Digital content standards

Standards for digital book content delivery in trade and library supply chains, and issues surrounding their use in the UK

a report prepared for

Book Industry Communication

28 March 2010

Executive Summary

This reports records the results of an exploration of the supply chain issues surrounding the emergence of standards for digital book content, at a time when e-books are showing signs of becoming an increasingly important slice of the book market across both consumer and non-consumer titles.

There are two closely-related standards for the content of digital books: ANSI/NISO Standard Z39.86, the Digital Talking Book; and the IDPF standards known collectively as .epub. Both these standard rely heavily upon existing web standards, including CSS, Unicode and a range of more-or-less well-established standard formats for images and audio. Both share the same "package" format, by which a collection of content, metadata, navigation and other components are organised in a single zip-compressed file. The Digital Talking Book format is, in effect, a subset of the .epub format, in that it only supports a single format for text, its own XML format that is specifically designed to support the production of accessible talking books containing mixtures of text and audio, whereas .epub supports both this format and XHTML. Most e-books published in .epub format are thought to use XHTML for text content.

These standards have been developed over ten or more years and can now in many respects be viewed as being fairly stable. However, given that the market for digital books is only now beginning to show signs of taking off, and cannot yet be considered to be mature, it is not surprising that some aspects of these standards remain relatively light on detail – possibly because there has until now been little clarity as to what the market actually needs in these areas.

This report identifies two areas of significant weakness in the .epub standard, which should be of concern, and where action by the supply chain, encouraged and moderated by organisations such as BIC, could begin to address these weaknesses.

The first significant weakness is in the inclusion of descriptive metadata in digital books. The .epub standard relies mainly upon Dublin Core (with a nod in one or two other directions, such as MARC), and does not recognise that the supply chain is already heavily committed to ONIX for description of conventional book products and will almost certainly wish to continue to use ONIX for description of digital products. There is an obvious need to address this weakness, either by defining how ONIX descriptions can be embedded directly in digital books, or by defining standard mappings from ONIX to the modified subset of Dublin Core specified by the .epub standard, or by both of these. BIC should seriously consider taking the initiative in this area, possibly by forming a digital product metadata group.

The second significant weakness is in the inclusion of rights information in digital books. The .epub standard allows for rights information to be included, but provides practically no guidance on how this is to be done. The assumption is that such information will be purely human-readable, and there is no provision for the inclusion of machine-actionable rights information. Given the results of the BIC strategy

meeting last year, in which there was a clear interest in BIC taking a lead on the issue of rights and digital products, some initiative in this area would seem to be called for, as part of which the expression of rights information in digital books could be considered. Such an initiative would enable BIC to develop a clear picture of requirements in the UK supply chain, thereby ensuring that UK interests are fed into – and fully taken into account in – any future efforts to standardise at the international level.

Table of contents

Preface

This report has been prepared for BIC by Francis Cave at the request of Peter Kilborn.

This is my first exploration of this area in such depth, and my thoughts have evolved and possibly matured during the process. The report has been written in parallel with the exploration, which will perhaps explain any apparent progression from relatively unformed and general ideas in the early part of the report to rather better formed and more specific ideas towards the end.

The Executive Summary was written last, and it is to be hoped that this contains my most considered view of what I have learned.

Francis Cave
March 2010

1. Background

For some time the steady march of digital information and communication technology has been relentlessly challenging and eventually replacing many of the forms of paper-based information and communication that have held sway since the invention of printing. BIC has been championing the elimination of paper-based communication from the supply chain. In many ways the book industry has been quick to embrace the "new technology", having largely "gone digital" not only in the supply chain, through the adoption of e-commerce (although there is a long tail of small players who have yet to make this step), but also in the editorial and production processes. BIC has not concerned itself so much with production processes, but digital standards have nevertheless emerged in that area, driven by the desire to reap the benefits of an open and global market in high quality pre-press and printing services. PDF is as much a success story in the pre-press and print services supply chain as ONIX and EDI are in the finished product supply chain.

But the technology revolution that has helped to make the production and supply of printed books increasingly efficient is now threatening to replace paper as the medium for publication of books, as has already effectively happened for journal publishing. The main threat may still be to the non-consumer markets for academic, professional and educational books, but the growing consumer market for e-book readers is starting to create significant demand for digital books in the general trade sector as well.

