

AAP's Subcommittee for Books Online Briefing Paper

Copyright © 2007 AAP, Inc.

Abstract

The Association of American Publishers Digital Issues Working Group has formed a Subcommittee for Books Online. This briefing paper describes the purpose and goals of that Subcommittee.

Table of Contents

- [1. Introduction](#)
- [2. Mission statement](#)
- [3. Usage scenarios](#)
 - [3.1. Large book retailer Web sites](#)
 - [3.2. Smaller book retailer Web sites](#)
 - [3.3. Search engine indexers](#)
 - [3.4. Individual Web page developers and third-party sites](#)
- [4. Technical goals](#)
 - [4.1. Deliver open and non-proprietary standards](#)
 - [4.2. Enable multi-server user experience and publisher-managed access](#)
 - [4.3. Address the "lowest common denominator"](#)
 - [4.4. Maximally leverage existing standards](#)
 - [4.5. Enable performant implementations](#)
 - [4.6. Enable declarations of semantic type for book components](#)
 - [4.7. Enable flexible content licensing agreements](#)
 - [4.8. Design for extensibility](#)
- [A. Working Group members](#)
- [B. Production notes](#)

1. Introduction

The AAP's Subcommittee for Books Online Working Group (the "Working Group") was formed to recommend industry standards for the online discovery, browse, search, and distribution of books presented in digital format for marketing purposes. This document provides the mission statement and specific goals of the Working Group. Examples are given to illustrate how the Working Group's deliverables will address key publishing industry issues in a practical and useful fashion.

The Working Group is addressing new content management challenges created by the growth of online services distributing digital book content. Managing content at multiple large outlets each with their own rules and interfaces is already a challenge; additionally, there is currently a lost opportunity in the thousands of smaller potential outlets that do not have the will or ability to build their own proprietary book archives.

As the number of online book browsing services increases, and to promote the creation of new online browsing services, the need for a simple, unified method for publishers to distribute book content to the Web has become clear.

The Working Group's Deliverable will be "Technical Specifications" for communicating to online book browsing services ("Syndication Partners") the content available in publisher digital repositories.

The following examples are the types of problems being addressed:

- **Retailers:** Online booksellers and other sites have developed successful programs for displaying digital book content which serve their audiences quite well. But as the demand for online book content increases and these kinds of services are duplicated by many sites: questions of content management such as quality, availability, versions, and updating will need to be addressed by the content owners. A standard means of distributing and updating digital book content will allow publishers to manage content distributed to multiple places in a cost-effective way.
- **Digital book viewers:** Development of software for reading digital book content is undertaken mainly by third-parties who are capable of building and delivering their own book archives since most publishers are not able to provide the book content from their repositories directly. A simple, standard way to access publishers' content will allow new kinds of digital reading applications to be developed without requiring the overhead of building a large-scale repository.
- **Search Engines:** Because of a desire to protect their investment in their copyrighted works, most publishers have not made books available in Web-native, "lowest common denominator" media like plain text and images. But because book content is a valuable resource for Internet users, search engines have enabled book search by creating digital book repositories of their own. An industry standard for access to publishers' book content will enable search engines to continue providing search and useful ways to experience the content, and allow publishers to manage the availability and quality of the content.

1.1. Procedural Goals

The Working Group's expectations on how to realize this mission are outlined below. The process is built around the distribution of three documents: the **Briefing Paper**, the **Requirements Document** and the **Standards Specification**. AAP proposes to work with the Book Industry Study Group and others to establish a process involving all interested parties that will lead to the development and continuing maintenance of these

book industry-wide standards for the online discovery, browse, search and distribution of books presented in digital format. AAP and BISG will announce their plans for this process in April 2007.

