February 2005

People Profile: Alan Dawson

Editor

Follow this and additional works at: https://docs.lib.purdue.edu/atg

Part of the Library and Information Science Commons

Recommended Citation
Editor (2005) "People Profile: Alan Dawson," Against the Grain: Vol. 17: Iss. 1, Article 22.
DOI: https://doi.org/10.7771/2380-176X.4746

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
prescriptive but to illustrate the choices made in specific circumstances, in the belief that this level of detail will be useful to others planning similar initiatives. All the issues mentioned below have arisen during digitization of just six substantial books.

**Digitization Issues**

1. **Preservation vs accessibility**
   
   **Issue:** The demands of digital preservation and user accessibility are not incompatible but involve different priorities and may require compromises.
   
   **Policy:** Priority is given to making the content freely available, easily usable and readily searchable via open-access standards, hence the choice of XHTML for publication format. The page design of the printed books is not transferred to digital format (other than the cover or title page), but the original text and structure is preserved with a high degree of accuracy, and presented in accordance with guidelines for electronic textbook design (Wilson & Landoni, 2002) and accessibility (W3C, 1999/2004).

2. **Equipment selection**
   
   **Issue:** Accessible eBook creation requires digitising a printed book using an effective but non-destructive process (unless the book is already held in digital form).
   
   **Policy:** A standard flatbed desktop scanner is used for relatively small books that are not noticeably damaged by being fully opened at each page. A digital camera is used for large or valuable books that may be damaged by repeated scanning. If digital preservation of images is considered important, and the book is not suitable for flatbed scanning, a specialist agency is used for capturing images at high resolution, though this significantly increases digitization costs.

3. **Capturing text and images**
   
   **Issue:** An efficient procedure is required to minimise scanning or photography costs, but different settings may be necessary for text and images.
   
   **Policy:** If a digital camera is used then a single photograph of each page will sometimes suffice. The resulting image file can be interpreted by specialist software (such as Abbyy FineReader) to create machine-readable text, with any pictures being ignored. The same image file can be cropped by image editing software (such as Paint Shop Pro) to remove text. However, better image quality is possible by taking a close-up of the image only. If a scanner is used then two passes are usually required; one to capture the image and one to capture and interpret the text.

4. **Object naming**
   
   **Issue:** A coherent system is required for managing and publishing image files, documents and Web pages.
   
   **Policy:** For each eBook a convention for file naming is defined and applied consistently to all component files, for ease of identification, cross-referencing, and generation of persistent digital object identifiers. This helps enable the automated creation of eBooks with embedded images.

**Text Management Issues**

5. **Text file format**
   
   **Issue:** Most optical character reading (OCR) software attempts to interpret formatting detail such as lists, tables, bold and italic text, superscripts etc., and to save this formatting information along with the text in a rich format such as RTF or Word. Although superficially useful, this formatting can be counter-productive, as it is prone to error, and does not directly translate to formatting via HTML markup.
   
   **Policy:** In most cases the results of OCR are saved in plain text files, with any formatting produced during OCR deliberately discarded. Structures such as lists and tables are later reproduced using styles in Word documents, from where they may be precisely converted to XHTML. The only formatting that is sometimes preserved during digitization is bold, italic and underlined text. Though rarely used in older books, this formatting can be accurately converted to Word and then XHTML markup.

6. **Proof reading**
   
   **Issue:** All OCR software is prone to error.
   
   **Policy:** All text is read and corrected by a specialist proof-reader. This is by far the most time-consuming step in the eBook creation process, but is regarded as essential for producing credible and high-quality eBooks. In order to avoid repeated handling of large and valuable books, image files of the text are sometimes printed and used as a surrogate original for checking purposes.

7. **Error correction**
   
   **Issue:** Most printed books contain some spelling or typesetting mistakes, factual errors, misleading punctuation, or other forms of error, which is possible to correct in the digital version.
   
   **Policy:** This is an issue where compromise is required between preservation and functionality. Limited error correction is considered justifiable as part of the process of producing useful machine-readable eBooks. Indispensible spelling or typographical errors are corrected, whilst apparent factual errors are reproduced unchanged. This policy is publicised along with the text. If a book includes an errata page or slip, the changes specified are applied to the digitised text, the errata retained, and a note inserted explaining that the errata are no longer applicable.

8. **Symbols and character sets**
   
   **Issue:** Many books contain symbols or foreign-language characters not found on English-language keyboards, whose inclusion can detract from text searchability.

continued on page 22