

AAP Requirements for Standardizing the Discovery of Digital Book Content

document version v 1.8
June 12, 2007

Revision History

Date	Version	Author	Description
11/21/2006	1.0	D.Haase (RH)	First draft at uniting use-cases from all sub-committee members
12/20/2006	1.1	D.Haase (RH)	Draft adding usage scenarios.
1/11/2007	1.2	D.Haase (RH)	Corrected typo errors
1/23/2007	1.3	D.Haase(RH)	Updated draft per work session (1/11/2007)
1/25/2007	1.4	D.Haase (RH)	Updated per work session (1/24/2007); <ul style="list-style-type: none">- added Web 1.0 Usage scenario;- added details to two use cases regarding the availability of content from publisher and,- added appendix to reference XML examples in spec documents from H/C and RH.
2/04/2007	1.5	A. Young (RH)	Updated to reference latest version of HarperCollins and Random House specifications.
4/3/2007	1.6	D. Haase (RH), A. Young (RH)	Updated to reflect edits from AAP Counsel
4/27/2007	1.7	D. Haase (RH), A. Young (RH)	Updated to reflect AAP technical specifications
6/12/2007	1.8	D. Haase (RH) A. Young (RH)	Final changes and formatting prior to AAP membership review

Table of Contents

1	Project Overview	1
1.1	<i>Purpose</i>	1
1.1.1	Promote and Distribute Digital Book Content	1
1.1.2	Manage Book Content	1
1.2	<i>Users</i>	2
1.2.1	Syndication Partners	2
1.2.2	Publishers	2
1.3	<i>Usage Scenarios</i>	3
1.3.1	Large Bookselling Partner (Amazon, Barnes & Noble, etc.).....	3
1.3.2	Small Bookselling Partner (Independent booksellers, stores that only sell a few titles alongside other merchandise, etc.)	3
1.3.3	Search Engine (Ask, Google, Yahoo!, Microsoft, etc.)	4
1.3.4	Social Networks and other third-parties (MySpace, blogs, fan websites, online reading software, etc.).....	4
1.3.5	Arrangements using only Web 1.0 / site map technologies (Internal publishers' web search technology, simple websites, etc.).....	4
1.4	<i>Deliverables</i>	5
2	General Requirements.....	5
2.1	<i>GR-01: Leverage Existing Standards</i>	5
2.2	<i>GR-02: Work within Prevailing Standards Bodies</i>	5
2.3	<i>GR-03: Use Widely Accepted Browser Technologies</i>	5
3	Functional Requirements.....	6
3.1	<i>Agents</i>	6
3.2	<i>Use Case Overview</i>	6
3.3	<i>Use Case Descriptions</i>	7
3.3.1	UC-01: Get Number of Keyword Matches in Archive	7
3.3.2	UC-02: Search Archive for Keyword	7
3.3.3	UC-03: Search Book for Keyword	8
3.3.4	UC-04: Enumerate Archive	8
3.3.5	UC-05: Enumerate Book	8
3.3.6	UC-06: Enumerate Onix Info	8
3.3.7	UC-07: Enumerate Tag Info	9
3.3.8	UC-08: Browse by Context Pages.....	9
3.3.9	UC-09: Browse by Sample Pages	9
3.3.10	UC-10: Get Book Media	9
3.3.11	UC-11: Get Full Page Media	9
3.3.12	UC-12: Get Thumbnail Media.....	10
3.3.13	UC-13: Get Searchable Text.....	10
3.3.14	UC-14: Get Word Coordinates	10
4	Appendix A – Use Case Comparison to Current Implementations by Publishers.....	11
5	Appendix B – Related Examples of Actual XML Implementations	12

1 Project Overview

In the Spring of 2006, the *Digital Issues Working Group* (DIWG) of the *Association of American Publishers* created the Sub-committee for Books Online. The goal of the Sub-committee for Books Online is to recommend industry standards for the discovery, browse, search, and distribution of digital book content over the World Wide Web. This document outlines the initial use cases and related requirements to achieve that goal.

