February 2005

Twenty Issues In eBook Creation

Alan Dawson

University of Strathclyde, alan.dawson@strath.ac.uk

Jake Wallis

University of Strathclyde, jake.wallis@strath.ac.uk

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

Part of the [Library and Information Science Commons](#)

Recommended Citation

Dawson, Alan and Wallis, Jake (2005) "Twenty Issues In eBook Creation," *Against the Grain*: Vol. 17: Iss. 1, Article 6.

DOI: <https://doi.org/10.7771/2380-176X.4730>

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.

Twenty Issues In eBook Creation

by **Alan Dawson** (Senior Researcher/Programmer, Centre for Digital Library Research, Department of Computer and Information Sciences, Livingstone Tower, 26 Richmond Street, University of Strathclyde Glasgow G1 1XH; Phone: +44 141 548 2379) <alan.dawson@strath.ac.uk>

and **Jake Wallis** (Centre for Digital Library Research, Department of Computer and Information Sciences, University of Strathclyde, Room 12.12, Livingstone Tower, 26 Richmond Street Glasgow G1 1XQ; Phone: 0141 548 5854; Fax: 0141 548 4523) <jake.wallis@strath.ac.uk> <http://cdlr.strath.ac.uk/>

Abstract

Discussion of eBooks in libraries and universities covers a range of issues such as selection, purchasing, licensing, management and user support. One topic that rarely arises is that of eBook creation. Most librarians see themselves as consumers rather than creators of eBooks. Yet libraries have a collective treasure trove of books and other materials that may be suitable for digitization. This article summarises numerous issues that arise when creating eBooks and publishing them in open-access XHTML format on the Web, and describes policies that can help resolve these issues.

Background

Models of commercial eBook distribution cause problems for many potential purchasers. One reason for the complexity of eBooks as products is the extent to which they unsettle traditional certainties within the publishing and library worlds about the nature of the book (Dillon, 2001). Commercial distributors and technology companies jostle to get a stranglehold on the market by developing proprietary delivery formats. The publishing industry's approach to online distribution focuses on controlling access, aiming to ensure that profit compares favourably with investment. The attitudes of libraries inversely reflect those of publishers, with concerns over restrictive access models and value for money, while sharing publishers' uncertainty regarding the level of demand. The intensely commercial environment in which eBooks are provided to libraries can distract from the treasures that many libraries already hold within their collections, and which could potentially be converted to eBooks.

A literature review of the issues associated with eBooks provides reams of material on licensing, digital rights management, commercial eBook publishing and eBook usability, but very little on the process of eBook creation. Similarly, there is no shortage of published guidance about digitization (e.g., Morrison, Popham & Wikander, 2000; Liu, 2004). Although this guidance mentions relevant protocols and standards for metadata, archiving, interoperability etc., it fails to give the practical detail required by those wishing to create accessible eBooks.

As the information environment continues to evolve, the trend is for users to focus on ease of retrieval, and to search for aggregated mate-

rial from diverse sources, freely available in digital format. Electronic publication can add value to resources (Lonsdale & Armstrong, 2000), however access controls and proprietary formats can inhibit the uptake of eBooks and their embedding in virtual learning environments. An over-emphasis on the printed book as metaphor in constructing eBooks (Diaz, 2003), and on traditional publishing mechanisms in their delivery, continues to limit the extent of their functionality and benefits to users.

Given the complexity of the commercial information environment, there is a refreshing simplicity to the mission statement of **Project Gutenberg** (<http://www.gutenberg.org/>) — "To encourage the creation and distribution of eBooks." **Project Gutenberg** is a collection of more than 13,000 classic works of literature, freely available in digital form and produced by volunteers. Its model and motivation for creation and distribution is quite different from those of commercial publishers.

