APPENDIX A. *Electronic Book Standards*

Anna Gold, February 2, 2001

OUTLINE:

Scope
Background
Current Status:
    Identifiers
    Metadata
    Document Structure
    File Formats
    Reading Devices – Hardware
    Reading Devices – Software
Discussion: CDL Principles

Scope: The scope of this section includes technical (hardware and software) protocols and standards for the presentation of electronic books, *apart from the management of digital rights.* Included in this topic are standards and protocols which determine the interoperability of electronic books, their presentation on reading devices, embedded textual mark-up, management of color and multimedia elements in electronic books, printing and downloading. Applicable existing and developing standards include:

- Identifiers, e.g. ISBN, ISSN, Document Object Identifiers (DOI)
- Metadata, i.e., exterior to the text
- Document structure, i.e. Document Type Definition (DTD) and embedded structure such as Extensible Mark-up Language (XML)
- File format, e.g. standards for encoding text or graphics
- Reading devices (software)
- Reading devices (hardware)

Background: Monographic electronic texts have been available on the Internet since the early years of UNIX file exchange, file transfer (FTP), gopher, and finally hypertext transfer (HTTP) protocols. The original purpose of all of these protocols was the exchange and delivery of files principally in the context of noncommercial scholarly and technical communication. These protocols are the standards that have, until recently, shaped the production and distribution of most electronic texts. Only recently have electronic texts been packaged and offered for distribution and sale as electronic books. Thus the divergence of electronic books from the protocols and history of electronic texts generally, is both caused and shaped by the motivation to shift the distribution of electronic texts out of this history and into a framework of commercial production, sale and public distribution.

Until the advent of commercial electronic books, electronic text publishing and distribution have relied on existing protocols of platform-specific and later platform-independent web browsing. These same protocols still undergird significant commercial electronic book publishing ventures, such as netLibrary. Early, and in a sense relatively primitive commercial electronic book publishing systems such as netLibrary have begged some of the standards questions that will need to be resolved by an electronic books industry. For example, netLibrary relies on URLs as identifiers; proprietary web-accessible databases for metadata; and local mark-up and file management protocols for structuring electronic texts. The presentation and use of electronic texts, including interactive elements, until the emergence of a distinctive electronic *books* industry, have been based on the capabilities of common personal workstation hardware equipped with freely-available web browsers and browser plug-ins.

Current Status: The emerging market for electronic books is driven by the rapid expansion of access to personal computers and reliable bandwidth, and the growing market for portable, wireless networked devices. These factors in turn have increased both the viability and the demand for web- or PC-independent reading systems, and pushed the envelope with respect to options for distribution, marketing, and proprietary protocols.

Browser-based electronic text-reading systems are constrained by the capabilities of browsers and personal computers, which is to say they are best suited to end-user delivery of brief texts that can easily be printed out. But portable reading systems for electronic books enable a shift away from the "locate-view-print-read" model of use, toward "identify-acquire-read-incorporate" behavior. As the end-user has become available as a primary market for electronic books, sellers have been faced with the absence of establish standards that will, on the one hand support commercial end-user distribution, and on the other hand, will enable added value for the consumer.

The industry thus has a strong motivation to rapidly develop new standards that will support the development of robust marketing and distribution systems for electronic texts. Interoperability of texts is not at the top of their list of desiderata, even though the publishing industry avows that they share an "ultimate vision of interoperability between formats," that would enable "translation services or the 'universal reader' [to]…enable ebook content to 'move' between software / dev
for Ebooks, AAP, 2000, p. 26).

That said, one view (the American Association of Publishers) is that the formatting of e-book content is an area that is currently "proprietary," and states that "there can be no effective standar
area (Ibid., p. 33).

Another view is presented by representatives of the e-book industry who propose to pursue this vision by developing what they call an "Open e-Book Publication Structure, " the stated purpose of which is to ensure that content can be viewed on any reading system which is OeB compliant. It is intended to allow publishers to provide their content without having to reformat it for each reading system, though it does not refer directly to reader hardware or software. For example, the proposed OeB standard supports capabilities for "flowing" text and text with variable and complex formatting.