1.1 Purpose

The standard will address the fundamental question: *How can book publishers distribute and promote discovery of digital content over the web in a manner that allows them to manage the availability and quality of the content?*

1.1.1 Promote and Distribute Digital Book Content

The standard should encourage the promotion of digital content at all levels of distribution – from the individual social-network user to the large, retail partner or search engine and any web developer in between. The messaging protocol will enable web developers to easily and consistently *discover, browse, search* and *display* digital book content made available by any publisher conforming to this standard. For the broadest promotion, the standard will be developed as an open-source messaging protocol, which will be freely published and made accessible to web developers of all sizes.

Therefore, the standard will not be a specification of the format of the digital content, but rather a specification for how to 1) discover what books are available, 2) browse a book's available content, 3) distribute the book's content to partners, and 4) enable a search of the book's text.

1.1.2 Manage Book Content

The proposed standard will allow a publisher to build systems to manage the quality, distribution, and availability of its book content in a way that works directly with its internal production process but also integrates with its major partners. The standard is designed to complement the existing partner-specific models of distribution. By implementing the proposed standards, publishers will be able to make their content available to more outlets in a way that fits their business while continuing to work within existing arrangements.

1.2 Users

The expected users of the standard fall into two main groups: one group of users is the *syndication partners* that display digital book content and the other group is the *publishers* that produce digital book content. The proposed standard is meant to define the interface between the publishers of digital content and the syndication partners' software that delivers the content to consumers.

1.2.1 Syndication Partners

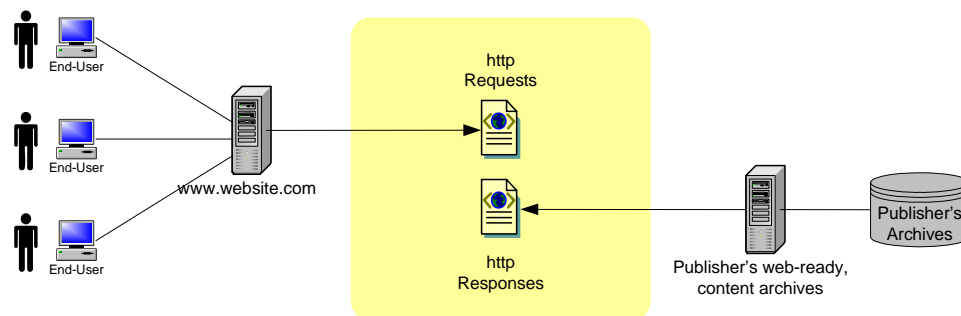
The standard will provide the opportunity for a wide range of distribution: from a novice individual adding a book to his or her blog, to the websites of the largest retail book-selling partners.

The standard will target the following distribution profiles in particular:

- *Large Retail Bookselling Partners.* Some large retailers have already built their own digital book archives and have successfully distributed digital content through their own channels. Creating a standard for digital content provides a second option. Publishers would assume the responsibility for creating and maintaining the archives, the costs of which would be assumed by the publisher. The retailers would be free of this investment while the publisher has the incentive of increased distribution and direct control of the quality, versioning and updates of the content.
- *Smaller Retail Bookselling Partners.* Many smaller booksellers are not willing or able to build their own digital book archives. But they would likely be able to build software to deliver book content on their web site, assuming a standard allowed them to access many archives in the same way.
- *Search Engines.* A common standard for exposing crawlable book content will allow search engines to access many publishers' archives in the same way, eliminating much of the need to build their own book archives. As in the case of the large retailer, a standard can free up search engines from the investment of building their own digital archives. Publishers could assume this responsibility with the incentive of wider distribution and direct control over the quality and version of what is being crawled.
- *Social Networks (blogs, etc.).* Currently, very little book content is available to these kinds of outlets, despite their growing popularity among avid readers. The standard will help these independent web developers to discuss their favorite books and feature the book's content thus expanding valuable word-of-mouth marketing which increases book sales.