The supply of e-books and other digital products presents many challenges: legal and commercial as well as technical. New business models will undoubtedly emerge and disappear again, as markets abandon the familiar territory of physical product supply for the as-yet uncharted waters of digital supply. Given all this uncertainty, what possible role can content standards play? Is there not a risk that they will quickly become obsolete? Will the major players adopt them or ignore them, hoping by doing so to lure market share with unique and attractive features, then hang on to it by locking their customers into proprietary solutions?

2. Content, the Internet and the Web

History offers some discouraging precedents, being littered with examples of good standards that fell by the wayside – consider the case of Betamax, the video cassette tape standard, for example. However, the overwhelming success of the Internet and the World Wide Web has made some key technologies so ubiquitous and relatively stable, that one can risk suggesting that certain technologies are unlikely to disappear in the turmoil of any shift from printed books to digital substitutes. Even if better technologies were found, the cost of changing the whole Internet to use them would be astronomical, so extremely unlikely to be contemplated.

The Internet and the Web have re-defined the way that we communicate. So many forms of communication and information delivery are now Internet-based, and as the

speed of broadband increases, this trend will accelerate. All electronic devices are ultimately likely to be – for some if not all of their functions – connected to the Internet, so that all forms of communication and entertainment will be deliverable over the Internet – and will increasingly be delivered that way for reasons of choice and efficiency. The key technologies of the Internet and the web are therefore likely to feature in all future delivery platforms, whether those be mobile phones, tablet PCs or big-screen televisions.

Here is a summary of the key content standards.

HTML The basic format for most web pages. There seems to be little sign of a replacement being proposed. HTML was developed in the 1990s on the foundation of the much older SGML standard, which, although it only became a published standard in 1986, was originally conceived in the late 1960s. HTML works moderately well on all devices, because the device decides how to format the text. Some control of presentation is possible using the style sheet technology **CSS**.

PDF The way that publishers have preferred to deliver content online, because it enables them to control how the content is presented. But a different PDF has to be produced for each screen format, if awkward scrolling is to be avoided while reading. Its popularity as a delivery format for digital products is diminishing.

Unicode The Internet is slowly responding to the shifting balance of its users from its origins in North America and Europe to a world in which the growth is in countries in Asia that have very different forms of written language. Unicode is enabling the Internet to evolve so that not only can content be in Chinese, Hindi or Arabic, but soon the web addresses used to get access to that content will be in Chinese, Hindi or Arabic characters too. Unicode is the technology that now makes it possible for all written languages to be communicated digitally, and increasingly not only computers but all devices will use Unicode, so that those devices are ready for consumers in emerging markets across the world.

Unicode is the default character set for all XML applications. Practically all web browsers support the use of Unicode-based fonts for web page display, and most computer operating systems now support the use of Unicode-based fonts.

JPEG, PNG, SVG Practically every digital camera in the world stores images in JPEG format. JPEG doesn't have the quality of TIFF, used in pre-press, but that is because it uses compression techniques that enable acceptable quality images to be delivered quickly across today's Internet connections. Although increases in broadband speeds will increase the demand for higher-quality images, some

compression is likely to remain desirable for optimum download speeds. Until a better form of compression is found, JPEG will remain the format of choice for photographic images on the web.

PNG – Portable Network Graphics – is used for simple bitmap drawings, logos and icons. This format is slowly supplanting the older **GIF** format, but the latter is likely to remain in common use for some years (despite patent issues).

SVG – Scalable Vector Graphics – is the web equivalent of Encapsulated Postscript (EPS), which is used widely for technical drawings, diagrams and other two-dimensional graphic objects such as mathematical and chemical equations. Support for SVG has grown very slowly among the popular web browsers, but there are no obvious alternatives. Scholarly, educational and professional publishers are the most likely to make use of this technology.

Audio and video formats

Audio and video content are the most demanding of bandwidth for delivery, and for this reason the competing demands for increasing quality and squeezing more content into each megabyte have ensured that there remains lively competition between various audio and video formats, each trying to offer the best combination of resolution and compression. The physical format war between DVD and Blueray contains echoes of the battle between Betamax and VHS. Data standards exist and have become very popular – MP3 for audio and MPEG for video being the obvious examples – but they have so far not succeeded in entirely killing off the proprietary alternatives, such as Microsoft's Windows Audio and Apple's QuickTime formats. The growth in live entertainment over the Internet, with audio and video being streamed in real-time on demand, presents technical challenges, the solutions to which show no sign of having been fully settled.

