The Hitchhiker's Guide to Chinese Encodings

Tom Emerson (湯姆·愛摩森)
Sinostringologist

18th International Unicode Conference, Hong Kong

Software Internationalization Services & Technology

Dedication

Rebecca M. Peckham
1944 - 2001

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

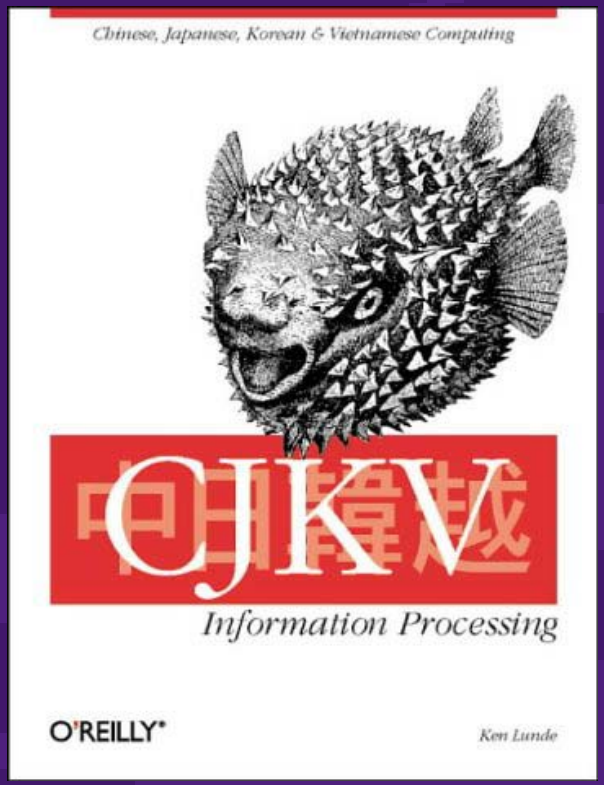
Who Am I?

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

BASIS TECHNOLOGY

Who Are You?

Don’t Panic!



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。

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.

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, ETen
  - 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.
  - Big 5+
  - CP950
- Unified
  - UTF-8, UTF-16, UTF-32
  - ISO 2022-CN and ISO 2022-CN-EXT
    - In a sense more complete than Unicode since it encodes multiple legacy character sets.

BASIS TECHNOLOGY

HZ

- Modal 7-bit, multi-byte encoding Developed by Lee Fung Fung at Stanford University, and defined by RFC 1843 and RFC 1842
- Encodes US-ASCII and GB 2312-80
- 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.

BASIS TECHNOLOGY

HZ

- This is 汉子.
  This is ~{::WS~}.

- 汉 is at 26.26 (0x1A.0x1A) and 子 at 55.51 (0x37, 0x33) in GB 2312:80
- Need to add 0x20 to each row-cell value to bring all row-cell points into printable range.

BASIS TECHNOLOGY

CN-GB

- Described in RFC 1922.
- 8-bit multi-byte encoding of GB 2312:80.
- Each code-point in GB 2312:80 has its 8-bit set: no mode switch is necessary.
- Instead of adding 0x20 to each row-cell value, just add 0xA0 (0x20 + 0x80) to generate the 8-bit value of GB 2312:80.
- This is 汉子.
  This is ºº×Ó.

BASIS TECHNOLOGY

ISO 2022 Overview

- ISO 2022 is a complex 7- and 8-bit encoding standard.
  - ECMA 35 and GB 2311-1990
- It allows characters from multiple character sets to be used within a single document.
- Achieved through the use of “designators” and “shifts”.

BASIS TECHNOLOGY

ISO 2022 Overview

- A "designator" specifies the character set associated with a particular "shift"
- ISO 2022 is a "modal" encoding: at any point in the input stream you are in a particular mode which indicates how to interpret the current character.
- ISO 2022 is line-based, resetting to 7-bit ASCII at the end of each line. You must include the designators for the character sets used on each line before the first use.
- Character sets are registered with ISO.

BASIS TECHNOLOGY

ISO 2022-CN

- Defined in RFC 1922.
- Supports three character sets:
  - 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-EXT

- Defined in RFC 1922
- 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, plane3 3-7 [ISO IR 183-187]
    Esc $ + I,J,K,L,M

BASIS TECHNOLOGY

EUC-CN

- Encodes GB 2312:80
- Essentially the same as CN-GB

EUC-TW

- Encodes eight character sets, ASCII and CNS 11643-1992 Planes 1-7
- ASCII is represented with 1 byte, CNS 11643-1992 plane 1 in 2 bytes, and CNS 11643-1992 1-7 (sic.) in 4-bytes.
  - 0x8E 0xAn row cell
    - Where n == the plane number being encoded

BASIS TECHNOLOGY

Big Five

- http://www.cmex.org.tw/big-5.html
- Corresponds to CNS 11643 planes 1 and 2

BASIS TECHNOLOGY

Transcoding Issues