1.2.2 Publishers

The proposed standard will provide a medium for publishers to determine how their digital content will be *discovered, browsed, distributed and searched*. By delivering their digital content via the proposed standard, publishers of all sizes can manage its availability and guarantee the quality of their content to the widest audience possible. Importantly, by having an established protocol to follow, publishers will gain an efficiency which will allow them to prepare their digital content more consistently and efficiently as opposed to adapting their content to the varying requirements of each syndication partner.



1.3 Usage Scenarios

The following five sections outline potential and expected implementations of the proposed standard:

1.3.1 Large Bookselling Partner (Amazon, Barnes & Noble, etc.)

One of the primary implementations of a standard for digital book content distribution is with major online book retailers. By providing these large partners with access to consistent and reliable tools for accessing book content, it allows them to display a wider selection of book content while enabling book publishers to manage updates and availability.

As an example, the website of a large bookselling partner could:

- Find all books available in a publisher's service, or all new books since a given date, and related ONIX information (**UC-04 and UC-06**)
- For each book, request all the pages available, the pages' designated types, and suggested sample pages (**UC-05, UC-07, UC-09**)
- For each available page, request the text, fullpage and thumbnails (**UC-11, UC-12, UC-13**)
- Import the text and page content into bookseller's own search index
- Display the book content in the bookstore webpages and sell the book direct to consumer
- Check back with the publisher's service for books that have been updated, added, or deleted (**UC-04**)

1.3.2 Small Bookselling Partner (Independent booksellers, stores that only sell a few titles alongside other merchandise, etc.)

Similarly, the standard can benefit the smaller bookselling partner as well. By providing retail partners with an open-sourced, straightforward and consistent messaging protocol, publishers give smaller bookselling partners an easy and attractive incentive to distribute and promote content on the web. At the same time, publishers maintain ownership and guarantee the exact quality of the content being distributed.

As an example, a small boutique bookseller could easily build a web-based customer experience, which displays publisher-managed content.

- Discover which books have online content available (**UC-04 and UC-06**)
- For books with matches, discover what pages are freely available and which are not (**UC-05**)
- If a page is viewable, retrieve the thumbnail and fullpage from the publisher and display to the consumer (**UC-05, UC-11, UC-12**)
- Switch to view the suggested sample pages or specially tagged pages (**UC-07, UC-09**)

1.3.3 Search Engine (Ask, Google, Yahoo!, Microsoft, etc.)

Another very important implementation of the standard is with online search engines. By offering a standard and consistent messaging protocol, search engines of all sizes and complexities can easily begin to crawl digital book content and start serving up the results to potential customers. Again, by defining a standard, publishers will be able to exercise control over their content, including what is available while increasing the distribution of book content to many partners.

A straightforward example for search engines would be the following sequence:

- Find all books available in a publisher's service, or all new books since a given date (**UC-04**),
- For each book, request all the pages available, the pages' designated types, and suggested sample pages (**UC-05, UC-07, UC-09**)
- For each available page, request the text, fullpage and thumbnails (**UC-11, UC-12, UC-13**)
- Import the text and page content into search engine's own search index.
- Display the book content in the search results and link to booksellers
- Check back with the publisher's service for books that have been updated, added, or deleted (**UC-04**)

1.3.4 Social Networks and other third-parties (MySpace, blogs, fan websites, online reading software, etc.)

Another popular target for the implementation of the standard will be the development of lightweight, online media tools that will distribute promotional elements of books in social networks and personal websites. By having an established standard interface for retrieving promotional book content, third-party developers can begin building online tools for viewing, annotating and sharing book content. As in the implementations above, a defined standard will help to ensure that the publishers retain ownership and management of the content.

An interested third-party could build a free online book application that is embedded in websites, blogs, etc. and would display promotional pages of a book:

- The application might first advertise with a thumbnail of the book's cover (**UC-12**)
- Then would identify which pages are freely available (**UC-05**),
- And finally display the full pages that are available (**UC-11**)

1.3.5 Arrangements using only Web 1.0 / site map technologies (Internal publishers' web search technology, simple websites, etc.)