3. Digital book content standards

3.1. Digital talking book

Among the benefits to be gained from the arrival of the digital book, one that has been recognised as having particular significance is to serve the needs of visually impaired readers in new ways that were never possible with printed books. Apart from the political drivers to improve rights and opportunities for those with disabilities, providing visually impaired readers with more reading options, at an economic cost, represents a real marketing opportunity for publishers, for whom the production of large print books was never a profitable activity.

The DAISY Consortium was formed in the USA in May 1996 to promote the development of a standard for digital talking books. DAISY (Digital Accessible

Information SYstem), which started life as a Swedish proprietary format for talking books, evolved quickly between 1997 and 2002, when DAISY 3 became a US Standard: ANSI/NISO Z39.86. The RNIB played an active role in the development of the standard and remains a key member of the DAISY Consortium.

A DAISY digital talking book is a collection of audio files with, optionally, some or all of the content of the book in a series of XML files for the text and image files (if there are any), together with other files (mostly in XML) containing navigation controls, descriptive metadata, style sheets and – crucially, if text is included – control data for synchronising the text and the audio.

The structure of a DAISY digital talking book in fact uses the same container format (OCF) as .epub digital books, which is described below. If a digital talking book contains text, this is packaged as a series of XML files using an XML schema (DTBook) specially designed for this standard.

The DAISY Consortium website provides a large number of resources to support those wishing to use the standard, including copies of the current Z39.86 specifications, as well as a guide to a wide range of tools and services supporting the production of digital talking books. For example, Adobe InDesign CS4 includes a "Save As..." option for saving text in DTBook format.

Given that the standard is now eight years old, and has been pretty stable (the structure guidelines were last revised in 2008, and the last XML schema changes were made in 2005), it seems unlikely that there will be much further development, unless some significant new requirements for talking books are identified.

For a discussion of possible ways in which BIC might be able to assist publishers wishing to supply products as digital talking books, see the end of the next section on .epub.

3.2. .epub

HTML has proved a fairly adaptable way of representing content, but it is not enough on its own to provide a standard means of packaging an entire digital book. HTML cannot be used to encode illustrations, audio etc, nor does it provide any kind of DRM protection, any standard method for inclusion of rich metadata, or any standard way of organising the content and metadata components of a digital book into a single package for delivery to a digital book reading device.

In 1999 the Open eBook Forum was formed to develop standards for the emerging digital book market. The first version of the Open eBook Publication Structure standard was published in September 1999. The Open eBook Forum changed its name to the International Digital Publishing Forum (IDPF) towards the end of 2005, and between June 2006 and November 2007 published the three specifications that define the present .epub format.

The three specifications that define the .epub format are:

- OEBPS Container Format (OCF) v1.0
- Open Publication Structure (OPS) v2.0
- Open Packaging Format (OPF) v2.0

3.2.1. OEBPS Container Format

The OEBPS Container Format standard defines a file system structure for organising the contents of a digital book within a single ZIP-compressed package.

OCF is in fact a general container format, which can be used for packaging any set of related resources in a single compressed file. As we have already heard, OCF is used by DAISY and is also expected to be used in the next version of OpenDocument Format (ODF), the standard format used by OpenOffice.org, the open source rival to the Microsoft Office suite.

The content of an OCF digital book package is organised into a series of files and folders:

- A folder META-INF containing meta-information about the digital book package,
- A folder OEBPS containing the content and style sheets of the digital book
- A package description (OPF) file and a navigation control (NCX) file that describes the organisation and navigational aspects of the digltal book content,
- Optionally, one or more additional containers for alternative forms of the same content, such as PDF.

Most of what is contained in an OCF package is of no particular consequence to the supply chain, but there are some significant items. The meta-information folder may contain descriptive metadata about the publication, and it may also contain in a separate file rights information about the publication. However, most .epub files appear to place such information within the OPF file, rather than as separate files at the package level. See below for more information on what descriptive metadata and rights information can be included in the OPF file.

