1 Insight and the Random House Digital Page Initiative

The Random House Digital Page Initiative is an on-going project to index, digitize, distribute, and set the terms for using book content online. As part of that initiative, Random House has developed a service that gives search engines and online retailers access to digitized book content over the Web. Publishers need to manage their published content, and an increasingly large percentage of this content is digital. The Insight service was developed to address the relationship between book publishers, their digital content, and the Web at large.

Insight addresses this relationship specifically by maintaining the publisher's proper ownership and management of the content and giving business partners access to easy-to-use, browser-compatible tools to search, view and retrieve digitized book content over the Web.

1.1 What is Insight?

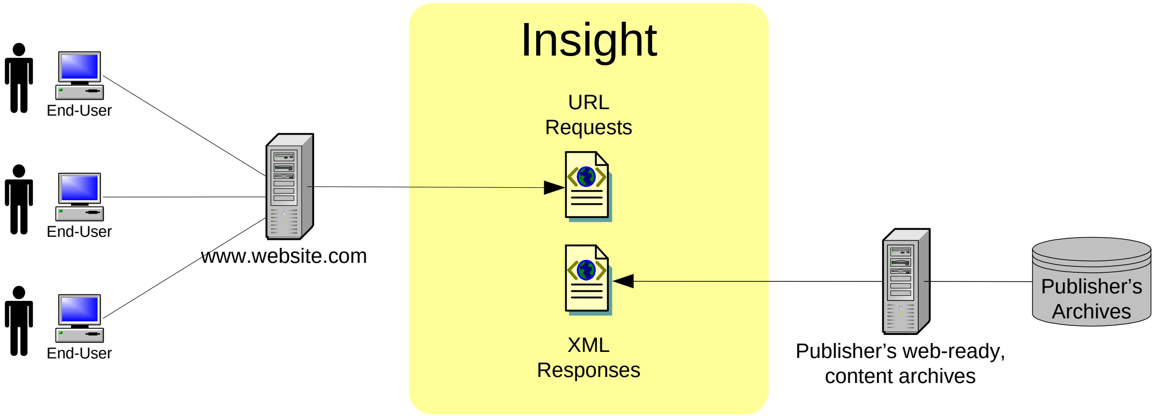
Fundamentally, Insight is a form of "middleware." Middleware is type of transparent software that brings other software together. In the case of Insight, it connects software from partner websites or search engines to the web-ready content archives made available by publishers. The publishers can then offer or sell the content to the partner, when requested by end users on the partner's website.

For a web developer at a retail partner, Insight is a lightweight and browser-compatible tool, requiring no special updates or plug-ins for the end-user. Insight uses familiar industry standards like JPEG images and XML formatting to display actual book page views as well as to provide keyword searches in the text of a title.

For the publisher, Insight is a tool to get the publisher's digital content onto the websites of retail partners, search engines, publicity outlets, authors, blogs, and readers. Insight leverages existing industry tools like ONIX to work with partners; it implements business rules to guarantee that ownership and management of the digitized content remains with the publisher; and it manages access to the content from third-party websites.

To those ends, Insight codifies the following:

- the rules of access to a publisher's digital archive,
- the tools a developer can use to search and view a publisher's archive, and
- the format in which the developer can expect to get the data returned.



1.2 How does it work?

Insight is a set of transactions for requesting book content, delivered in an online industry standard, XML. These XML transactions are defined by URL requests that can be downloaded from the Web and embedded into a partner's website or a search engine. A more detailed description of these transactions can be found below in Section 3 of this document and in the Insight Service Specification document available at: http://randomhouse.biz/webservices/insight/spec.php.

The embedded transactions are submitted over the web to a publisher's Insight server in the form of simple URL requests. The publisher's Insight server authenticates the user, tracks the request, and responds with the appropriate data, formatted in either XML or as page media such as JPEG images.

1.3 Who uses Insight?

The tools of Insight target the following types of users:

Online Retail Partners – In conjunction with existing book title resources like ONIX, Insight enhances the websites of online retail publishing partners by providing page views of actual book pages as well as "search inside" functionality for text.

Search Engines – Insight provides a secure gateway for search engine spiders like Google to crawl book content at the publisher's discretion.

Social Networks – The transactions of Insight can easily be wrapped into a tool for use in online social communities like MySpace.com, etc.

1.4 How do I get started?

Publishers: The management and ownership of the content remains in the hands of the publisher.

- Prepare the content data in the desired format; e.g., JPG, PDF, indexed text, etc.
- Decide which parts and/or how much of the book will be made of available to Insight and at what price.
- Message this information to an Insight-compatible system.

Development partners:

- Set up an Insight partnership with the publisher to define and authenticate access to the service.
- Download and review the Insight Specification document (http://randomhouse.biz/webservices/insight/spec) and begin enhancing websites/search engines code with features for keyword searches and full page or thumbnail page views.

2 Digitized Content

The transactions of Insight support any type of media available. Insight will always pass a content-type parameter along with the data itself. Digitized book content however, often falls into two primary categories: *image* and *text*. A priority for the development of Insight was browser-compatibility for these categories with web-friendly formats like XML text and JPEG, thereby avoiding the need for browser plugins or Java applets.

2.1 Structure of Insight Elements

The first release of Insight relies on the following hierarchy for content elements. The on-going project is compiling a plan for additional groups of content.

Archive – a collection of books made digitally available by a publisher. The availability of an archive and its contents remains at the sole discretion of the publisher.

Book – a collection of pages made available by the archive’s publisher. Each book is uniquely identified by the ISBN. Digital book content is returned as a collection of page images (e.g., JPEGs) or as a link to searchable text.

Sample Pages – a group of pages from a book that were designated as featured sample content for that book. Examples include the table of contents, the first pages of chapters, or the first page of the index.

Page – page content is returned as an image of words, excerpts and paragraphs physically bound by the actual page in print (JPEG, GIF, PDF). Pages are uniquely identified by the pageID.

Text – a searchable string of words or paragraphs excerpted from the actual page text.

Search Results – a list of matching pages for a given keyword phrase, with small text blurbs highlighting the matching text from a page.

3 Insight Example Use Cases

(A) Whole Archive Keyword Search Summary

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 “Ulysses?”*

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

(B) Whole Archive Keyword Search Results

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

Request: *Which books contain the keyword, “Ulysses,” and what is the context in which it first appears?*

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

(C) Book Keyword Search

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, “Ulysses,” 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 Ulysses appears as well as the excerpted text in which it appears.

(D) Book Full-Page Transaction

Get full-page media (e.g., JPEG) 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.

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

Response: A full page representation of the page corresponding to pageID 256 of *The Iliad*. Insight responds with JPEG images, but the service could respond with another media type.

(E) Book Thumbnail-Page Transaction
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.

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

Response: A thumbnail representation of the page corresponding to page 256 of *The Iliad*. Again, Insight currently responds with JPEG images for this request, but it could respond with another media type.

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

Request: *Where can I find links to the five pages before and after page 256 of* The Iliad*?*

Response: A list of thumbnail and full-page URLs to pages 251-255 and 257-261 of *The Iliad*.

(G) Sample Book Page Transaction
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*.

(H) Error Response
For any of the requests in the service, various error values and alerts may be sent as a response. These can include:

- when a publisher has made a page or book unavailable,
- when content is not available for a particular user, or
- when the request was not formatted correctly.