A fifth target for the standard will be the partners who avail themselves only of Web 1.0 technologies (e.g., plain HTML and images). To enable this scenario, the standard must be able to support the publishing of content on static, linked pages, access to which requires no web "service" technology.

The following shows how the proposed use-case requirements would support this model:

- An application (or person) creating the sitemap would first need to select a book from a list of available books (**UC-04**).
- For that book, the application would need to discover what pages are available, what formats of page media are available, and where they can be located (**UC-05**).
- Having located the appropriate information for the title, the application would create a directory and series of HTML pages using page media such as plain text and page images, based on the available material (**UC-11, UC-12, UC-13**).
- The URL for that simple directory would be made available on a website for crawling or browsing.

The standard should eventually recommend an HTML sitemap into which book content can be structured, thus allowing even the simplest search engines to crawl the available content link by link.

1.4 Deliverables

This document, the “AAP Requirements for Standardizing the Discovery of Digital Book Content”, represents the second of three deliverables scheduled for release from the subcommittee:

1. the first document was the “AAP Subcommittee for Books Online Briefing Paper”;
2. this document, “AAP Requirements for Standardizing the Discovery of Digital Book Content”;
3. the third document will be the recommended technical specification to support these requirements.

The subcommittee is following a process, outlined in the briefing paper, that will deliver these documents into a larger discussion forum at the BISG where all interested parties can come together to finalize the specification.

2 General Requirements

The proposed standard will satisfy the following general requirements.

2.1 GR-01: Leverage Existing Standards.

Where possible, the proposed standard will leverage existing standards in the book industry with particular attention to the existing tags and terminology of ONIX from the BISG. Also closely considered will be the efforts of the IDPF, existing standards for Web services, and of course specifications from the 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 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.

2.2 GR-02: Work within Prevailing Standards Bodies

The standard will seek the adoption and ratification of the appropriate standards bodies, BISG, OASIS, IDPF, ISO and others as necessary.

2.3 GR-03: Use Widely Accepted Browser Technologies

The proposed standard will be free from proprietary technologies where possible and instead leverage standardized internet technologies like JPEG, XML, and HTTP. Within a framework that will offer a high degree of compatibility, the standard will define both the protocol needed to request digital content from publishers as well as the format in which developers can expect to get a response.