3.2.2. Open Publication Structure

The Open Publication Structure standard defines the rules for constructing valid assemblies of content, style sheets and embedded fonts for a digital book. It also defines the principal file formats and XML schemas that can be used for the content.

Text content must normally conform either to the XHTML schema or to the Digital Talking Book (DTBook) schema. It is also possible to include XML that conforms to other schemas, but there is no guarantee that a digital book reading device will be able to handle arbitrary XML, so in this case a "fall-back" representation in XHTML or DTBook must be provided.

The standard requires that all reading devices should be able to handle style sheets in the format specified in the standard, which is closely based upon the web style sheet language CSS. However, the standard also supports the inclusion of style

sheets in other formats. For example, .epub digital books prepared for Adobe Digital Editions make use of an XSL style sheet to define page properties (margins, columns per page, etc) that cannot be expressed in CSS.

3.2.3. Open Package Format

The Open Package Format standard defines an XML file format for the file within a .epub container package – typically 'contents.opf' – that describes the content of the ebook. A digital book description typically contains the following:

- Unique package name / identifier
- Metadata
- Manifest
- Spine
- Guide (optional)

3.2.3.1. Identification of digital books

Each digital book is supposed to have a unique name or identifier. The form of identifier is not fixed, and no particular scheme is recommended or endorsed by the standard, although examples given in the standard include 'ISBN' and 'DOI'.

3.2.3.2. Digital book metadata

The metadata format specified for the OPF file is based upon Dublin Core, which can be extended using the XHTML <meta> element.

Three metadata elements are mandatory in all OPF files:

- <title>
- <identifier>
- <language>.

Other metadata elements can be included as tabulated below. Many but not all of these elements have qualifying attributes. See the table for details.

Dublin Core metadata element	**Description**	**Qualifying attributes**
<title>	The title of the publication. Mandatory. May be repeated.	None. It is up to the reading system to determine which title to display and how.
<identifier>	A unique identifier. Mandatory. May be repeated.	An optional 'scheme' attribute may be included to specify the identification scheme used.
<language>.	A language of the publication. Mandatory. May be repeated. Must comply with IETF RFC 3066, which allows both 2-letter and 3-letter	None.

Dublin Core metadata element	Description	Qualifying attributes
	language codes from ISO 639 Parts 1 and 2 respectively.	
<creator>	A primary contributor to the publication. Non-mandatory. May be repeated. Publications with multiple authors should include multiple <creator> elements.	An optional 'role' attribute may be used to define the role of the creator, using 3-letter MARC relator codes (e.g. author='aut'). An optional 'file-as' attribute may be used to define a normalised form of the creator's name for machine processing (e.g. indexing) purposes
<subject>	A subject or keyword descriptor for the publication. Non-mandatory. May be repeated	None. The standard does not attempt to provide any method for specifying the subject scheme.
<description>	A description of the publication's content. Non-mandatory. May be repeated.	None. The standard does not attempt to provide any method of qualifying descriptions.
<publisher>	The name of the publisher as defined by the Dublin Core Metadata Element Set. Non-mandatory. May be repeated.	None. There is, for example, no mechanism for specifying an imprint as distinct from the publisher, nor for specifying a co-publisher.
<contributor>	A secondary contributor to the publication. Non-mandatory. May be repeated.	Same as for <creator> – see above.
<date>	Date of publication, unless qualified. Non-mandatory. May be repeated. Dates may be expressed in various formats, but follows the format restrictions defined by W3C, which requires the use of punctuation between the elements of a date, time or date-time.	An optional 'event' attribute can be used to qualify the date as being the date of 'creation', 'publication' or 'modification'.
<type>	A code value specifying the category, function, genre or "aggregation level" of the content of a publication. Non-mandatory. May be repeated.	None. The standard suggests that "best practice is to select a value from a controlled vocabulary", but there is no mechanism for specifying the controlled vocabulary.
<format>	The media type or dimensions of the resource. Normally used to specify the digital book format as a MIME type. Non-mandatory. May be repeated.	None.
<source>	Used to specify a resource (e.g. another publication or possibly a work) from which the digital book has been derived. Non-mandatory. May be repeated.	None.
<relation>	Used to identify an "auxiliary resource"	None.