The "interoperability" (that is, the non-proprietary usability) of printed texts is a foundation for providing library services. *Libraries cannot assume that the priority we place on interoperability of electronic books will be shared by electronic book publishers and distributors*. Rather, the highest priority in the area of standards development for the publishing industry is clearly that of digital rights management or DRM, with a close second going to developing standards for metadata and identifiers – but only insofar as these will support the development of retail markets and the delivery of products to end-users.

What follows below is an outline of the standards proposed to date by the electronic books industry, in the categories of interest: identifiers, metadata, document structure, file formats, and reading devices (hardware and software).

Identifiers:

The publishing industry understands that Document Object Identifiers (DOI) can be applied effectively in the realm of electronic texts. Guidelines drafted by the AAP have suggested that electronic monographs require adopting the DOI as a new numbering standard, and use it conjunction with the ISBN and other established product codes for legacy purposes. Their proposal is that a single ISBN and DOI should cover "all renderings of an ebook," and that various "renderings" (such as different file formats) be

managed through the publishing industry's ONIX metadata. The AAP envisions that DOIs might be *saleable* pieces of ebooks by extending the ebook DOI via the use of nodes" (Numbering Standards for Ebooks, AAP, 2000, p. 7, emphasis added). It is envisioned that the uniquely identified components or pieces of an ebook might include book parts, chapters, sections, sub-sections, figures, tables, etc. (Ibid., p. 33).

Other identifiers recommended by the AAP are handled as metadata recommendations, e.g. version control (a requirement "put forth by representatives from the Education Market" to handle textual corrections and changes below the threshold of editions; and software release numbers for required software to read the text; or the as-yet unsupported association between an ebook and an abstract work, or "WorkID" (Metadata Standards for Ebooks, p. 23).

Metadata:

AAP: The AAP has issued a comprehensive statement of proposed metadata standards (Metadata Standards for Ebooks, 2000) with a comment period ending December 31, 2000. As noted above, these standards focus on extending the book industry's ONIX standard used to support book sale and distribution. Notable in this document is the suggestion that "electronic content may be sectioned and distributed / sold as individual 'chunks' of content," and the corollary need for metadata to identify and describe such "chunks" (Ibid., p. 6, see also pp. 8-9). They further analyze the need for metadata into:

- "discovery" metadata intended for public consumption in locating and purchasing titles;
- "core" metadata also intended for public consumption, consisting of data that will enable certain user functions such as cataloging and file management; and
- "private" metadata that would not be available to the public, but would be used to support the book selling process by specifying rights, format, and "return" metadata from consumers (data provided by the consumer in exchange for services or incentives). The AAP recommends *against* standardizing this private metadata with the exception of DRM standards once they are developed.

The AAP's proposed metadata standards are worth studying particularly as they relate to the "composite" and component-based products, customized to individual needs, course-packs, and the like; and as an alternative set of metadata that may or may not be exchanged with libraries for the purpose of creating enriched catalog records and points of access to electronic books.

OeB: Unlike the AAP proposed standard, the OeB metadata proposals reference several standards developed by or with library input, including Dublin Core and the US MARC relator code list. Dublin Core metadata is to be expressed in XML, with an "x-metadata" element, also expressed in XML, to allow content providers to express arbitrary metadata beyond that allowed by Dublin Core (OeB Publication Structure, pp. 12-18).

Document Structure:

AAP: The AAP proposed standards do not address document structure per se, except to the extent that it allows for the identification of saleable document components through nested identifiers and discovery metadata. Consequently, the AAP standards do not push for or advocate the interoperability of electronic books in the same way that the OeB standards do.

OeB: As currently written, the OeB standard calls for the following levels of document structure: the lowest level is the "document;" next is the "publication", consisting of a collection of documents and

other files; and at the highest level is the "package," consisting of the metadata associated with and modifying the "publication." The "package" consists of the following elements:

- unique identifier ("package identity");
- metadata (DC and extended);
- list of files in publication ("manifest"): consists of "items" each with "id" and specified MIME type and "fallback" types;
- primary linear order of (relevant) files in publication ("spine"): only one spine per publication; items from manifest are ordered via "itemref" tags that refer to item id's;
- alternate sequences of files ("tours"): optional; identified by "site" titles; and
- description of structural components of publication ("guide"). Components are denoted as "reference" elements, uses "types" from Chicago Manual of Style (cover, title-page, toc, index, glossary, etc.)

