The Hitchhiker's Guide to Chinese Encodings

Tom Emerson (湯姆·愛摩森)
Senior Computational Linguist

20th International Unicode Conference
Washington, D.C., USA

strategy • process • technology • results

www.basistech.com

BASIS TECHNOLOGY

Overview

- Who am I and why am I here?
- What do we mean by “Chinese?”
  - Simplified vs. Traditional
- Chinese *Character Sets*
- Introduce *Chinese Encodings*
- Driving Forces
- Reality vs. Idealism
- Transcoding Issues

BASIS TECHNOLOGY

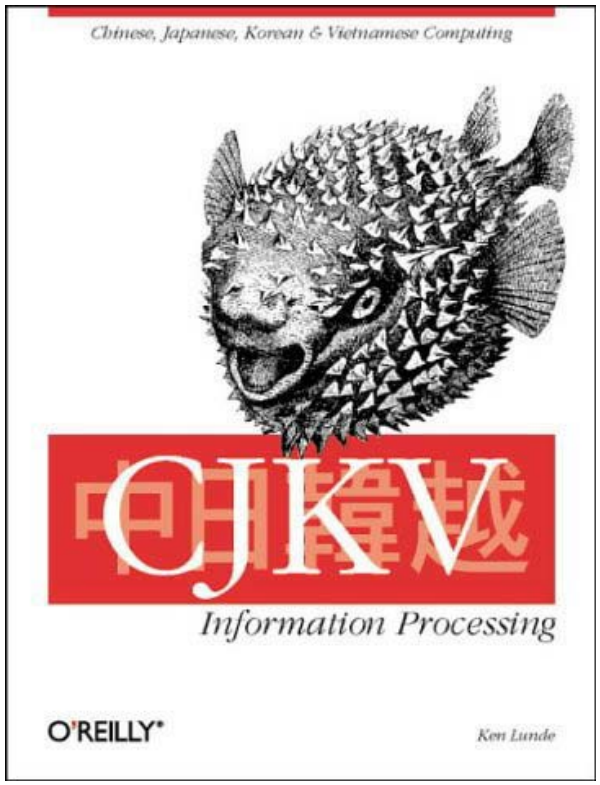
Who Am I?

- “Sinostringologst” at Basis Technology
- Lead developer for our Chinese Morphological Analyzer and our Chinese Script Converter
- Background in both Computer Science and Linguistics

Who Are You?

Don’t Panic!

BASIS TECHNOLOGY



The Blowfish Book is *the* reference for anyone working with CJK character sets. Get it.

BASIS TECHNOLOGY

What is “Chinese?”

- For our purposes we are interested in the written language.
- In general this means Mandarin.
- Topolects (方言) sometimes define their own hanzi for local words, usually for names.
- Hence “Written Cantonese” doesn't make a lot of sense.

BASIS TECHNOLOGY

Simplified vs. Traditional

- “Simple” and “Full” Form
- Mainland China and Singapore use “Simplified Chinese”
- Hong Kong, Taiwan, and Macao use “Traditional Chinese”

BASIS TECHNOLOGY

Simplification

- Fewer strokes
  - Easier to learn
  - Easier to remember
  - Easier to write
- Compare: 台 vs. 臺
- Simplification is not recent
  - Some simplified characters in current use date to the pre-Qin period (pre 246 B.C.E.)

BASIS TECHNOLOGY

Simplification

1956: *Scheme for Simplifying Chinese Characters*

1964: *The Complete List of Simplified Characters* (2236 characters)

1986: *The Complete List of Simplified Characters, 2nd*

BASIS TECHNOLOGY

Simplification

國際大購併所產生的國際經濟趨勢、將至少主導未來十年經濟的發展。從去年度開始的購併潮中、主角都是行業之首。

国际大购并所产生的国际经济趋势、将至少主导未来十年经济的发展。从去年度开始的购并潮中、主角都是行业之首。

BASIS TECHNOLOGY

Character Sets vs. Encodings

- Non-Coded Character Sets
  - A non-coded character set represents a list of characters that are defined by an organization as the standard set that one is expected to know.
  - Tōngyòng (7000), Chángyòng (2500), and Cìchángyòng (1000)
- Coded Character Sets
  - A coded character set assigns a unique number (“code point”) to each abstract character in the repertoire.
  - A coded character set does not make any statements about how its code-points are represented on a computer.

BASIS TECHNOLOGY

Character Sets vs. Encodings

- An encoding specifies how the code points in a coded character set are to be represented and transmitted with a computer.
- A single character set can have multiple encodings.
- Sometimes the distinction is blurred: Big Five and GB18030-2000 both define a character set and an character encoding.
- Generally laid out in one or more 94x94 grids. Each character is indexed by its row-cell address (qūwèi) within the grid.