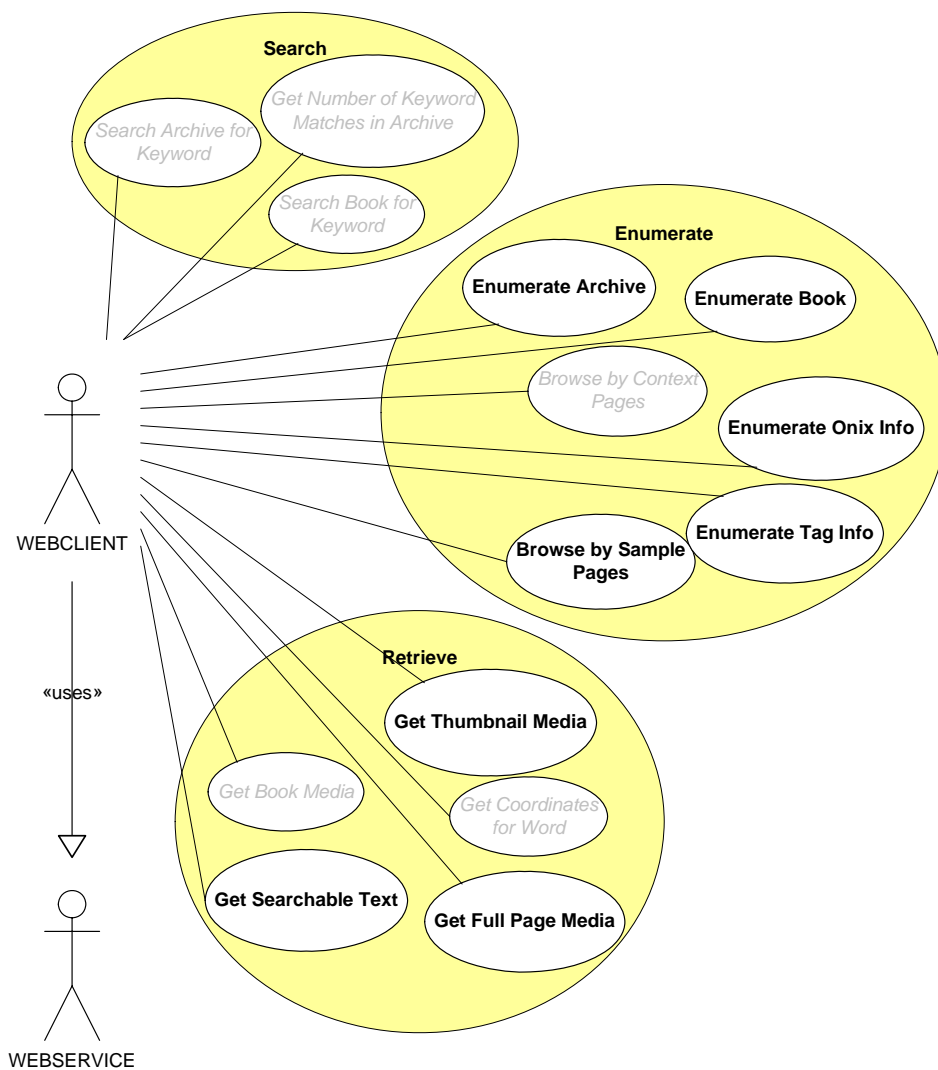
3 Functional Requirements

3.1 Agents

WEBCIENT	Any software client application availing itself of web-ready, digital book content through the internet: <i>the website of a retail partner, search engines, digital communities, internal archiving systems, etc.</i>
WEBSERVICE	A publisher's digital content server with an http interface delivering web-ready, digital content. It is precisely the details of the "http interface" which are under design.

3.2 Use Case Overview

The AAP proposes the fourteen use cases¹ shown in the diagram and table below.



¹ For the full table on how these use cases map to the AAP technical recommendation and implementations by both Random House and Harper Collins, see Appendices A. It is important to the DIWG to create a standard that will serve as many publishers' needs as possible. The DIWG therefore invites and encourages the adaptation of specifications from other publishers into the final requirements.

Table of Use Cases

SEARCH	UC-01	Get Number of Keyword Matches in Archive
	UC-02	Search Archive for Keyword
	UC-03	Search Book for Keyword
ENUMERATE	UC-04	Enumerate Archive
	UC-05	Enumerate Book
	UC-06	Enumerate ONIX Info
	UC-07	Enumerate Tag Info
	UC-08	Browse by Context Pages
	UC-09	Browse by Sample Pages
RETRIEVE	UC-10	Get Book Media
	UC-11	Get Full-Page Media
	UC-12	Get Thumbnail Media
	UC-13	Get Searchable Text
	UC-14	Get Word Coordinates

3.3 Use Case Descriptions

The descriptions below describe the fourteen use cases proposed by the AAP in terms of request / response pairs.

Search Functionality

3.3.1 UC-01: Get Number of Keyword Matches in Archive

Search the entire publisher's archive to get a total count of books and pages that contain the keyword.

Request: *How many books and pages from the archive contain the word "Achilles?"*

Response: The keyword Achilles appears in 12 books and on 678 pages of the archive.

3.3.2 UC-02: Search Archive for Keyword

Search the entire archive to get a list of book titles and excerpts that contain the keyword.

Request: *Which books contain the keyword, "Achilles," and what is the context in which it first appears?*

Response: A list of 12 book titles in which the keyword Achilles appears as well as the first page and pageID in which it appears.

3.3.3 UC-03: Search Book for Keyword

Search a specific book title to get a list of search results and links to pages that contain the keyword, within the specified range.