More internal structure can also be added to the publication or document via XML.

File Formats:

The issue of electronic book file formats involves issues of hardware and software interoperability, accessibility (ADA compliance), and functionality.

XML: The OeB proposal calls for a nonproprietary, open XML-based format. They state in their FAQ that they foresee the development of tools to convert proprietary formats (PDF, Quark, Pagemaker, etc.) into OeB's interoperable XML format. The OeB specification further requires that conforming reading systems support, minimally, the following file formats: XML, CSS, JPEG, and PNG. The specification allows other kinds of files (e.g., a QuickTime movie), but requires that a "fallback" version be provided for such files in one of the other formats (e.g. JPEG).

The OeB XML format has been implemented by Questia, and _____ and is supported by the Microsoft Reader software.

HTML : HTML has been the open file format of choice for many electronic text projects and for at least one important commercial e-book venture, netLibrary. The format permits hyperlinking, variable font displays, embedded images and other embedded files. In addition to self-contained book-length texts, HTML has been used for linked collections of texts, and for producing extended and contextualized core texts (e.g. Mendelweb).

PDF : PDF is used widely to create a largely tamper-proof replica of a printed page. Software is available that easily converts scanned images and word processing files to PDF, and free web browser plug-ins make it a quick and cheap way of publishing documents on the web. However, unlike HTML or XML, PDF is a proprietary standard, owned by Adobe, who has partnered with Glassbook to offer e-book software designed to use PDF.

The OeB stance toward PDF is that Open eBook reading systems are not required to support PDF. The OeB FAQ states that "an alternate representation of the content" must be provided for any embedded PDF file, to be used by reading systems that lack PDF support. However, like many others in the industry, the OeB standard anticipates that commercial PDF-to-OeB conversion tools will become available.

Plain Text (ASCII):

For reading on virtually any device, some believe that plain (ASCII) text has no rival; others consider it a lowest common denominator that is dysfunctional both for scholarly works and for texts that require non-American alphabets.

An illustration of the argument for plain text is provided by Project Gutenberg:

> "The Project Gutenberg Etexts should so easily used that no one should ever have to care about how to use, read, quote and search them ... This has created a need to present these Project Gutenberg Etexts in "Plain Vanilla ASCII" as we have come to call it over the years. The reason for this is simple. . .it is the only text mode that is easy on both the eyes and the computer. However, this encourages others to improve our etexts in a variety of ways and to distribute them in a variety of the available media….Once an etext is created in Plain Vanilla ASCII, it is the foundation for as many editions as anyone could hope to do in the future. Anyone desiring an etext edition matching, or not matching, a particular paper edition can readily do the changes they like without having to prepare that whole book again. They can use the Project Gutenberg Etext as a foundation, and then build in any direction they like."
> (http://promo.net/pg/history.html#thepgphil-2)

Plain text does not permit any significant text formatting (whether for substantive or aesthetic purposes), reading, or "packaging" of texts other than in a serial linear way. On the other hand, ASCII makes simple text searches possible that would not be for users of image-only or non-digital versions of the same content. (Some e-book publication efforts combine plain text (often created with OCR) to support textual searching, with PDF or graphical images to provide a facsimile of a printed page.)

Some hand-held devices such as the Palm Pilot are limited to viewing plain text or simple HTML. In response to this, the Scholarly Technology Group at Brown University has developed a means of autoconverting XML data for use on palmtops:
http://mama.stg.brown.edu/projects/indexcard/displaycard.php3?card=40)

Reading Devices: Hardware

Until recently, electronic books, like electronic serials, have been made available on multi-purpose workstations running web browser clients. Although it is theoretically possible with this technology to read offline, the majority of commercial electronic journal reading is done either while connected to the provider site, or the item is printed out and read on paper offline. The relative length of electronic monographs defies this model, and encourages the development of devices to replicate some of the virtues of printed monographs, including portability and network-independence. And, as Eamonn Neylon writes (D-Lib, January 2001), with e-books "the commodity that is being considered is finite."

