Data Technologies

R and File Formats

Plain Text Formats

- Fixed-width
- Delimited
- CSV
- XML (more later)

Background Reading

Chapters 7 and 8 of “Introduction to Data Technologies”
(***Now in HTML!!!***)

The “R Data Import/Export” Manual.

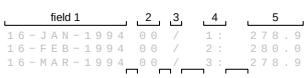
The `XML` documentation

The `ncdf` documentation

Computer Memory

- bits, bytes, and words

- basic types in R



File Formats

- Text versus binary
- Open versus proprietary
- Self-describing (or not)
- High-level versus low-level data models

Binary Formats

- Stata, SAS, SPSS, ...
- netCDF
- Spreadsheets

Example: netCDF

netCDF provides a public standard binary format and software libraries to read and write the format, so it is at least possible for any other software to read and write the format.

variables An n-dimensional array of values. Values can be integers, reals, or characters.

dimensions The size of a dimension.

attributes Extra information (metadata).

The format is "self-describing" (e.g., can ask how many variables, what the variable names are, ...).

R and XML

- Reading XML
- Writing XML

R and Plain Text

- Reading text
- Writing text

Assignment and Project

- `identical()`
- `as.integer()`
- `as.data.matrix()`

R and netCDF

- Reading text
- Writing text

XML

- Syntax
- Design
- Data integrity (DTD)