

How to be a CSI (encoding Crime Scene Investigator)



Y!

How to be a CSI (encoding Crime Scene Investigator)

Tex Texin
Internationalization Architect
Yahoo Inc.


How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 2

Y! Objectives for Crime Scene Investigation

- Have some fun
- Prevent death by bullet points
- Introduce strategies for analyzing character corruption
- Compare TV Show and Software investigation of mutilated character corpses
- The latest version of the presentation is at www.i18nguy.com/How-To-Be-A-CSI.pdf

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 2

Y! Our CSI Heroes
TV CSI vs. Software CSI



Gil Grissom
Catherine Willows
Nick Stokes
Sara Sidle
Greg Sanders
Warrick Brown
Dr. Al Robbins
Capt. Brass

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 3


Y! Our Software CSI Heroes



How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 4

How to be a CSI (encoding Crime Scene Investigator)

Crime Discovery On TV



- First steps
- Close off crime scene and gather clues
 - Take lots of odd pictures
 - Collect potential forensic evidence
 - Identify the victim(s) (or what's left of them)
 - Interview witnesses
 - They lie! They presume! They are prejudiced!
 - Put down either the jaded or newbie cop
 - If it's suicide then how do you explain...

How to be a CSI (encoding Crime Scene Investigator)
copyright © 2006 Tex Texin All rights reserved.
5

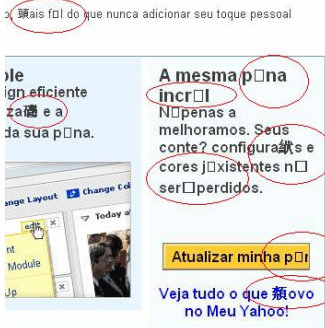
Crime Discovery in Software

A woman calls 911 yelling: “There has been a Mojibake! Someone committed Mojibake!”*

Characters have died.

Decomposition has already set in.


Who did it?



*Mojibake = Japanese for garbage characters.

How to be a CSI (encoding Crime Scene Investigator)
copyright © 2006 Tex Texin All rights reserved.
6

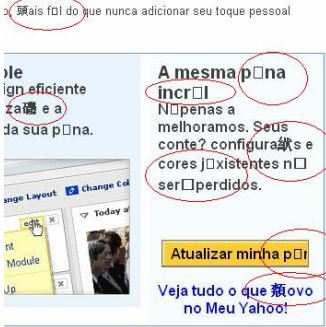
Crime Discovery in Software



- Characters Corrupted
 - Wrong characters
 - White or black boxes
- Collect forensic evidence
 - Screen captures, Keystrokes, Environment
- Witnesses/Users: “It happens when I...”
 - “The entire application is busted”
 - They lie, They presume, They’re prejudiced
- The newbie or jaded developer
 - “That vendor always does that. Our code is fine.”

How to be a CSI (encoding Crime Scene Investigator)
copyright © 2006 Tex Texin All rights reserved.
7

Collecting Images



- Often included in bug reports
 - Better than nothing
 - You can detect patterns
 - But often not helpful
- Better to collect:
 - Expected bytes/chars
 - Actual bytes/chars
 - Fonts referenced, available
 - Expected/detected encoding
 - Keystrokes
- A way to reproduce

How to be a CSI (encoding Crime Scene Investigator)
copyright © 2006 Tex Texin All rights reserved.
8

How to be a CSI (encoding Crime Scene Investigator)

Y! Example Case History

- Developers refused to debug based on an urban legend-
 - “Crime happens. Can’t catch these guys.”
 - “It’s Microsoft IE’s fault- Our software is fine”
- Instead of:
`<meta http-equiv=Content-Type content="text/html; charset=utf-8">`
- The code said:
`<meta http-equiv=Content-Type content="text/html; charset=utf8">`
(Note the missing hyphen in “utf8”)

How to be a CSI (encoding Crime Scene Investigator)

copyright © 2006 Tex Texin All rights reserved.

9

Y! Analysis On TV

- Autopsy: More data collection
- Possibilities: What could have happened?
- Conjecture: How could it have happened?
- Reconstruction: Does it fit the data?
- Research: Does other data fit the theory?
 - Always ask the most narrowing question last
- Experiments
 - Measurements and proof of concept
 - Decay rates under abnormal circumstances

How to be a CSI (encoding Crime Scene Investigator)

copyright © 2006 Tex Texin All rights reserved.

10

Y! Analysis on TV

Anatomy
and
decay
science
Autopsy



Typical computer queries

- How many trucks on the road?
- How many on route 1?
- How many on route 1 at 3pm?
- How many stopped at Joe’s bar?

Trajectory
Recon-
struction



Experiments with controls

How to be a CSI (encoding Crime Scene Investigator)

copyright © 2006 Tex Texin All rights reserved.