Multi-purpose devices:

PC workstations and laptops can both be used as e-book readers either by using web browsers, PDF viewers, or e-book reader software. When e-books can be downloaded, in theory a laptop can become a portable reading device even when not connected to the Internet. Wireless networking may make this consideration less significant. Palm and Pocket PC can be used in a similar way, and software has been written specifically to enable reading e-books on PDA's (Peanut Reader).

Dedicated devices:

A few portable dedicated reading devices are available, with different advantages and disadvantages.

RCA: The RCA REB eBook is a dedicated device available in monochrome or color displays, with prices currently ranging from $299 to $699. The display is "book-size" and high quality, and has a built-in 33.6 KB modem to facilitate downloading. It is about the same size and weight as a hardcover book. The NYT reports that "once a straightforward if lengthy setup is completed, buying books becomes -book software from Gemstar-TV Guide International (see below).

Frankline eBook Man: This device is still in prototype and does not have a downloading system. The company expects to use the Reader for reference books that will be downloadable from the web, and will use Microsoft Reader. Price ranges from $129-$229; it has a large screen and is lighter weight than most organizers. The lack of backlighting means there is low contrast in the display.

Reading Devices: Software

Many electronic books are available for reading and downloading via ftp and web clients. While this ensures interoperability, the software functionality is very limited: for example, within-text and cross-text search options are limited; variations in type size are possible, but type face options are limited; and user annotation is not possible. Electronic book software promises to give readers new capabilities, such as adding bookmarks, margin notes, highlighting, powerful searching, and improved font and page display. E-book reading software may be device-dependent or (more-or-less) device-independent.

Microsoft Reader: This is available as a free download. It uses "ClearType" display technology for improved readability. It allows the user to download an entire text into the reader hardware for portable viewing. The average size of a Microsoft Reader e-book file ranges from 200 KB to 600 KB; a 300-page novel is approximately 250KB. Reader capabilities include adding bookmarks, margin notes, and search. It is available for Windows-based PC's and a version ships preinstalled with Pocket PC devices. It is also slated to be available on the Franklin Electronic Publisher's eBookman device. They have no plans to make the Reader available on any other platforms (Linux, Macintosh, Palm). The Reader requires that Internet Explorer 4.01 or later be installed on the user's PC.

There are conversion tools available for HTML and Microsoft Word documents and Microsoft claims that others will become available. They also claim that various media file formats are currently supported ("text, images, and audio") and that they will support streaming video and other media formats in future releases. There is a free Microsoft Reader add-in for Microsoft Word that allows you to convert a Word document to the Reader format. The reader does not support PDF, and Microsoft argues that their dynamic layout capability is superior to PDF.

Major Microsoft Reader e-book file retailers include: Barnes&Noble.com (currently including Cliff's Notes, popular fiction, Harvard Business Review titles, and a selection of "e-book originals" – titles only available as e-books. Here is a user response to the e-book original, Orpheus Emerged, by Jack Kerouac.

> "Orpheus Emerged isn't like the other eBooks on Barnes & Noble's virtual shelves. In LiveReads' innovative hands, it's become part literature, part conceptual art, part educational software. Hip graphics, colors, and typefaces evoke the cigarette-smoke-and-cheap-wine-tinged atmosphere of the New York intellectual underground of the 1940s (or at least what we now think of as that atmosphere). Informative hyperlinks illuminate Kerouac's endless literary allusions and point to relevant passages his contemporary journals. The book includes an introduction by Robert Creeley, essays about Kerouac and the Beat movement, timelines and bibliographies, and even excerpts from an audio version of the book and video clips from The Source, a documentary film

about the Beats. Now this is what eBooks were supposed to be." -- Wade Roush - eBookNet

Other non-commercial publishers using the Microsoft Reader software include the NISEE project at the Earthquake Engineering Library, and the University of Virginia.

A NYT writer notes that the download "ritual" for Microsoft Reader and for downloading related e-book files is cumbersome.

Glassbook: Glassbook is also available as a free software download, available for PC's only. It is designed to display Adobe's PDF file format. It offers capabilities that include display rotation for laptops (to provide the "landscape" experience of book reading); search; bookmarking, zoom function; highlighting and annotation; two-page side by side viewing; sharpen text; read aloud to permit hearing spoken text; and copy and print from clipboard and print eBook pages (these last two capabilities are subject to publisher's permission).