Request: *What pages of the book, The Iliad, contain the keyword, "Achilles," and what is the context in which it appears?*

Response: A list of links to the 58 pages with pageIDs of the book, The Iliad, on which the keyword Achilles appears as well as links to the context in which it appears.

Enumeration Functionality

3.3.4 UC-04: Enumerate Archive

Get a list of links to the available media of all the books in the archive. Entire books may or may not be available for viewing by the requesting system. The control of which books that are available for viewing will be indicated by the publisher: which books are available, which books are no longer available, or which books have been updated. The request can be optionally specified to include parameters that restrict the result set to a specified range and/or by the book's date-modified value. Similarly, specifying the request to a single ISBN is also optionally available.

Request: *Where can I find links to the searchable texts, sample chapters, advance readers' copy, and other media for the first 100 books in the archive made available since November 1, 2006?*

Response: A list of the first 100 book elements made available in the archive that include the ISBN, the last date modified and a links to the media for each book.

3.3.5 UC-05: Enumerate Book

Get a list of links to the page media (thumbnail, full-page images, searchable text, etc.) of the available pages of a book. Individual pages of a book may or may not be available to the requesting system; therefore, the control of which pages of a book are available for viewing, etc., remains in the hands of the publisher who will indicate which – and how many – pages are available. Requests can be optionally restricted to a range of page results.

Request: *Where can I find a list of links to the thumbnail, full-page and searchable text of the first 50 pages from The Iliad.*

Response: A list of 50 page elements that contain URLs to the thumbnail and full-page images as well as to the searchable text of the first 50 pages of The Iliad. Get a list of pages and its available media from a specified book.

3.3.6 UC-06: Enumerate Onix Info

Get a list of ONIX information available for a specified book.

Request: *What information is available in ONIX format for The Iliad?*

Response: The ONIX-formatted XML record for the specified book.

3.3.7 UC-07: Enumerate Tag Info

Get a list of book-level information and media available for a specific book.

Request: *What media and information is available for The Iliad?*

Response: A thumbnail, sample pages and audio excerpt is available for The Iliad.

3.3.8 UC-08: Browse by Context Pages

Get a list of links to thumbnail and full-page images for a specified number of pages before and after a specific page in the book. This use case enables browsing forward and back, or jumping a few pages in either direction.

Request: *What are the links to the five pages before it and after page 256 of The Iliad?*

Response: A list of thumbnail and full-page URLs to pages 251-261 of The Iliad.

3.3.9 UC-09: Browse by Sample Pages

Get a list of links to thumbnail and full-page images for a group of pre-determined sample pages available for the specified book (e.g., cover, backcover, TOC, etc.), as chosen by the publisher.

Request: *Where can I find links to all of the sample pages made available from The Iliad.*

Response: A list of thumbnail and full-page URLs to pages the front cover, table of contents, first index page, and first pages from sections of The Iliad.

Retrieval Functionality**3.3.10 UC-10: Get Book Media**

Get book-level media (e.g., Review Copy, Sample Chapters, Audio, etc.) of a specified book.

Request: *The PDF Review Copy of The Iliad.*

Response: A PDF representation of the review copy of The Iliad. This transaction may deliver other media types. The available media types for a page will be clearly specified by the publisher in other use cases.

3.3.11 UC-11: Get Full Page Media

Get full-page media (e.g., JPEG, PDF, etc.) of a specified book by page number. These pages are of sufficient quality for reading, but the publisher decides the quality, size, and media type. The availability of page media may be determined by the publisher based on the requesting system or any other criteria.

Alternative versions of full-page media can also be delivered. For any given full-page media, the publisher may also concurrently offer alternate versions of the same page; e.g. page image without illustrations.

Request: *The full-page image of pageID 256 of The Iliad.*

Response: A full page representation of the page corresponding to pageID 256, or the alternative content for pageID 256, of The Iliad. The service may respond with JPEG images or other media types. The available media types for a page will be clearly specified by the publisher in other use cases.