BASIS TECHNOLOGY

Character Sets

- Simplified
  - GB 2312-80
- Traditional
  - GB 12345-90
  - CNS 11643
  - Big Five, Big Five Plus, EThen
  - GCCS and HKSCS
- “Generic”
  - Unicode/ISO 10646/GB 13000-1992 (Unicode 1.1)
  - GB 18030-2000

BASIS TECHNOLOGY

Encodings

- Simplified
  - HZ
  - CN-GB
  - EUC-CN
  - CP936
- Traditional
  - EUC-TW
  - Big Five et al.
  - CP950
- Unified
  - Unicode
  - GB 18030:2000
  - ISO 2022-CN and ISO 2022-CN-EXT
    - In a sense more complete than Unicode since it encodes multiple legacy character sets.

BASIS TECHNOLOGY

Encodings

- My hovercraft is full of 鰻魚
  - 鰻 is at 87-9 (0x57 0x09)
  - 魚 is at 51-67 (0x33 0x43)
- My hovercraft is full of 鰻魚
  - 鰻 is at 1-92-13
  - 魚 is at 1-62-03

BASIS TECHNOLOGY

HZ

- Modal 7-bit, multi-byte encoding
- Encodes US-ASCII and GB 2312-80
- Defined by RFC 1843 and RFC 1842
- Developed by Lee Fung Fung at Stanford University
- MIME name is HZ-GB-2312

BASIS TECHNOLOGY

HZ

- Default mode is ASCII with an escape sequence used to switch to GB mode.
  - "~{" switches into GB
  - "~}" switches out of GB encoding
  - "~\n" is the continuation character that can be used at the end of a line.
  - "~~" is the tilde in ASCII mode

BASIS TECHNOLOGY

HZ

- My hovercraft is full of 鰻鱼
  My hovercraft is full of ~{w)Sc~}
- Need to add 0x20 to each row-cell value to bring all row-cell points into printable range.
  - 鰻 is at 87-9, 0x57-0x09
  - Adding 0x20 yields 0x77-0x29, or w)

CN-GB

- Non-modal 8-bit, multi-byte encoding
- Encodes US-ASCII and GB 2312-80
- Defined in RFC 1922
- MIME name is CN-GB
- AKA “GB-2312”. Just say no.

BASIS TECHNOLOGY

CN-GB

- My hovercraft is full of 鳗鱼
  My hovercraft is full of \xF7\xA9\xD3\xE3
  My hovercraft is full of ÷©Óã
- 鱼 is at 51-67 (0x33 0x43)
  - Adding 0x20 brings the row-cell value into printable range, 0x53-0x63
  - Now set the high-bit by adding 0x80, 0xD3-0xE3

BASIS TECHNOLOGY

CN-GB and HZ Summary

- Both encode GB-2312 row-cell values directly.
- To convert HZ to row-cell, subtract 0x20 from each byte.
- To convert row-cell to HZ, add 0x20 from each value.
- To convert CN-GB to row-cell, subtract 0xA0 from each byte.
- To convert row-cell to CN-GB, add 0xA0 to each value.

BASIS TECHNOLOGY

ISO 2022 Overview

- ISO 2022 is a complex 7-bit, modal encoding standard
  - ECMA 35
  - GB 2311-1990
  - CNS 7654-1989
- Multiple character sets can be encoded in a single document
- Character sets are selected through “designators” and “shifts”
  - Shifting characters
  - Single shift sequence

BASIS TECHNOLOGY

ISO 2022 Overview

- A “designator” specifies the character set associated with a particular shift into double-byte mode
  - Consists of the designator sequence
    - \x1B \x24
  - Followed by the shift type
    - \x29 - \x2B
  - Then the character set designation.

BASIS TECHNOLOGY

ISO 2022 Overview

- A “shifting character” switches between single- and double-byte modes
  - SO shifts to double-byte mode (\x0E)
  - SI shifts to single-byte mode (\x0F)
- A “single shift sequence” invokes double-byte mode for only the next two bytes.
  - SS2 (\x1B \x4E)
  - SS3 (\x1B \x4F)

BASIS TECHNOLOGY

ISO 2022 Overview

- Character sets are registered with ISO's International Registry for Escape Sequences
  - http://www.itscj.ipsj.or.jp/ISO-IR/

BASIS TECHNOLOGY

ISO 2022-CN

- Modal 7-bit, multibyte encoding
- Encodes four character sets:
  - ASCII
  - GB 2312-80
  - CNS 11643-1986 Plane 1
  - CNS 11643-1986 Plane 2
- Defined in RFC 1922.
- MIME name is ISO-2022-CN

BASIS TECHNOLOGY

ISO 2022-CN

- Designators for the double character sets are:
  - GB 2312-80 [ISO IR 57]
    Esc $) A
  - CNS 11643-1986, plane 1 [ISO IR 171]
    Esc $) G
  - CNS 11643-1986, plane 2 [ISO IR 172]
    Esc $ * H

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻魚
  My hovercraft is full of \x1B$)A\x0Ew)Sc\x0F
- Yikes!