Gemstar: Designed to be read on RCA brand devices by Thomson Multimedia. Exclusive distributor of devices for their format. Patented.

Mobipocket: The Mobipocket PDA Reader format is intended to be readable on all PDA platforms, including Palm OS, Windows CE/Pocket PC, Psion, Epoc R5 and R6, and Franklin eBookman. It claims to be OeB compliant. Software is available for publishers to convert HTML and text files into the Mobipocket format (cost $149-999, with the $999 "Professional" version providing "full security capability). The Mobipocket Reader is available as a free download.

Peanut Reader: Available as a free download, it works on any personal portable organizer running Palm Computing's operating system (manufactured by Palm, Handspring, or Sony). Font displays are crude, and displays work best with a backlighted color screen.

Discussion, Relation to CDL Principles

1. Importance of librarian and W3C involvement in developing standards, esp. in areas of identifiers, metadata, and interoperability of files and structures. Hasn't been a problem yet for ejournals, but definitely is a problem for ebooks.
2. Potential role of university publishers, including UC Press, in promoting interoperable standards. Microsoft Reader currently being used by some university e-book publishing ventures.
3. Impact of market for reading devices and software on universities, e.g. for course readers. The Wilensky proposal / idea of a portable personal library for every graduating student.
4. Impact of disaggregation of the printed book package on cataloging, lending, etc. Will the printed book continue to serve as the structural model for e-books?
5. Born-ebooks: when "e" is not a choice: implications for how libraries can provide access when there is no printed version (e-only book publication).

APPENDIX B: Digital Rights Management Systems

Karen Coyle, 3/12/01

Digital Rights Management systems are technologies (either hardware, software or a combination) that enforce controls over intellectual property. These can be controls over the copying of works, or on their access and use.

RIGHTS

Copyright law confers a limited set of rights on copyright holders, mainly those that involve the production of copies of works. (There are other rights, such as performance rights for music. These are less relevant to our relationship with e-books, however, and aren't included in this discussion.) All other rights not explicitly given to copyright holders belong to the public.

The Digital Rights Management (DRM) systems that I have seen take precisely the opposite approach. DRMS define the "rights" of users, and any rights not explicitly allowed by the technology are not conferred. That is, the default is that nothing is permitted in terms of copying, access or use of the intellectual property. The DRM system then defines and allows any exceptions to that default.

Note that although the allowed activities are commonly called *rights*, these are not rights in the sense of "something to which one has a just claim" or "an entitlement based on law or custom." Instead, these are contractual agreements and probably should not be called "rights."

USE

In general, copyright law governs only the copying of works. DRM systems, on the other hand, may exercise controls over the access to or use of works. These can limit access by user or by time ("for 2 hours"; "from 3-1-00 to 3-1-01"; "weekends only"). They can also limit the access or actions to particular devices, i.e. limit printing to a class of printers that can retain watermarking. Actions or access can be associated with payments as well, so that printing, or even reading, could have per-page or per-time charges.

COPYING

Some DRM systems allow limited printing or copying from the works, often referred to as similar making fair use copies. Ironically, many DRM systems are not able to prevent making copies of the entire digital "container" for the content. Once delivered to a general-purpose computer all files can be copied. However, if the copied files are essentially disabled for access, as they are under some DRM systems, then the copies are considered insignificant from the publisher's point of view. The use of dedicated reading devices (such as the Rocketbook Reader, now RCA 1100) gives publishers control over copying because these devices can be designed to communicate only with the publisher's e-book sales site and the files are only usable on those devices.

FIRST SALE

The first sale doctrine states that ...

> "... the owner of a particular copy or phonorecord lawfully made under this title, or any person authorized by such owner, is entitled, without the authority of the copyright owner, to sell or otherwise dispose of the possession of that copy or phonorecord." Title 17, 109 (a)

Many people think that first sale allows lending, but in fact it disallows any control by the copyright holder over secondary use. This makes lending possible, but it also implies much more by making a specific limitation on the rights of the copyright holder.

