Linguistics: Adding Value to e-Publishing and e-Content
Executive Summary

Constant change is the overriding factor shared by thousands of electronic publishing companies that provide information through Web sites. Traditional information distribution mechanisms (email, postal services, etc.) must now coexist with the chaos of the Web and its billions of pages. Smart, agile publishers and information aggregators – grouped under the headings e-Publishers and infomediaries – are learning that their success is dependent on monetizing content while protecting its inherent value from creeping “Napsterization” – all the while bucking the “everything for free” culture of the Internet.

So, the problem becomes how to charge a premium for information that the online world assumes should be free of charge. The answer lies in adding significant, obvious and relevant value by delivering specific, targeted information that saves users the time, effort and resources required to sift through the Web’s massive data storehouses themselves.

Of the products and technologies targeting this market, only Inxight’s e-Content Publishing Platform has the appropriate combination of functionality and underlying architecture to achieve the three key e-Publishing objectives: return on investment, improved productivity and increased value of information.

As the Web’s overall content spirals upward into the realm of exabytes and beyond, the sheer volume of information that requires high-speed, intelligent search and retrieval software has presented an opportunity for editors to embrace new technology systems that replace former error-prone methods. Inxight’s e-Content Publishing Platform presents e-Publishers with the best available choice.

e-Publishing Industry Overview

Electronic publishing is growing at a staggering rate. Up from $146 billion in 1999 (Outsell, Inc.), industry insiders suggest that the number could now be as high as $300 billion, and the expansion of the Web and the Internet as information distribution engines can only spur further growth.

IDC estimates that 200 million pages are added each month, while an estimated 100 million become obsolete. The latter figure demonstrates clearly why information management on the Web is so difficult. Obsolete pages proliferate misinformation, although there’s usually no way for users to gauge the validity of any given page.

Timeliness is a key element in certain e-Publishing areas, notably the distribution of news content. No one is going to pay a premium for yesterday’s news. In other areas, though, time is less important.

Looking at the bigger picture, the Web has imposed a set of issues unknown in the publishing industry a decade ago, including the commoditization of information, the explosive growth of content and content aggregators, and the focus on profitability – to the chagrin of those who would contend that the Web must continue to operate at no cost to the user.
While the Internet and Web are generally regarded as free sources of information, business realities have created new mandates to generate revenue and profits, driving the publishing industry to control intellectual property assets more closely. Market factors affecting publishing today include:

- Growth and expansion of e-Publishers/infomediaries
- High expectations for quality and the need for specific content
- User expectations and demand for real-time information
- New players, roles and commoditization lead to slimmer margins
- Information needs to be available anytime, anywhere
- Content vendors need to adapt their content to support expanded media

The Players
These days, the e-Publishing distribution system is dominated by “infomediaries.” These include many of the leading syndicators from the “old economy,” such as Reuters and Dow Jones, along with the new breed of aggregation companies that serve as go-betweens, linking content providers with mass audiences. The latter group includes YellowBrix, iSyndicate, Factiva and Screaming Media. They don’t produce their own content, nor do they have the kinds of direct relationships with information consumers that newspapers and magazines create. Rather, they serve as matchmakers for both ends of the distribution chain, linking content with knowledge consumers.

Infomediaries separate content from a specific interface or distribution channel, and make it available through an infinite number of channels. They aggregate content from many primary publishers – just as the traditional syndicators did – but instead of rolling the content into a single one-size-fits-all package, they offer slices of the content to a Web site or intranet that wants to add third-party content to its mix.

The e-Publishing cast looks like this:

- Infomediaries, including syndicators who collect information from selected sources and resell it to other buyers. Leading players include Factiva, The New York Times Syndicate, and iSyndicate. Aggregators are also part of the infomediary group. They don’t own the content, but collect multiple information types from multiple sources and re-distribute it to buyers. Key players include Lexis Nexis, Dialog and Screaming Media.
- Buyers, in this case an organization that buys content on behalf of end users and publishes it on an information portal. The buyer could be a private company or a commercial portal like msn.com or yahoo.com.
- End users, or information consumers. Players include individuals browsing Web sites for private use or business users collecting decision support information for commercial use.