11

Y! Analysis in Software

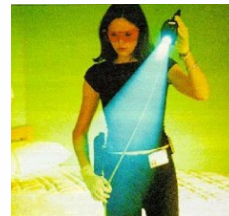
Core
dumps,
architecture
and
debugging



Typical computer queries

- Names of encodings
- Relations between encodings
- Variations of encodings
- Quality of encoding conversions

Data flow
analysis
And points
of
conversion



Known data injection

How to be a CSI (encoding Crime Scene Investigator)

copyright © 2006 Tex Texin All rights reserved.

12

How to be a CSI (encoding Crime Scene Investigator)

Y! Analysis In Software

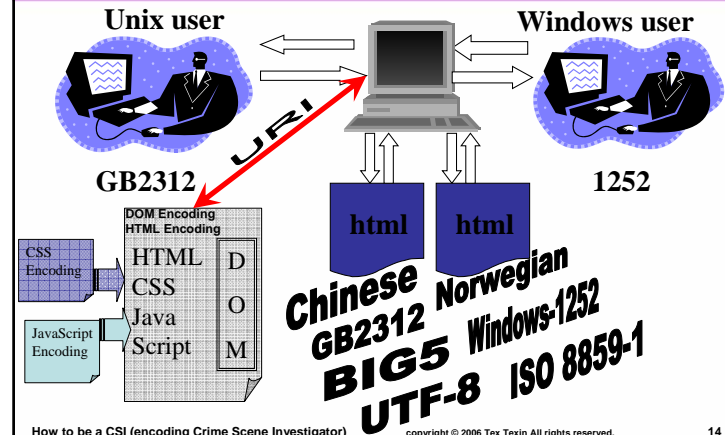
- Architecture: The Web is complex
 - Clients, Browsers, mix technologies (DOM, java, css, php, active-x, applets, etc.), *ml standard+non-standard behaviors, OS dependencies, etc.
 - Servers, protocols are also mixed
 - Language, locale, encoding, are ambiguous
 - Negotiation tactics are unreliable
- But must be understood & kept up-to-date
 - See Web Internationalization Tutorial, et al.

How to be a CSI (encoding Crime Scene Investigator)

copyright © 2006 Tex Texin All rights reserved.

13

Y! Character Encoding Negotiation



How to be a CSI (encoding Crime Scene Investigator)

copyright © 2006 Tex Texin All rights reserved.

14

Y! Architecture



- Data flow and points of (mis)conversion
 - Interfaces are common failure points
 - Between Legacy, new and 3rd party software
 - API, web services, protocols, devices
 - Data sources, sinks
 - Types of misconversion
 - Missing conversion Source ➡ Source
 - Incorrect conversion Source ➡ X
 - backwards conversion Target ➡ Source
 - double conversion (Source ➡ Target) ➡ Target



How to be a CSI (encoding Crime Scene Investigator)

copyright © 2006 Tex Texin All rights reserved.

15

Y! Architecture

- Other contributors to conversion errors
 - Missing or incorrect encoding identifier
 - False detection
 - e.g. ASCII vs UTF-7
 - Too little data to detect
 - Text not like corpus used for detection statistics
 - Encoding name variations
 - "I didn't know it was a crime. I just did what I was told!"

How to be a CSI (encoding Crime Scene Investigator)

copyright © 2006 Tex Texin All rights reserved.

16

Y! Encoding Name Variations

- E.g. ISO 8859-1 vs Windows-1252

0	0 - 1F Control Codes																															
20	!	"	#	\$	%	&	'	()	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
80	80-9F Control Codes in ISO 8859-1																															
A0	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı		
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 17

Y! Encoding Name Variations

- E.g. ISO 8859-1 vs Windows-1252

0	0 - 1F Control Codes																															
20	!	"	#	\$	%	&	'	()	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
80	€	‚	ƒ	„	…	†	‡	•	‰	Š	‹	ƒ	Ž	‘	”	”	•	–	—	™	›	œ	ž	ˆ	˜	˘	˙	˚	˛	˜		
A0	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı		
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	

1252-specific characters

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 18

Y! Encoding Naming Variations

- Several variations of big-5, shift-jis, etc.
 - E.g. Microsoft, Apple added chars to shift-jis
- Then there is big5-hkscs
 - Hong Kong Supplementary Character Set
- Application dependent names
 - Windows IE expects “KSC5601” or “Korean”, not cp949, windows-949
 - Firefox would expect euc-kr
- Sometimes the error is font not encoding

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 19

Y! Encoding Name Variations

- Poor naming conventions in the industry
- Application dependent

Language(s)	Traditional Encoding Name	iconv Name	ICU Name
Chinese Traditional	BIG-5	BIG5-HKSCS	ibm-1375_P100-2003
Japanese	EUC-JP	EUC-JP	ibm-33722_P12A-1999, ibm-954_P101-2000
Korean	EUC-KR	CP949	windows-949-2000, ibm-1363_P118-1998
Chinese Simplified	GB2312	GBK	windows-936-2000
Western Europe	ISO-8859-1	CP1252	windows-1252
Greek	ISO-8859-7	CP1253	windows-1253

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 20

How to be a CSI (encoding Crime Scene Investigator)

Y! Encoding Conversions

- Correct conversion is not always obvious:
 - 0x5C is it a file separator “\” or currency “¥”?
 - (or Won, or other currency)
 - Half-width or Full-Width?
- Sometimes you need different converters



Y! Missing Character Conversions

□	□	,	f	...	†	‡	‰	§	€	□	□	□	□	,	“	”	,	–	—	™	›	œ	□	□	Ÿ					
ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı	ı					
À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ
à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ

West European Code Page ISO 8859-1

Р	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	Г	
Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	Ÿ	
А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Э	Ю	Я
а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	э	ю	я

Russian Code Page 1251

Converts to “?” without error code.