The authors of DRM systems often state that their technology "allows" lending and therefore is in accord with first sale. In fact, their technology violates first sale by maintaining control over secondary uses. I admit that it may be very difficult to both secure digital objects and abide by first sale, but we need to be clear when first sale is not applicable.

FAIR USE

Fair use is a very "squishy" concept and courts have been careful not to set any pre-defined limits to what copying is considered "fair." There are complex criteria to determining fairness and each case must be decided on its own merits in relation to those criteria.

Many DRM systems allow some copying and some set hard limits (one paragraph can be copied to the clipboard, one page can be printed, etc.). These hard limits cannot be considered fair use because they make none of the situational determination that goes into a fair use defense. Although the ability to do some copying is highly desirable, it should not be confused with the fair use defense included in copyright law.

SOFTWARE AND HARDWARE CONTROLS

Copyright depends on law for its enforcement. This leads to an imprecise application of control over copies and copying, but it can be argued that this imprecision is intentionally allowed by the law.

DRM systems employ technological controls that can be very precise. If users are limited to copying one paragraph per hour, it will not be possible to copy three paragraphs in two hours no matter what the circumstances. This is a fundamental difference between technology and law, and one of great debate in terms of the eventual social consequences of relying on technology in areas where we have previously used legal institutions.

Among the dangers of technological controls is that they may not function properly, that their restrictions may be hidden from users, or that users will accept unreasonable controls that are part of a complex and desirable electronic package. A technological control that fails will deny access regardless of the legal contract that the user has agreed to. And users may be unaware of the extent of controls prior to or even after making a purchase. Libraries rightly fear that some technological controls may hinder their role as information providers.

Controls being developed today often have a hardware component. For example, when electronic books are downloaded from the Internet to a personal computer, the software mechanism that allows the user to open and read the book is probably keyed to a unique CPU or hard drive identifier. If the electronic book is copied to another computer or another hard drive, the rights management mechanism will detect this and will not allow the book to be opened. These controls are particularly worrisome because of the rapid rate of hardware obsolescence in today's environment. Presumably, as a user upgrades hardware these protected files will no longer be readable. In the Glassbook Reader FAQ on the Glassbook web site, the question about moving a file from one computer to the other gets this response:

> "Currently we are working on incorporating a lend/give feature into the Glassbook Plus Reader, which will allow you to transfer a book to another PC. In the interim, if you want to read a book on a certain PC you will need to download it directly to that PC."
>
> http://www.glassbook.com/support/faq2.htm#qt6

Note also that ebook readers and their rights management systems can interact with other functions on a personal computer. Glassbook warns that no debuggers can run on a system while the Glassbook reader is open:

Can I run the Glassbook Reader and a debugger at the same time?

Our security implementation does not allow debugging of any program while the Glassbook Reader is running. However, as long as the Glassbook Reader is not running, debugging is not a problem.

http://www.glassbook.com/support/faq2.htm#qt16

DRM Systems Today

The development of DRM systems is still in its early phases and few sophisticated systems exist. Here are some of the systems in use and in development:

- Adobe Acrobat Web Buy. This system controls access to a PDF document through the use of a certificate file. Implementations can use hardware IDs to limit access to a single hardware configuration so that the PDF file and certificate cannot be copied to another machine and used there. http://www.adobe.com/products/acrobat/webbuy/main.html
- EBX. The Electronic Book Exchange (EBX) Working Group is an organization of companies, organizations, and individuals developing a standard for protecting copyright in electronic books. This standard defines a trust system that would accommodate a variety of rights languages but allow the licensing of works and control over access.
- XrML. XrML is a DRM language specification developed by Xerox and currently owned by ContentGuard, a company formed by Xerox and Microsoft. It has not yet be implemented but is being considered by some eBook standards bodies. http://www.xrml.com
- ODRL. The Open Digital Rights Language is only now being formed, but it is being considered as a project by the World Wide Web consortium and could evolve into a WWW standard for digital rights. http://odrl.net/
- ONIX. ONIX is a book industry standard for communicating product information. ONIX began as a standard for hard copy books but is moving into the eBook area and will include DRM for digital materials. http://www.editeur.org/onix.html