- **Briefing Paper.** At the highest level, this non-technical document defines (a) the mission of the standards specification, (b) why a specification is needed and (c) the parameters of what should and should not be specified in the process. The Briefing Paper will have the widest audience and serve to market the efforts of the Working Group and promote the need for an agreed-upon set of standards for digital book discovery and distribution.
- **Requirements Document.** This is an introductory technical document that defines the exact use-cases deemed essential by the working group. The use-cases will detail what needs to be accomplished in the subsequent Standards Specification. While the audience for this document shall in no way be limited, the stakeholders in this document are expected to be developers and business managers directly involved in the discovery, browse, search and distribution of digital book content over the web.
- **Standards Specification.** This is a technical document that will represent the group's work to date. The documents will specify the precise, recommended interface for each use-case requirement as previously defined in the Requirements Document. The final Standards Specification, developed after coordination through the Book Industry Study Group, will be available for complete public distribution. It is proposed that the Book Industry Study Group will establish a process involving all interested parties that will lead to the development and continuing maintenance of the final standards.

2. Mission statement

The mission of the AAP's Subcommittee for Books Online is to define standards for the discovery, browse, search, and distribution of book content online for marketing purposes. The Technical Specifications will define ways to communicate book content and related data to new kinds of partners: search engines, online retailers and distributors, other third-party web sites, and online communities (collectively "Syndication Partners").

The Working Group's goals are to solve the content management problems for distributing book content on the Web and to increase the presence of books on the Web in a framework that suits publisher needs.

The Working Group will recommend technical specifications which are platform-independent, open standards that are freely implementable by any publisher or publishing partner. The recommendations of the Working Group are not intended as requirements for any party involved and will not preclude participation in existing proprietary arrangements: a publisher satisfied with its current working relationships with online book browsing services will not need to make any change.

The group recognizes that the creation of a digital book archive and the implementation of Web-based software for distribution could require resources beyond what many publishers will be willing to commit. We expect that a clear and simple standard will enable third-parties to offer cost-effective services for conversion, consolidation, and distribution and the group will be soliciting information from representative organizations (such as eBrary, Lightning Source, NetLibrary, O'Reilly Safari, and many others).

To achieve its mission, the group will identify existing standards to leverage, clearly document relationships to parallel efforts, and limit any new introduction of standards to the simplest means of achieving the goal. Particular attention will be paid to the existing tags and terminology of ONIX (Online Information Exchange) from the Book Industry Study Group (BISG). Also closely considered will be the efforts of the International Digital Publishing Forum (IDPF), existing standards for Web services, and of course specifications from the World Wide Web Consortium (W3C), such as XML and HTTP.

The group will work to identify standards that support the project, to list related standards that address parallel issues, and to document the reasons why any new standards must be introduced. As an example, ONIX serves the purpose of communicating data about books, and the Automated Content Access Protocol (ACAP) can be used to communicate data about the permitted use of book content, so this project will complement those two efforts by providing a standard means to discover and deliver the digital book content itself.

3. Usage scenarios

To help illustrate situations where the deliverables of the Working Group will be useful, several usage scenarios are presented below. It is among the goals of the Working Group to enable these scenarios.

3.1. Large book retailer Web sites

Large book retailers have initiated their own development for browsing and consuming digital content online. For several years now, such sites have been delivering online tools that both allow the consumer to search the content inside books as well as provide an online reading environment for the content.

These online booksellers provide the full-service book retail experience. All online aspects of the book search and purchasing experience are hosted at one site: the search engine, the browsing/reading application, the content server providing page scans for the reader, and the e-commerce system for purchase of the book. The Working Group's deliverables will enable such a scenario to be recast in new ways: for example, a bookseller may choose to use the book search services of a separate search engine and access sample digital book content from a publisher's repository, and so on, with each organization providing the appropriate contribution in accord with its usual responsibilities.

The Working Group's deliverables will enable such a bookseller, if it so chooses, to deliver this same cohesive user experience and set of tools, but with the change that the book content would be published and managed from the book publishers' archives, not from the retailer's archive.

3.2. Smaller book retailer Web sites

The Working Group's technical specification would enable smaller book retailers to provide an online book browsing experience in a cost-effective fashion. These retailers could leverage a publisher's Web content repository to deliver authorized content to the retailer's site.

