

**Data Technologies**

**R and File Formats**

## Background Reading

---

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

The “R Data Import/Export” Manual.

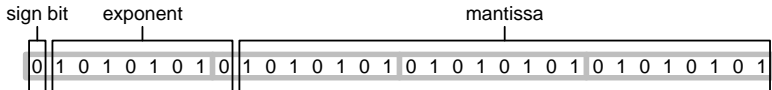
The XML documentation

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

## Plain Text Formats

---

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

field 1											2	3	4	5												
1	6	-	J	A	N	-	1	9	9	4	0	0	/	1	:	2	7	8	.	9						
1	6	-	F	E	B	-	1	9	9	4	0	0	/	2	:	2	8	0	.	0						
1	6	-	M	A	R	-	1	9	9	4	0	0	/	3	:	2	7	8	.	9						
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27

<u>field 1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>
16 - JAN - 1994	00	/	1 :	278 . 9
16 - FEB - 1994	00	/	2 :	280 . 0
16 - MAR - 1994	00	/	3 :	278 . 9

## 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 Plain Text**

---

- Reading text
- Writing text

## R and netCDF

---

- Reading text
- Writing text

# XML

---

- Syntax
- Design
- Data integrity (DTD)

## R and XML

---

- Reading XML
- Writing XML

## Assignment and Project

---

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