An initiative based at the **University of Strathclyde's Centre for Digital Library Research (CDLR)** is developing methods for producing eBooks more akin

to **Project Gutenberg** than commercial models, whilst using its expertise to add value in ways that a voluntary collaboration such as **Project Gutenberg** cannot. This

eBook project is helping to prevent the marginalisation, and assist in the preservation, of valuable material available in **Glasgow's** library collections, and to publish them via the **Glasgow Digital Library** (<http://gdl.cdlr.strath.ac.uk/>). Despite difficulties of sustainability, the initiative manages to side-step many of the problems related to the commercial distribution of eBooks by making content freely available over the Web. However, this creates a new set of problems that need to be addressed.

eBook Formats

One good reason why libraries are wary of eBooks is the variety of proprietary formats used. This makes their management and support difficult, and is alien to the principles of universal access that underpin the Web. It is therefore not surprising that librarians may welcome eBooks provided in portable document format (PDF). Although this is also proprietary, it is widely used for ejournals and other forms of online publishing, so will be familiar to most Web users. PDF preserves all formatting, and

is useful for printing, so some may regard it as an ideal format for eBook publishing. However, PDF has many disadvantages too, which are often overlooked but worth summarising:

- its creation requires commercial software
- its use requires installation of a browser plug-in

- it is not compliant with current international accessibility standards (W3C, 1999/2004)

- access is slower than standard Web pages
- scrolling is uneven and irritating
- arrow keys and navigation keys (PageUp, PageDown, Home, End) may not work
- two-column format is awkward to read on screen

- searching is often ineffective or not available

- embedded links and cross-references are not available

- copying and pasting may not be possible
- individual sections can not be easily located and repurposed

- metadata can not be easily embedded and retrieved (Senserini et al, 2004)

- index entries are of limited value

- it does not facilitate the different perspectives offered by the non-linear narrative of hypertext (Ferris, 2002)

PDF is a convenient means of digitally distributing documents designed to be printed, but has few advantages over other proprietary eBook or document formats such as **Microsoft Reader** or **Microsoft Word**.

The obvious question that arises is why PDF is so widely used for eBooks and other documents in preference to the basic Web document format — HTML — which has none of these drawbacks. For commercial publishers, the restrictions inherent in PDF are appealing, allowing them to retain tight control over the appearance and re-use of their products. However, for non-commercial publications, PDF may be chosen simply because it is perceived to be easier to produce. While this is arguable, the fact is that use of HTML introduces numerous issues that do not arise with PDF. Rather than deal with these issues, in order to provide a more flexible and usable format, many content providers default to offering PDF despite its disadvantages for users.

The remainder of this article summarises some issues that have been encountered in producing eBooks in XHTML format (XML-compliant HTML), and the policies adopted by **CDLR** to resolve them. The aim is not to be

continued on page 20

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

against the grain people profile

Alan Dawson

Senior Researcher/Programmer
Centre for Digital Library Research
Department of Computer and Information Sciences
Livingstone Tower, 26 Richmond Street
University of Strathclyde Glasgow G1 1XH
Phone: +44 141 548 2379 <alan.dawson@strath.ac.uk>

BORN & LIVED: Born Liverpool, 1952. Lived in Wirral, Southampton, Lancaster, Bath, Leicester, Liverpool, Glasgow, Perthshire etc.

EARLY LIFE: Ruined at age 9 by being made to jump a year at school.

FAMILY: Single, no children.

EDUCATION: Grammar school for boys, B.Sc Psychology, teacher training.

FIRST JOB: Trainee accountant.

PROFESSIONAL CAREER AND ACTIVITIES: More a haphazard series of jobs than a career, mostly in universities.

IN MY SPARE TIME I LIKE TO: Walk, climb hills and listen to music, sometimes simultaneously.

FAVORITE BOOKS: HHGTTG, How to be idle.

PET PEEVES/WHAT MAKES ME MAD: The list would be longer than the article.

PHILOSOPHY: I take my responsibilities seriously, so I try not to have any.

MOST MEANINGFUL CAREER ACHIEVEMENT: Getting out of bed in the morning, otherwise none.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: Freedom from the requirement to write articles.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: I have no idea. If I had had any foresight I would have capitalised on my Internet expertise in the early 1990s.