For example, a reading customer visits the retailer's site and is considering purchasing a specific book. The retailer has made a prior arrangement with the publisher to allow the first fifty pages to be freely readable online. In such a dual server situation (retailer front-end and publisher back-end), the fifty image files of the approved pages of the book could be fetched from the publisher's repository by the reader application running on the retailer's Web server. In this case, the retailer does not need to absorb the cost of scanning books and maintaining the page scans repository. The publisher maintains the content and controls the quality of the page images. Additionally, the publisher can centrally track usage of the page files, which provides useful marketing information.

3.3. Search engine indexers

Search inquiries of the book (e.g., searches to find a particular word or phrase within the book) will also be enabled by the group's deliverables.

The Working Group will support indexing of book content for publishers large and small. The Working Group will make recommendations to enable crawlable access both through Web Services (XML, "Web 2.0") and through Web Publishing Guidelines (HTML, "Web 1.0").

The Web Services will make transactions available to search engines that automate the discovery of what books and book pages are available in the publisher's archives. The Business arrangements between individual publishers and clients of the service, which will allow for different levels of access, security, and other details, will be handled within the Web services model, leveraging existing standards wherever possible.

The Web Publishing Guidelines will recommend how to expose a book's content on the Web using nothing more than HTML and simple image files such that the book can be hosted on the simplest of static HTTP servers. The key goal is that no new software would be needed at the publisher's Web site, and the publisher would not need to learn any new technologies. Additionally, a representation of a book would be indexable by the simplest search engine Web crawling spiders, perhaps even a search engine running at the publisher's site.

The Web Publishing Guidelines will express intellectual property usage rules in an industry standard way, via "no-cache" directives in the HTML pages and "robots.txt" files. This is not the strongest of security solutions, but it is simple, it is probably sufficient for limited amounts of content, and it allows publishers to express their intellectual property concerns in a way that they can reasonably expect legitimate organizations to respect.

In both models, access can be restricted using whatever mechanisms are available in the publisher's servers. And Web usage data can be gathered by the publisher in both models.

In both the "Web 1.0" and "Web 2.0" models, the group's goals in the search engine usage scenario are to make the publishing of indexable book content on the Web a straightforward effort and to make the crawling of that content a simple task.

3.4. Individual Web page developers and third-party sites

To help promote books by enabling book browsing on the Web at large, individuals should be able to easily reference the books and pages stored within publisher repositories. For example, an individual reader may want to link from his/her blog, fan web site, or MySpace site to a specific page of their favorite book. The Working Group will recommend ways that URLs can be constructed to point to books, and perhaps to components of books (for example, individual pages or page ranges). Additional relevant content will also be made available such as book cover images and other media.

4. Technical goals

To support the mission, the Working Group will define technical specifications to enable the delivery of book page text and images in a way that will support and respect intellectual property rights.

Initially the Working Group will define a collection of request and response message pairs for obtaining information about and the content contained in collections of books through Web services. Also specified will be implementation best practices for interoperability in situations where books are represented simply as HTML and/or images, without the requirement for any XML-based technologies besides XHTML.

Coherent, useful, implementable specifications will be defined. The specifications will have multiple mechanisms for extensibility. Each publisher can do their own adaptations as necessary, or work with the Working Group to help update the technical specifications.

The following sections enumerate several specific goals of the Working Group.

4.1. Deliver open and non-proprietary standards

The standards will be free of intellectual property rights (IPR) claims, licensing restraints, and royalties. The standards will be made available under an unmodified, popular open source license.

4.2. Enable multi-server user experience and publisher-managed access

The standards will enable the browsing of books on the Web to evolve to where search engines can treat books like other commonly-indexed media. That is, the standards should enable search engines to create a crawlable, full text index of a book yet not make the full content available to customers. Book retailers should be able to programmatically access the book content that publishers decide to make available to help sell the book

4.3. Address the "lowest common denominator"

The Working Group's deliverables will maintain compatibility with existing Web infrastructure. In the context of the Web, the Working Group defines the lowest common denominator as scriptless Web browsers, unsophisticated Web crawlers that can parse only HTML and traverse only "<a href=..." links, and simple XML parsers. In the case of media on the Web, the group defines the lowest common denominator as text and image files served by a Web server.

