Digital Document Formats

Griffith Feeney

What is *Format*?

- Literally, its the way in which information in the document is translated into the long line of “bits” (0’s and 1’s) read by the computer
- Operationally, it specifies the kind of hardware and software we can use to access a digital document, *i.e.*, to view it, copy it, print it, and so on

Why is Format Important?

- You won’t be able to read, or copy, or print, or *do anything* with a document unless it is in a format that your computer hardware and software can process
- Different formats have different properties that make them more useful or less useful in particular contexts

Format Distinctions

- **Proprietary** *vs* **non-proprietary**
- **Application-specific** *vs* **Application-independent**
- **Platform-specific** *vs* **Platform-independent**
- **Text** files *vs* **binary** files
- These dichotomies not always strictly observed, some “fuzzy” boundaries

Production Formats and *Archiving/Distribution* Formats

- Formats we use to *produce* documents of various kinds reflect our choice of software applications for internal use
- Formats we use to *archive* and *distribute* documents are tailored to the needs of the recipients of the documents
- We need to *translate* production formats to archiving/distribution formats

Format for Archiving/Publication

- We need to select a suitable format for archiving and publication from among the several competing options
- We need effective procedures for translating our internal production documents to this “export” format for archiving and publication

Available Formats

- Ordinary text file
- Postscript
- T_EX
- HTML
- SGML
- “Portable document formats”: Adobe .pdf, Corel Envoy

Ordinary Text File Format

- Just about universally readable on any computer system without special software
- The *de facto* “universal document format”
- Best format for some purposes, but too limited for many other purposes
- Can’t handle even simple formatting, can’t handle images, can’t handle hypertext links

PostScript Format

- A “back door” format that emerged on the internet as a *de facto* standard
- PostScript is a *page description language*, a special purpose computer language used to command printers to draw marks on paper
- Any production application that can print to a PostScript printer can produce a postscript file; standard extension is *.ps*

T_EX Format

- A computer typesetting language created by computer scientist Donald Knuth
- Originated with demands of computer typesetting of mathematical formulas
- Given by Knuth to the American Mathematical Society, which makes it freely available

HTML Format

- HTML = Hypertext Markup Language is the language of the World Wide Web
- The phenomenal growth of the WWW has made HTML an important format
- HTML files are “marked up” text files that allow “hypertext” links to other documents
- Can contain images and some text formatting, but limited in later respect

SGML

- Standard Generalized Markup Language
- “The mother of all mark up languages”
- HTML is merely a special case
- An extremely powerful tool, but except for HTML, not yet widely adopted
- May emerge as an important general distribution format in the future; for now it is limited to very specific channels

Portable Document Formats

- Proprietary formats, strictly speaking
- Created and by companies that sell software for creating documents with these formats
- “Viewers” free on the Internet, so “anyone” can access the documents without paying software license fees
- Market leader is **Adobe Acrobat**; Corel **Envoy** is another example of the genre

Adobe Acrobat

- Product of Adobe Systems, Inc, 345 park Avenue, San Jose, California 95110-2704 USA
- Download free viewer free at www.adobe.com
- ‘Acrobat’ software, used to create “.pdf” files, about $200, $50 for educational institutions in the US
- “Use any Windows application” to create .pdf files or translate Postscript to .pdf

How it Works
(is supposed to work …)

- Create a document of any kind in any Windows application (versions available for other platforms as well)
- Print the document, but instead of directing it to the printer, direct it to the ‘printer driver’ supplied by Acrobat
- The result is a .pdf file that anyone with an Acrobat Viewer can access

Another Portable Document Format

- WordPerfect Suite 7 comes with **Envoy**, a portable document facility similar to that provided by Adobe's **Acrobat**
- The newer versions of WordPerfect also come with significant SGML facilities
- Microsoft is dominant at present, but WordPerfect is worth keeping an eye on

Review of Key Points

- What ‘format’ means
- Why it is important
- Production *vs* distribution formats
- Standard distribution formats
- ‘Portable document’ formats
- Adobe Acrobat software and .pdf format
- Rapid evolution in progress

Questions?
Comments?
Discussion?