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻魚
  My hovercraft is full of \x1B$)A\x0Ew)Sc\x0F
- The designator for the GB 2312-80 character set

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻魚
  My hovercraft is full of \x1B$)A\x0Ew)Sc\x0F
- The S0 shift

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻魚
  My hovercraft is full of \x1B$)A\x0Ew)Sc\x0F
- The printable GB 2312-80 row-cell values for each hanzi (should look familiar!)

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻魚
  My hovercraft is full of \x1B$)A\x0Ew)Sc\x0F
- SI shift back to ASCII

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1
  - The designator for GB 2312-80

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1
  - S0 into GB 2312-80

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1
  - The printable GB 2312-80 row-cell values for the simplified hanzi

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1
  - SI back to ASCII

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1
  - The designator for CNS 11643-1992 Plane 1

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1
  - S0 into CNS

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1
  - The printable values of the CNS row-cell values for the full-form hanzi

BASIS TECHNOLOGY

ISO 2022-CN

- My hovercraft is full of 鰻鱼 and 鰻魚
  ... of \x1B$)A\x0Ew)Sc\x0F and \x1B$)G\x0E|-^#\x0F
- Here we've mixed characters from GB 2312-80 and CNS 11643-1986 Plane 1
  - SI back to ASCII

BASIS TECHNOLOGY

ISO 2022-CN-EXT

- Modal 7-bit, multibyte encoding
- Encodes the four character sets in 2022-CN, as well as:
  - ISO-IR-165 (aka CCITT)
  - CNS 11643-1992, Planes 3-7
- Defined in RFC 1922.
- MIME name is ISO-2022-CN-EXT
- Also supports several other character sets, such as GB/T 12345-90, though these aren't registered with ISO. See RFC 1922 for details.

BASIS TECHNOLOGY

ISO 2022-CN-EXT

- Supports nine character sets:
  - GB 2312-80 [ISO IR 57]
    Esc $) A
  - ISO-IR-165 (aka CCITT) [ISO IR 165]
    Esc $) E
  - CNS 11643-1986, planes 1 - 2 [ISO IR 171-172]
    Esc $) G, Esc $ * H
  - CNS 11643-1992, plane 3 - 7 [ISO IR 183-187]
    Esc $ + I,J,K,L,M

BASIS TECHNOLOGY

ISO 2022-CN Finale

- Designator sequences *must* appear on each line containing characters from a double-byte character set
- SO is ambiguous:
  - GB 2312-80 or CNS-11643 Plane 1?

BASIS TECHNOLOGY

EUC

- Extended Unix Code
- Defined in ISO-2022/ECMA 35 as an 8-bit encoding form.
- Defines four “code sets”
  - Code Set 0 is one-byte, ASCII
  - Code Sets 1 – 3 are language dependent, and may be two or more bytes wide
- Two shift characters are defined:
  - SS2 (\x8E) prefixes each character in code set 2
  - SS3 (\x8F) prefixes each character in code set 3

BASIS TECHNOLOGY

EUC-CN

- 8-bit, non-modal double-byte encoding
- Encodes GB 2312:80 in Code Set 1
- MIME name is EUC-CN
  - AKA GB2312, GB_2312-80, GB, ISO-IR-58
- Same as CN-GB described in RFC 1922

BASIS TECHNOLOGY

EUC-TW

- 8-bit, modal multi-byte encoding
  - ASCII in Code Set 0
  - CNS 11643-1992 Plane 1 in Code Set 1 (2 bytes)
  - CNS 11643-1922 Planes 1 – 7 in Code Set 2 (4 bytes)
  - Code Set 3 is not used
- MIME name is EUC-TW
  - AKA CNS11643

BASIS TECHNOLOGY

EUC-TW

- CNS 11643 Planes 1 – 7 are encoded in Code Set 2
  - SS2 \xA*n* *R* *C*
    where *n* is the plane number, 1 – 7
    R and *C* are the 8-bit hex row-cell value
- All non-ASCII text is fixed width.