The Working Group will therefore address both "Web 1.0" and "Web 2.0" paradigms. Specifically, in a Web 1.0 context a book's content would be made available in HTML and image files that can be indexed by simple Web crawlers. For the Web 2.0 situation, collections of books, their text, and the page media would be made available via simple XML Web services.

4.4. Maximally leverage existing standards

The Working Group's deliverables will be based on open standards and be platform independent. For, example, XHTML renderings will have a place within the standards. Wherever appropriate, core XML technologies will be used (e.g. XML 1.0, XML Base, etc.). Also, [XML Schema Datatypes \(XSD\)](#) will be used for datatyping, as much as possible. No specific programming language or platform will be assumed.

The group is evaluating a number of related specifications to leverage for use in this project. Among them are the ONIX set of standards for communicating book data and the IDPF's Open eBook and Open Container Formats. At present no existing standard formats appear to fully meet the group's goals, but the analysis is on-going.

It is explicitly not a goal of the Working Group to define new image file formats or page layout file formats that would compete with technologies such as Portable Network Graphics (PNG), Portable Document Format (PDF), or Joint Photographic Experts Group Interchange Format (JPEG).

The Working Group's deliverables will be agnostic with regard to identifiers, leaving it up to individual publishers to identify the content items (with ISBNs, DOIs or other numbers).

The group will work to identify standards that support the project, list related standards that address parallel issues, and document the reasons why any new standards must be introduced.

4.5. Enable performant implementations

The mechanisms defined by the Working Group will be such that implementations can be realized for high traffic situations. For example, the representation of volumes, books, and documents will be Web cache friendly. That is, the statically renderable components of a book will be mapped to HTTP in a cache-friendly fashion.

4.6. Enable declarations of semantic type for book components

The standards will establish semantic labels ("tags") for components of books (for example, "foreword," "preface," and "table of contents") based on the ONIX protocol.

Publishers can use as few or as many semantic tags in their content as they see fit. The set of values is expected to increase over time as more are suggested to the group.

4.7. Enable flexible content licensing agreements

Publishers will make their own agreements with syndication partners concerning usage rules and permissions. The group's recommendations will provide a way for syndication partners to discover what content has been made available, but authentication and authorization, while extremely important issues, are left out of scope of the current project. Different publishers' server deployments will implement their own mechanisms for authentication and authorization.

4.8. Design for extensibility

There will be mechanisms to accommodate future growth. Specifically, media types (that is, page file and image formats) as well as book part semantic labels will be marked up in XML such that new ones can be added without rendering obsolete previous deliverables.

A. Working Group members

This briefing paper was prepared and approved for publication by the AAP's Subcommittee for Books Online Working Group. The participants of the Working Group at the time of publication of this document are as follows.

- Ed McCoyd, AAP (*AAP Liaison*)

- Leslie Hulse, HarperCollins (*Co-chair*)
- Adam Young, Random House (*Co-chair*)
- John Tigue, Rosetta Solutions (*Technical Team Chair*)
- Kelley Allen, Random House
- Marjorie Fowler, University of North Carolina
- Saju George, Holtzbrinck
- David Haase, Random House
- Chris Hart, Random House
- Theresa Horner, HarperCollins
- Tina Jordan, AAP
- Daniel Lee, Yale University Press
- Peter McCarthy, Random House
- Corey Podolsky, Oxford University Press
- Richard Rothstein, HarperCollins
- Patricia Schroeder, AAP
- Laura Stevens, Thomson
- Keith Titan, Random House
- Ted Treanor, Rosetta Solutions

B. Production notes

John Tigue served as editor of this Briefing Paper.

This document was encoded in a strict subset of [DocBook 4.4](#), an industry standard XML mark-up language. The PDF version was produced with Rosetta Solutions Inc.'s RosettaMachine©. The fonts used were Courier, Helvetica, and Times (New) Roman.

[This MS Word .doc version of the briefing paper was produced by RosettaMachine as xhtml and then that was imported into Microsoft Word 2003. From there the content was saved as a .doc file. The goal in this case is simply to provide a mechanism for AAP members to provide feedback within a file. No effort was made on the aesthetic front for this version. The final Briefing Paper will be released as a PDF file.]