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 pro-

cess, 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 it 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. Indisputable spelling or typographical errors are corrected, while 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

Policy: Where characters in ordinary words can easily be represented by keyboard characters, e.g., the diphthong æ can easily be typed as ae, then the plain text version is used. However, in place names, personal names and other proper nouns, the original form of the word is retained and represented by using HTML entities. Although this makes the names difficult to search for, this is usually compensated for by including the names in an index.

9. Typefaces and typographic conventions

Issue: Books may use typefaces that are not available to Web users. Many older books use typesetting conventions that do not translate well to online publication, e.g., upper-case headings, full stops after headings, and open quotation marks at the start of every line within long quotations.

Policy: Any use of bold, italic or underlined text is retained in the eBook. However, the font used for printing is not retained. External stylesheets are used to specify fonts currently regarded as suitable for online usage, such as Trebuchet or Verdana. Typographic conventions that were common in older books are regarded as artefacts of publishing that do not need to be preserved in eBooks that use more modern formatting conventions. It is therefore considered acceptable to change the case of headings, punctuation in headings, and the appearance of quotations, provided the text itself is unchanged.

Image Management Issues

10. Image file format

Issue: A compromise is required between image quality, file size and usability.

Policy: JPEG format is considered adequate for eBook publication, but if digital preservation is important as well as online publication, images are archived as lossless TIFF files and converted to JPEG for publication. A resolution of 300 dots per inch is often used for preservation, but lower resolutions of 150dpi or 200dpi are adequate for some types of image, e.g., line drawings.

11. Image captions

Issue: Images in books may or may not have captions.

Policy: Any existing captions are faithfully reproduced, though not necessarily in exactly the same position in relation to the image, and are used as the Alt text for tags. If no printed caption exists, one is created for use as Alt tag text, to meet accessibility criteria, but is not displayed on the published page. Lists of captions may be used to create a linked list of illustrations.

12. Image positioning

Issue: Images in books may be printed on a different page or section to the referring text.

Policy: Images can be moved. It makes more sense to include an image in the appropriate place in an eBook than to faithfully reproduce its original position in the book.

Publication Issues

13. Content structuring

Issue: Books may or may not be conveniently divided into chapters and chunks of a

against the grain

people profile

Centre for Digital Library Research
Department of Computer and Information Sciences
University of Strathclyde, Room 12.12, Livingstone Tower
26 Richmond Street Glasgow G1 1XQ
Phone: 0141 548 5854 Fax: 0141 548 4523
<jake.wallis@strath.ac.uk> <http://cdlr.strath.ac.uk/>

Jake Wallis

BORN & LIVED: Born in Stirling (Scotland), grew up in Belfast (Northern Ireland) and have lived in Glasgow (Scotland) for the past 14 years.

EDUCATION: University of Glasgow - MA(HONS) 1995, Postgraduate Diploma in Information Technology 1997, University of Strathclyde - MSc in Information and Library Studies 2001.

FIRST JOB: Was as a gardener.

PROFESSIONAL CAREER AND ACTIVITIES: Researcher at the Centre for Digital Library Research, University of Strathclyde. I'm on the editorial team of the journal Library Review.

IN MY SPARE TIME I LIKE TO: Listen to music, read and fly kites (not all at the same time).

FAVORITE BOOKS: *On the Road* by Jack Kerouac, *Down and Out in London and Paris* by George Orwell.

MOST MEANINGFUL CAREER ACHIEVEMENT: Working at the Centre for Digital Library Research.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: Finding ever more innovative ways of tailoring information resources to the needs of individuals.

suitable size for online access (Wilson & Landoni, 2002).

Policy: Wherever possible, the natural structure of a book (chapters, sections and paragraphs, but not pages) is used to determine the structure of the corresponding eBook. Similarly, structures such as lists, tables, notes and quotations are retained in the eBook. However, in order to produce Web pages of reasonable length, eBooks may have a finer structure than the original, with additional section or paragraph breaks inserted if necessary. If a book has a running heading, applying to one or more pages, this may be adopted to function as a section heading. New section headings may be inserted as a last resort if existing ones are misleading, and should be identified to users as a component of the eBook rather than the original book. In some books an extensive index helps compensate for inadequate section headings.