3.3.12 UC-12: Get Thumbnail Media

Get thumbnail media (e.g., JPEG) of a specified book by page number. The thumbnails are useful for displaying search results, to indicate the kind of content on the page (full text, pictures, etc.). They are not intended for reading.

Alternative versions of thumbnail media can also be delivered. For any given thumbnail page media, the publisher may also concurrently offer alternate versions of the same page; e.g. page image without illustrations.

Request: *The thumbnail image of pageID 256 of The Iliad.*

Response: A thumbnail representation of the page corresponding to pageID 256, or the alternative content for pageID 256, of The Iliad. The service may respond with JPEG images or other media types. The available media types for a page will be clearly specified by the publisher in other use cases.

3.3.13 UC-13: Get Searchable Text

Get the “plain” text from a given page that the publisher has made available. The availability of page media may be determined by the publisher based on the requesting system or any other criteria.

Request: *What is the text in a searchable format for page 32 of The Iliad?*

Response: The plain ASCII text available for page 32 of The Iliad.

3.3.14 UC-14: Get Word Coordinates

Get a list of links to the two-dimensional coordinates for each word of a search phrase on the full-size page image.

Request: *What are the coordinates of each instance of the word Achilles on the full-size page images of The Iliad?*

Response: A list of coordinates from each page of The Iliad on which the word Achilles is found.

4 Appendix A – Use Case Comparison to Current Implementations by Publishers.

The following table maps the AAP’s proposed use cases to the transactions defined in the AAP technical specification. For those familiar with existing implementations, it also maps the use cases to the Harper Collins and Random House service specifications. The DIWG invites and encourages comparison to a broad range of publishers’ repositories and will update the comparison as these specifications are provided by other publishers.

AAP Specification		Harper Collins	Random House
Search			
UC-01	TR01: Get Number of Keyword Matches from Archive		A-Whole Archive Keyword Search Summary
UC-02	TR02: Search Archive for Keyword	3.2 – Search Catalog	B-Whole Archive Keyword Search Results
UC-03	TR03: Search Book for Keyword	3.1- Search	C-Book Keyword Search
Enumerate			
UC-04	TR04: Enumerate Archive	1.1- Enumerate Catalog	I – Book Index
UC-05	TR05: Enumerate Book	1.2- Enumerate Title	J- Page Index
		1.12- Scanned Pages 2.8- Page Level - Scanned Page	J – Page Index
		1.9- PDF Page Images 2.6- Page Level - PDF Page Image	J – Page Index
		1.6- JPG Page Images 2.2- Page Level - JPG Page Image	J – Page Index
		1.7- Thumbnail Images 2.4- Page Level - Thumbnail Image	J – Page Index
UC-06	TR06: Enumerate ONIX Info	1.3- ONIX	
UC-07	TR07: Enumerate Tags	1.8- Tag Information 2.5- Page Level - Tag Information	
UC-08	TR08: Browse by Context Pages		F - Book Page Context
UC-09	TR09: Browse by Sample Pages	3.3- Sample Pages	G - Book Sample Pages
Retrieve			
UC-10	Get Book Media	1.10- HTML Page Rendition	M – Book Media
UC-11	Get Full Page Media	2.7- Page Level - HTML Page Rendition	D – Full Page
UC-12	Get Thumbnail Media		E – Book Page Thumbnail
UC-13	Get Searchable Text	1.5- Text	J - Page Index, K - Page Text
		2.1- Page Level – Text	K - Page Text
UC-14	TR14: Get Word Coordinates		L - Page Text Coordinates
Misc			
	Error Response	4.1 – Error Reply	H - Error Response

5 Appendix B – Related Examples of Actual XML Implementations

The AAP recommendation for the ONIX implementation of these required use cases is documented in “AAP Technical Recommendation for Standardizing the Discovery of Digital Book Content.” The following documents are listed below for context.

HarperCollins

Digital Warehouse XML Interface Specification, version IX, January 2007

Available upon request from AAP.

Random House

Insight Service Specification, version 2.3, March 2007

Available upon request from AAP.