



What is Digitization? Core Issues

Stephen Miller

AAC / ÖAW

Texts & Files WS 2005/06

Universität Wien



Interested Bodies

- Computer Science
- Literary and Linguistic Community
- Governments
- Library Science
- Industry



PreHistory

- “Machine-Readable” Texts
- Origins of digitization in the Humanities
- Authorship Attribution
- Literary Stylistics
- Corpora



Corpora

- Brown Corpus
 - 1 million words (1964)
- BNC
 - 100 million words (1994)
- Bank of English
 - 450 million words (January 2002)



Hypertext

- Hypertext does not begin with the WWW
- The WWW is an example only of hypertext
- PreHistory
 - Intermedia (Dickens Web)
 - Hypercard (Apple Macintosh)
 - StorySpace (Eastgate Systems)



Digital Information

- Not all digital information is “new”
- Availability in digital format is no guarantee of longevity
- Accumulated in a random manner
 - no difference from analog material



Multiplicity of Formats

- Multiplicity of Media

- Text
- Data
- Images
- Video
- Sound



Death of Formats

- “Vanished” software
- New versions of software



Death of Media

- Change in format / hardware
 - Punched Cards
 - Paper Tape
 - Magnetic Tape
 - Floppy Disks
 - Hard Disks
 - Video Disks
 - Video Tape



Decay of Media

- “Slow Fires of Decay”
 - Pioneer 10
 - 21 month mission ended 30 years later...
- Analog material also decaying
 - Sound recordings



“Brittle Books”

- 19th century print technology used cheap pulp for mass production
- Libraries now contain decaying books
- Libraries also face storage crisis
 - discarding of books and journals



Microfilm Mania

- Utopian visions of the library of the future
 - Microcards
 - Microfiche
 - Microfilm
 - Esp for newspapers
 - Destruction of the originals...



Is PDF the answer?

- Retention of “look and feel” of the original
 - allows dual print/electronic distribution
- Portable Document Format
 - specification is published
 - proprietary format
 - owned by Adobe
 - freeware browser (“Reader”)



Dream of Digitization

- “Two for the Price of One”
 - Preservation
 - Content Release



Digitization / Preservation

- Digitization is not Preservation
 - the physical object remains...



Preservation / Digitization

- Preservation is more than Digitization
 - the digital object now becomes the focus of preservation...



Core Issue

- How to manage all of this?!



Solution

- Must be platform independent
- Must be software independent
- Must be backed by standards



Technical Challenges

- Agreed formats for **Storage**
- Agreed languages for **Representation**
- Agreed procedures for **Reformatting**
- Awareness of **Migration** (“Refreshing“)
- Issue of **Resource Discovery** (Metadata)



Other Challenges

- Effective workflow and management
 - establishing “best practice”
- Embedding in institutional structures
 - a digitization programme is not an “add-on extra”
- Financing of digitization programmes



Implications

- Access to digital resources
 - who owns the content?
- Part of a wider debate over dissemination /access to scholarship
 - eJournals
 - Budapest Open Access Initiative
 - OAI
- Globalisation of Knowledge



Etienne Louis Boullée