14. Contents pages

Issue: Printed contents pages may include different chapter or section headings from those used in the body of the book.

Policy: The methodology for eBook creation involves automatic generation of contents pages that link to the relevant headings used within the book. This ensures consistency and functionality, but may mean that discrepancies in the original contents pages are lost.

15. Notes

Issue: Many books have footnotes whose lo-

cation is dependent on specific pagination.

Policy: It is more important for notes to be associated with the referring text than to appear in exactly the same position, or with the same number, as in the original book. Footnotes are therefore converted to endnotes, appearing at the end of eBook sections. If necessary the symbol used to identify notes is changed to a unique character, to enable automatic linking from note to reference. Notes may also be renumbered to ensure uniqueness.

16. Cross-references

Issue: Books may include internal cross-references, to a page or image or section in the same book, or external references to other publications.

Policy: Links are a fundamental function of the Web, and eBooks should include them where appropriate so that they function properly, even if this requires making minor changes to wording of the text. References to page numbers may therefore be reworded to refer to sections in order to work effectively as links. Similarly, text references to images 'opposite' or 'on the next page' may be changed to read 'above' or 'below' etc., so that the text makes sense in an online context.

Cataloguing and Indexing Issues

17. Title tags and metadata granularity

Issue: In order for eBook sections to be

continued on page 24

Twenty Issues In eBook Creation from page 22

meaningfully retrieved via search engines, each Web page requires a title tag, and perhaps other metadata that varies from page to page yet retains its identity as part of a book.

Policy: The database-driven methodology used for eBook creation ensures that all eBook sections have customised title tags, including the book title and the section heading. This has proven extremely effective in assisting retrieval via search engines (Dawson, 2004).

18. Names and authority files

Issue: The names used to refer to people and places change over time, so that many proper names in older books differ from the forms of names in current usage. This poses problems for indexing and searching. For example, islands of the St. Kilda group currently known as Boreray, Soay and Stac Lee are referred to in one eBook as Borrera, Soa and Stack Lee (Kearnton, 1897).

Policy: The forms of names given in the original book are left unchanged in the eBook. The use of extensive indexes alleviates the problem to an extent, but further research is continuing into how to apply the concepts of established names and authority files to eBook collections.

19. Summaries

Issue: Funding was received to write summary biographies for one historical eBook (Maclehose, 1886). These summaries are useful to users but do not form part of the original text.

Policy: The summaries are included before each chapter, but displayed in a different font and style from the main text. An explanatory note is included to make clear these were not part of the original book.

20. Indexes

Issue: Book indexes that refer to page numbers do not translate well to eBooks that use sections as the basic structure. Simply reproducing existing indexes is of limited value and does not take advantage of the eBook publication medium.

Policy: The CDLR aims to preserve the "access richness" of the original texts (Diaz, 2003) by digitising and fully inter-linking their back-of-the-Book indexes, which are regarded as useful and significant components of the original works. The original index entries are retained but are enhanced by converting page number references into active links to relevant Web pages and paragraphs. Research is continuing into methods for automating this process and for creating aggregated indexes from related works. This exploration may offer significant scholarly benefit, as an aggregated index of several texts can assist fresh comparative, analytical and structural perspectives on the content, facilitating interpretations not previously available.

Further Issues

Various other issues arise during publication where decisions need to be made on pragmatic and aesthetic grounds rather than as matters of policy, for example concerning page design, navigation, use of stylesheets, retention of page numbers, and making eBooks searchable (indi-

vidually and collectively). There are also many issues concerning the automation of production methodology and the creation of indexes which have not been covered here but will be described in subsequent publications.

Principles for eBook Creation

The policies described above were not specified in advance but were defined as necessary during the production process in an attempt to balance the requirements of online usability and digital preservation. The policies are not arbitrary but are based on the following underlying principle:

The original substantive text must be retained unaltered. However, minor changes may be made to structural elements of the text that are a product of printing in book format, rather than inherent in the original work, in order to enhance the value of publication in eBook format.

