Encoding Schemes

Joe Celko

copyright 2015
We encode data to get it into a database as symbols
- Alphabets
- Numbers
- Symbol

Various way to formally manipulate the codes
- Math for numbers
- String operators

You can put “direct” data into databases these days, but most of it is still encoded
- Music, pictures, etc. are hard to search
UNICODE requires that all language character sets include a “minimal ASCII” subset

ISO Standard encodings are all based on this subset

Latin Alphabet – no accents, no case sensitivity
  - If a position can be numeric, then some alphas are disallowed

Digits – the base ten model for the majority of encodings

Punctuation – comma, dash, period, slash, underscore
  - Marks like #, @, &, etc. have special meaning in various languages
Display is important

Fixed Length versus Varying Length
- Fixed length is part of validation
- Varying length requires Regular Expressions

People read text in “bouma” or “chunks”
Groups of 3 are best, but up to 5 can work
- Ex: 512-845-7871 is a US phone number
- 5128457871 is a string of digits
Bad Encoding Schemes -1

- **Does not allow for growth in its domain**

- **Georgia Auto tags in the 1970's**
  - Georgia auto tag type codes started as one digit on a punch card
  - Commemorative tags got popular - every college, veterans group, popular cause wanted one
  - The codes became a mess of special multi-punches on the cards that had to be translated in the file system.

- **American Honda**
  - “We will never have more than 10,000 dealerships in the United States”
Bad Encoding Schemes -2

- Ambiguous codes

- ISBN
  - International Standard Book Number
  - 10 characters and four parts (language, publisher, book number, check digit)
  - Language, publisher and book number are variable length
  - There have been ISBNs that can be parsed two ways

- A “miscellaneous” code that gets used a lot is a bad sign
Bad Encoding Schemes -3

- **Lack of support for exceptions**
  - Unknown values
  - Missing values
  - Non-applicable value
  - Miscellaneous or unclassified
  - Overflows, underflows, division by zero, etc.
  - Errors in one field
  - Errors in more than one field (pregnant male)
  - Computable but not known

- **SPARC committee listed 14 kinds of missing data** (Interim Report 75-02-08)

- **Someone else published 22 kinds of missing data**
Bad Encoding Schemes -4

- If you think designing encoding schemes is not important, do math in Roman Numerals for a week
- Try living without alphabetical order for a week
- Find a book in library organized by color instead of Dewey Decimal Classification
- Queries and aggregations can be made much easier with a good encoding scheme
- Calculations are more accurate, too.
Enumeration Codes

- Lists the values and assigns a name or tag number to them

- This is a Nominal Scale under another name

- Good idea to order the code symbols in some order for use
  - Chronological - which values appear first in time
  - Procedural - steps in the order of a task
  - Physical - rainbow color order
  - Sort their codes in alphabetical or numerical order
Measurement Codes

- Column in the database is known to represent units in a certain scale

- The value is expressed in the unit of the column
  - cannot do math on mixed units without conversion

- Worst design to have the unit and measure in the same value
  - dollars shown with $ in the column
  - fine for display, not for storage

- Have a related column which tells you the unit being used and let it drive conversions -- (23.45, ‘US $’), (54.75, ‘Euro’)
Abbreviation Codes

- A shorten version of the name of the value being encoded
- One to one mapping
- Can be figured out by a human reading it
- Can be variable or fixed length
- Three-letter Airport names
  - Pretty good for major airports LAX, BOS, ATL, etc.
  - Pretty weird for minor airports -- anything in Alaska’s back country
Algorithmic Codes

- Use a procedure to encode a value
- Not immediately human readable
- Encryption
- Rounding numbers
  - There are all kinds of rounding functions, but that is another topic
- Hashing functions
Hierarchical Codes -1

- Partition the set of values into disjoint subsets, then partitions the subsets until some final level is reached

- Usually numerics, but can be mixed alphanumerics
  - Library of Congress Classification is mixed
  - Dewey Decimal Classification is numeric

- ZIP code partitions on geography
Hierarchical Codes -2

- Can put something in the wrong part of the tree
  - Dewey Decimal has logic under philosophy and not math

- Can fail to allow enough space
  - Dewey Decimal puts all Eastern religions in one bucket

- Item can fall into more than one code -- Church architecture and the worship service can be religion and/or architecture
Vector Codes

- Made up of parts that cannot be separated from the whole entity, but have some meaning
  - Parts are not a complete fact

- Components can be dependent or independent on each other

- Dates -- year, month, day

- ISO Tire sizes -- width, material, diameter

- Social Security Numbers
Concatenation Codes

- Variable number of parts that are concatenated together
- Components can be ordered or unordered
- Keyword lists on documents
- Check lists
- Called “facet codes” in Europe
- Not in favor any more with computers
  - Used on old machine shop tags; each step was initialed as it was done
Guidelines

- Use existing ISO, national or industry-specific standard codes
- Avoid inventing your own encodings
  - Do you want to maintain them yourself?
  - Will anyone else use them?
- Allow for expansion in the codes
- Use explicit exceptional value codes
- Keep a translation of codes for the user in the database
  - Very common auxiliary tables