Lecture 5 – Document Processing

Karl R. Wilcox
K.R.Wilcox@reading.ac.uk

Objectives

- **Discuss Document Processing**
  - **And some associated topics**
- **Understand why it is not the same as Word Processing**
- **Today's practical**
  - **Further formatting, replacing, spelling and grammar checking**

101 Things to do with a document

- **Q) What is the purpose of the document -**
  - **To be viewed on a screen?**
  - **To be printed and bound?**
  - **To be printed and displayed?**
  - **To be placed on a web site?**
  - **To form the content of an e-mail message?**
  - **To be archived as a matter of record?**
  - **To be re-published in a different medium?**
  - **To be used as data in a computer system?**
  - **…**
- **A) Any and all of the above**

The B52 Stratofortress

- Designed and built during the 1950's
- Has undergone continuous upgrades and modifications
- Currently 200+ in service, at modification level 'H'
- Further upgrades planned
- Expected to remain in service for a further 30 – 40 years

- Total service life of the aircraft (and it's documentation!) will be ≈ 100 years

Issue – Longevity

- Some documents exist for a very long time
  - Sometimes for archiving
  - Sometimes still being updated
- Other examples
  - Legal documents
  - Census data

- When choosing the medium for long lived documentation
  - The feature set of currently available word processors is not a major consideration
  - 18 month upgrade cycles are NOT welcome

The Updating Problem

- Consider a large document
- The document is important
  - It may contain legal, operational, or reference information
  - It is vital that the information is up to date
- The document has a wide circulation
  - But not massive, say 100 – 25,000 copies
- Changes are frequent
  - But not too frequent, say one every 1 – 6 months
- The content affected by the change is a small fraction
  - Say just a few percent of the total pages

Updating Examples

- Almost all “technical” manuals
  - Aircraft maintenance
  - Vehicle maintenance
  - Trouble shooting guides
- Legislation
  - Primary and secondary legislative documents
  - Guidance associated with those documents
- Standards documentation
- Sales and marketing information
- Part works

Two Options Considered

Consider 2,500 copies of a 500 page document, of which 25 pages are updated every 3 months

- Reissue the entire document
- 2,500 x 500 x 4
  = 5 million pages per year

- Reissue change pages + TOC etc.
- 2,500 x 50 x 4
  = 500,000 pages per year

Issue – Incremental Updates

- There can be considerable savings through incremental updates
- But incremental updates imply:
  - Page numbering by sections, not continuous
    - e.g. Page 2 – 14 – 3
  - Table of Contents and Index must follow likewise
  - There may be "knock-on" updates to following pages
    - Double sided publication is quite common
  - There may be a need for a "List of Effective Pages"
    - Each page has a version number
    - The LEP shows the version for each page

The Re-use Problem

- Consider a set of related documents
  - They may be for different variants of a “product”
  - They may have different target audiences
- There may be a high proportion of common content
  - Large amounts of re-used text
  - Graphics may also be re-used
- Common text and graphics should be edited ONCE
  - Edited in one place, re-used in others

The Graphics Problem

- **Graphics re-use can be difficult**
  - **May need different sizes, levels of detail**
  - **May need to reformat for different media**
- **Bitmap formats**
  - **Bitmaps are (potentially) very large**
  - **They do not scale well (up or down)**
- **Vector formats**
  - **Vector / Object based graphics better for re-use**
  - **Scalable, layered, editable text**
  - **There are too many vector formats to choose from!**
    - **CGM (versions 1-4), TIFF (lots of variants), WMF, SVG**

Issue – Content Reuse

- Some documents are made up from “fragments”
  - Fragments of text (“boilerplate”)
  - Graphics (in the broadest sense)
- But some content is specific to the document
  - Introductory material
- Some elements are “generated”
  - Table of contents
  - Table of figures
  - List of Effective Pages
- How are changes and versions managed?

Other Issues

- **Multiple Authors**
  - **Many large documents have more than one author**
    - **Simultaneously**
    - **Over the lifetime of the document**
- **Structured review and approval cycle**
  - **Documents may be subject to a specific review process**
  - **There may be a formal approval process**
- **Multiple output formats**
  - **Documents may be simultaneously published**
    - **On the web**
    - **On paper**
    - **On CD**

Can We Use A Word Processor?

- **The PC Word Processing Industry has not been good at producing long lived documents**
  - **Except ASCII text!**
- **Word processors are NOT good at incremental updates**
  - **Locating all changed pages (only *editing* changes are marked)**
  - **Generating LEPs not really feasible**
- **Word processors are NOT good at shared content**
  - **Bookmarks use absolute (filename) addressing**
- **Multiple output formats, formal review, not well supported**

What Can We Use?

- **Separating content from format helps a lot**
  - **Allows re-use & multiple output formats**
  - **Allows selection & ordering of content arbitrarily**
- **But – you can only format based on the markup**
  - **Need to mark up as much detail as possible**
    - **E.g. proper nouns, foreign languages, emphasis**
- **Document (content) management systems do address these issues**
  - **Designed for incremental updates**
  - **Manages shared content**
  - **Enforces review and approval cycles**
  - **May use COTS software tools as the "Front End"**

Conclusions

- Word Processing is NOT Document Processing
- A 5 page document is significantly different from a 500 page document
- 5 people jointly working on a document is significantly different from 50 collaborating on a document

- There are some publishing problems that exhibit ALL of the issues we have discussed today
  - For example aircraft technical documentation