Y! Encoding Conversions

- The family secret: Versioning
 - Unicode is an evolving standard
 - As characters are added the opportunity for improved conversions exist
 - “Which Unicode version is the converter for?”
 - If I convert data today how will it compare with the same data converted 3 years ago?
 - How long after death before rigor mortis sets in and when do the maggots come?



Y! Breaking Multibyte Characters

- By not treating all bytes as one character
 - Don’t insert bytes in the middle
 - Don’t delete 1 byte of a multi-byte char.
 - Be careful with block boundaries
 - Don’t treat individual bytes as characters
 - E.g. uppercase a trailing byte
 - E.g. Treat trailing byte 0x5C as “\” in file pathnames

Y! Making Mojibake

Splitting multi-byte characters

	日		本		語	
Byte type:	L	T	L	T	L	T
Bytes:	9	F	9	7	8	E
	3	A	6	B	C	A

Byte types:
L = Lead Byte
T = Trail Byte

Inserting “a” (61) in second byte

	殿		至		語	
Byte type:	L	T	L	T	S	L
Bytes:	9	6	F	9	7	8
	3	1	A	6	B	C

a

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 25

Y! Encoding Pathology and Forensics

- Piecing together who did what to whom
 - Knowing what was expected and what resulted, and comparing with typical patterns of failure, we can deduce what occurred
- Do all characters fail or certain ones?
- How do they fail? Wrong characters, Question marks, Blackboxes, etc.

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 26


Y! Encoding Forensics and Pathology

- Question Marks
 - Conversion to an encoding that doesn't have the characters replaced by “?”.
- All non-Ascii characters are mojibake
 - Missing, wrong, backwards, or 2x conversion
 - Expected “ç”. Actual: “Ã§”
 - Expected 0xE7 (ç). Actual: 0xC3 0xA7 (Ã§)
 - Note that U+00E7 in UTF-8, “ç” is 0xC3 0xA7
 - Conclusion: UTF-8 bytes displayed as ISO8859-1

How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 27

Y! Encoding Forensics and Pathology

- Only certain characters misconverted
 - Variant encoding, out of date conversion, data flow, incomplete font
 - E.g. Only euro, trademark, smart quotes, OE ligature,...
 - Used iso-8859-1 to utf-8 conversion instead of windows-1252 to utf-8



How to be a CSI (encoding Crime Scene Investigator) copyright © 2006 Tex Texin All rights reserved. 28

Encoding Error Patterns



- **Comparing conversion error patterns is like identifying the gun that fired a bullet by the shared striation marks**
- **It's encoding DNA**
- **Often there are obvious candidate patterns**
 - “Round up all the usual suspects”
 - **A stoolie (er... tool) for testing conversions**
 - Given an input string, try the usual and expected error conversions, See if any of the results match
 - **Other tools: Encoding validator, byte to %hh**

Encoding Research

- How many trucks stopped at Joes bar?
- Look for variants you might be incurring
 - www.iana.org/assignments/character-sets
 - Conversions supported
 - See iconv, ICU, or doc for your converter
 - Check for font versions/updates

Data Injection



- Testing with live data feeds
 - like being on a stakeout- Hoping the criminal will come by while you wait
- Create a pseudo-feed or simulate data entry for “controlled experiments”.
 - Known, repeatable values simplify debugging.
 - Less threatening for debugging foreign languages
 - Inject directly into subsystems to eliminate other components as “suspects”.

Setting Traps for Criminals

- Use validation routines at key points of data acceptance or conversion
 - Especially useful for UTF-8, which has illegal byte patterns.
 - Reference:
www.w3.org/International/questions/qa-forms-utf-8.html

How to be a CSI (encoding Crime Scene Investigator)



Conclusions: How to be a CSI

- Gather forensic evidence
- Be objective, be thorough
- Understand the architecture, the data flow, the points of conversion
- Identify the potential patterns of failure

- “Concentrate on what cannot lie. The evidence” -Gil Grissom



Thank You!

¿Questions?

Note: Images from the CSI TV show and those of CSI cast members are courtesy of CBS Broadcasting Inc. All Rights Reserved, and/or the actual copyright owners.

The owner and copyright for these photos were generally not documented from the sites I downloaded them. If notified I will update this file to include the correct copyright.