BASIS TECHNOLOGY

Big Five

- 8-bit, non-modal double-byte encoding
- Industrial Standard, *not* a national standard
  - However, it has become the de facto standard for encoding Traditional Chinese text
- Corresponds to CNS 11643-1992 planes 1 and 2
- Includes an extension mechanism, allowing vendor defined and user-defined character blocks
- MIME name is CN-BIG5

BASIS TECHNOLOGY

Big Five

Character Range	Contents
0x8140 – 0x8DFE	User-Defined Area 3
0x8E40 – 0xA0FE	User-Defined Area 2
0xA140 – 0xA3FE	Big 5 Symbols and Controls
0xA440 – 0xC67E	Big 5 Level 1 Hanzi
0xC6A1 – 0xC8FE	Vendor-Defined Area
0xC940 – 0xF9D5	Big 5 Level 2 Hanzi
0xF9D6 – 0xF9FE	Vendor-Defined Area
0xFA40 – 0xFEFE	User-Defined Area 1

BASIS TECHNOLOGY

Big Five

- Big Five has been extended by many organizations.
  - ETen
  - Microsoft CP950
    - One row of ETen
  - Big Five Plus
    - Unicode 2.x Ideograph repertoire
  - GCCS & HKSCS
  - HKUST EUDC
  - …

BASIS TECHNOLOGY

Big Five

- All of the extensions share the VDA and UDA blocks.
- Big Five means different things to different people, depending on what set of extensions they have in mind.
  - Big Five on Windows is not Big Five on MacOS
  - CP950 is not Big Five
    - It isn't even Big 5 with ETen extensions, sometimes!

BASIS TECHNOLOGY

Big Five Plus

- Big Five extension adding the remaining hanzi from Unicode 2.0.
  - Uses parts of UDAs 1, 2, and 3
  - Rarely, if ever, seen
    - Wenlin and TwinBridge Chinese Partner

BASIS TECHNOLOGY

GCCS

- Government Chinese Character Set
- Developed by the Hong Kong government 1994
- Includes 3099 characters
  - Based on Big Five
  - Encoded in UDAs 1 and 2
    - Conflicts with Big 5+ and other vendor extensions to Big Five
- Replaced in 1999 by HKSCS, though it is still used by some sites.

BASIS TECHNOLOGY

HKSCS

- Hong Kong Supplementary Character Set
  - Specification developed in 1999 by the HKSAR
  - New characters can be submitted to the HKSAR for inclusion.
  - Repertoire mandated in Hong Kong
- Two allocation schemes
  - Big Five
    - An update to GCCS
    - Still has the problem of stomping on other Big Five extensions such as Big 5+ and ETen
  - Unicode
    - Mappings split between BMP, Plane 0 PUA and Plane 0 and Plane 2 ideograph blocks

BASIS TECHNOLOGY

HKSCS

- Unicode Allocation (cont.)
  - There are different mappings for HKSCS and Unicode 2.x, 3.0, and 3.1.
  - Even with Ideographic Extension Block B in Plane 2 there are six characters that must be encoded in the Plane 0 PUA.
  - And more may be added…
- HKSCS support required by SAR government
  - ISO 10646 with HKSCS extensions is the official representation

BASIS TECHNOLOGY

GB 18030:2000

- Chinese National Standard
  - Released March 2000
  - Second printing, with corrections, May 2001
- Descendent of GB 2312 and GBK
  - Backwards compatible with GB 2312
  - Adds the character repertoire of Unicode 3.0
  - Ready for additional Unicode code-points in Plane 1 and beyond
- Now required to sell software in the PRC!
- Mixed 1-, 2-, and 4-byte encoding

BASIS TECHNOLOGY

GB 18030:2000

- New software *must* support GB 18030:2000 to be sold in China.
  - Products released after 1 September 2001 must satisfy the “relevant” requirements.
  - Products released between 31 March 2000 and 31 August 2001 must be updated.
  - “Standard Conformity Technology Standardization Research Institute”.
  - What if you don’t? Beats me.
    - But is it worth the risk?
- Easiest way to support GB 18030:2000 is to Unicode enable your Products!

BASIS TECHNOLOGY

Transcoding Issues

- The repertoire of these character sets can be different
  - GB 2312:80 and CNS-11643 are dramatically different
  - GB/T 12345:90 and CNS-11643 sometimes differ in the traditional variant used
- Big Five means different things to different people. This can lead to nightmarish scenarios.
  - CP950 may really be HKSCS depending on what the user has done on their machine.

BASIS TECHNOLOGY

Transcoding Issues

- Zhu Rongji, China's Premier, writes his name as 朱镕基
- 镕 is not encoded in GB2312
- 鎔 is a variant of 镕, but you cannot use it to write his name
- 镕 *is* encoded in GBK, AKA CP936
- But most Chinese web pages specify "GB2312" as their charset. D'oh.

BASIS TECHNOLOGY

Questions

- Final version of these slides will be available at http://www.basistech.com/info/