Dublin Core metadata element	Description	Qualifying attributes
	and its relation to the publication. Non-mandatory. May be repeated. There is no information on how this element is to be used.	
<coverage>	Used to contain information about the spatial, temporal or other coverage information. It is recommended to follow Dublin Core Metadata Element Set guidelines. Non-mandatory. May be repeated.	None.
<rights>.	A statement about rights, or reference to one. It is recommended that a copyright notice should be directly included here. Non-mandatory. May be repeated.	None.

The <meta> element can be used to embed additional metadata, using the name-content pairs defined in the XHTML standard, e.g.

```
<meta name="prize" content="2010 e-Booker short list"/>
```

3.2.3.3. Manifest

This section of the OPF file lists all the content components and style sheets, including for each component its path and filename within the overall package and its MIME type. Embedded fonts are not specified in the manifest, but may be referred to directly from within style sheet files.

3.2.3.4. Spine

This section of the OPF file lists the items to appear in the digital book's human-readable table of contents, in the order in which they are to be presented.

This section of the OPF file is not to be confused with the navigation control file – typically 'toc.ncx' – which specifies how the digital book is to be "played" by the reading device and will typically contain information about components that are not explicitly listed in the human-readable table of contents.

3.3. Supply chain issues raised by digital book content standards

There is currently much discussion in the trade about not only how to identify and to describe digital books in supply chain communications. There has been little discussion about the technical formats used for digital books, on the basis that these have hitherto been determined by the reading devices. The emergence of standards for the content of digital books means that it becomes possible to have a meaningful

discussion of the key features of these standards and whether they really meet the needs of publishers and the supply chain as a whole, as well as readers.

As indicated in the preceding sections, there is a growing convergence on the use of the .epub standard for formatting digital books. This standard sensibly builds upon existing standards where possible (e.g. HTML, CSS), even if in some respects the choices are not entirely in line with existing book supply chain conventions (e.g. the use of Dublin Core for digital book metadata).

It is in the areas where the .epub standard is least well-defined that it may well be in the interests of the supply chain to fill in some gaps.

3.3.1. Descriptive metadata

The first and most obvious gap is in the inclusion of metadata in a digital book. It should be possible for a publisher to embed whatever metadata they wish in a digital book, so that a full description is available to anyone who wishes to read it.

The .epub standard defines two possible locations for descriptive metadata:

- As a separate XML file 'metadata.xml' within the container package
- As a metadata section within the OPF contents description file 'contents.opf'.

In the latter case the .epub standard specifies that the metadata must be Dublin Core metadata, which has significant limitations. The separate file at the container level does not have these limitations and could, for example, contain a complete ONIX record for the digital book.

3.3.2. Rights information

The .epub standard does not attempt to standardise the way in which rights are expressed within a digital book.

The .epub standard defines two possible locations for rights information:

- As a separate XML file 'rights.xml' within the container package
- As a rights element within the metadata section within the OPF contents description file 'contents.opf'.

In neither case does the standard do much, if anything, to state how rights are to be expressed, beyond a vague recommendation that a copyright statement should be included directly in the metadata section of the OPF file and not indirectly (by reference to an external resource).

3.3.3. Possible actions

This suggests that there are possibly two areas in which the supply chain could collaborate to develop "refinements" – possibly in the form of local profiles – of the

.epub standard, to more precisely meet the needs of the supply chain in the UK (or ultimately more widely).

A digital book metadata group could be established to develop best practice guidelines for the inclusion of descriptive metadata in digital books. Such a group could have several distinct objectives:

- to specify how a complete ONIX record for a digital book can be included in a .epub digital book;
- to specify extended rules for embedding descriptive metadata in OPF files using the Dublin Core elements defined in the .epub standard;
- to specify a consistent method for mapping a subset of ONIX record metadata to OPF Dublin Core metadata elements;

The same group, or a separate group, could be tasked with developing best practice guidelines for the inclusion of rights information in digital books.

Other possible areas for collaboration might include the development of best practice guidelines for file-naming conventions for use in .epub packages. But this would be of less obvious value in the supply chain for finished digital book products than it would in the up-stream manufacturing processes. However, it may be worth some follow-up research to determine whether there are other areas of uncertainty about how to apply the .epub standard, which publishers in particular might be willing to address by some kind of collective action.

Such activities as the above could feed the development of an international consensus on some of these issues, and would presumably in due course lead to revisions of the .epub standard itself.