Implementation of this principle leads to the following set of rules:

All text must be carefully proof-read and verified before publication.

Indisputable spelling or typographic errors may be corrected.

Text that exists purely for internal navigation, e.g., cross-references to page numbers or illustrations, may be reworded.

Additional section headings, index entries or summaries may be inserted, provided they are flagged as being part of the eBook but not the original work.

Cross-references to other publications may be converted into live links, provided the wording of the original text is retained.

Punctuation and case may be amended if necessary to enhance online publication.

Images, notes and tables may be moved to ensure they appear in the right section, or to smooth formatting issues.

The policies adopted and described above are intended to reflect these rules and principles.

Conclusions

Experience of creating usable eBooks in non-proprietary format for publication via the Glasgow Digital Library has shown that eBook creation is more complex than might be expected. The mechanics of publication can be simplified and automated by adopting an efficient methodology. While Google has started working with major libraries to carry out automated digitization of their collections on a large scale (Google, 2004), there is great scope for smaller libraries to digitise historical materials of local relevance not available elsewhere. In doing so, numerous policy decisions need to be made regardless of the production process, but these can be simplified if they are based on an underlying principle that reflects an appropriate compromise between digital preservation and usability. Creating open-access searchable eBooks does require more thought and effort than creating eBooks in PDF or image format, but the resulting publications are far more flexible and useful, are more easily retrieved via search engines, and better meet the needs and expectations of most Web users.

References

Dawson, A. 2004. Creating metadata that work for digital libraries and Google. *Library Review* 53, no. 7:347-350. <http://cdlr.strath.ac.uk/pubs/dawson/ad200402.htm> (accessed 15 December 2004).

Diaz, P. 2003. Usability of hypermedia educational eBooks. *D-Lib Magazine* 9, no. 3. <http://www.dlib.org/dlib/march03/diaz/03diaz.html> (accessed 22 November 2004).

Dillon, D. 2001. E-books: the University of Texas experience, part 2. *Library Hi Tech* 19 no. 4:350-362.

Ferris, P. F. 2002. The effects of computers on traditional writing. *Journal of Electronic Publishing* 8 no. 1. <http://www.press.umich.edu/jep/08-01/ferris.html> (accessed 22 November 2004).

Google Inc. Google checks out library books. Google press release, 14 Dec 2004. http://www.google.com/press/pressrel/print_library.html (accessed 15 December 2004).

Kearnton, R. 1898. *With nature and a camera: being the adventures and observations of a field naturalist and an animal photographer*. Univ. of Strathclyde Centre for Digital Library Research. <http://gdl.cdli.strath.ac.uk/keacam/> (accessed 14 December 2004).

Liu, Y. Q. 2004. Best practices, standards and techniques for digitizing library materials: a snapshot of library digitization practices in the USA. *Online Information Review* 28 no. 5:338-345.

Lonsdale, R.E. and C.J. Armstrong. 2000. New perspectives in electronic publishing: an investigation into the publishing of electronic scholarly monographs. *Program* 34 no. 1:29-41.

Maclehose, J. 1886. *Memoirs and portraits of one hundred Glasgow men who have died during the last thirty years and in their lives did much to make the city what it now is*. 2d ed. Univ. of Strathclyde Centre for Digital Library Research. <http://gdl.cdli.strath.ac.uk/100men/> (accessed 14 December 2004).

Morrison, A., M. Popham, and K. Wikander. 2000. *Creating and documenting electronic texts: a guide to good practice*. Arts and Humanities Data Service. <http://ota.ahds.ac.uk/documents/creating/> (accessed 13 December 2004).

Senserini, A. et al. 2004. Archiving and accessing Web pages. *D-Lib Magazine* 10 no. 11. <http://www.dlib.org/dlib/november04/hodge/11hodge.html> (accessed 14 December 2004).

W3C. Web accessibility initiative (1999-2004). <http://www.w3.org/WAI/> (accessed 14 December 2004).

Wilson, R. and M. Landoni. 2002. *Electronic textbook design guidelines*, Eboni project. <http://ebooks.strath.ac.uk/eboni/guidelines/>