بـ أبو العباس الجنوبي. وهدد فيها بضرب المصالح الأميركية في كل مكان ما لم تخرج القوات الأميركية من الجزيرة العربية



Related Problem: Intelligent Text Extraction

- Sometimes, we have a large blob of data with unknown types and quantity of text in it.
- English text extraction is easy:
 - Look for sequences of 4 or more valid ASCII values
 - 37% of possible byte values are valid.
- What is a valid Unicode / code page sequence?
 - Roughly 80% of possible 2-byte values are valid Unicode.
 - Some sequences are less likely (based on language and linguistics)
- Language-specific models to reduce false positives are needed



Summary

- Odyssey shows that digital forensics benefits from:
 - Text normalization
 - Name translation
 - Keyword translation
- Future:
 - Incorporating name matching technology
 - Advanced extraction of non-English text from data
 - Incorporation of information retrieval techniques
 - *Ranking*
 - *Filtering*
 - *Clustering*
 - *Mining*
 - Cross-drive correlation of entities and keywords



Brian Carrier

Director of Digital Forensics

carrier@basistech.com

<http://www.basistech.com>