Industry Dynamics
As the nature of the publishing business changes – and specifically the move to e-Content – so, too, do the drivers and inhibitors that dictate industry dynamics. As we’ve already seen, the sheer volume of information is growing exponentially. As a result, applications that categorize, tag and cull Web-based information today must encompass terabytes of data. In the not too distant future, applications must be capable of performing the same set of functions on to ‘exabytes’ of data (one million terabytes). The demand, therefore, is for information management architecture that scales to accommodate a virtually infinite set of information derived from multiple sources.

These sources include corporate ERP/CRM applications, document management systems, external research materials, news feeds, interpersonal email, and data housed on personal computers. Metadata – the “rules” by which information is categorized and culled, is scrambled
and non-standardized. In addition, commoditization has impacted the value of e-Content and made it difficult to create enough perceived value to charge money for it.

The obvious needs are for highly refined filtration systems and tight, standardized organizational methodologies. In short, information delivery needs to “get personal,” sorting through the maze of online content to zero-in on the requirements of specific users and organizations. Let’s now examine the technologies that make that possible.

**Technologies**

Just as HTML has led the pack in describing Web presentation, XML (eXtensible Markup Language) is growing as the industry standard used by application developers as the mechanism to define key information within documents. Used by a variety of enterprise applications, XML can be used as the common link between data from disparate sources. XML is essential in architecting a robust enterprise system, and allows for reuse of information across multiple, integrated applications.

As *PC Week* (now *eWEEK*) stated in 1999, “Like HTML, XML derives from the granddaddy of all markup languages: SGML (Standard Generalized Markup Language). SGML is a meta-language, or a system for defining markup languages such as HTML. XML is also a meta-language, a subset of SGML designed for use on the Web. As with SGML, you can use XML to define different markup languages for specific uses, particularly for data representation.”

Adding to the confusion, different XML derivatives are constantly being promulgated as standards by various industry groups. Among the markup languages in publishing, whether NewsML, PRISM, NITP, each leverages XML, but modifies it for specific applications (e.g., NewsML for news publishers and ICE <Information and Content Exchange> as a protocol to automate content syndication. Now, a look at how Inxight’s underlying technology and application-level systems combine to deliver the most efficient e-Publishing content delivery solution available today.

**The Inxight Business Case**

Strong technology and a rich feature set is one thing; mapping these capabilities to address real business needs is another. Fortunately, Inxight’s e-Content Publishing Platform excels at delivering the benefits that define a successful e-Publishing solution: return on investment; productivity increases; and increased value of information.

**Increased ROI**

Decrease Operation and Administrative Costs:

Time is money – especially in the case of publishing content. The older the news, the less valuable it becomes. So, e-Publishers/informediaries invest a significant amount of money on human capital to produce and maintain content. These days, much of that content is derived from the Web and, as discussed earlier in this paper, searching billions of Web pages for pertinent, related information is the single most difficult, time-consuming activity in the e-Publishing industry. By automating the process, most publishers can decrease their operational and administrative costs, and realize significant savings with an investment payout in less than one year.

A small news agency, for example, that manages 300-plus news articles a day, will spend at least $1,000 a day on the time editors will take to categorize, summarize and tag an electronic article. Considering that it takes five minutes to manually categorize, tag and summarize a typical article, the cost added to the article is about $3.00 just to prepare it for electronic distribution. By automating this process, the savings would amount to at least $365,000 each year.
**Increased productivity**

Another advantage of e-content automation technologies is enhancing productivity and efficiency. For editors using the Inxight e-Content Publishing Platform, the benefit is the time they can save on tagging documents, enabling them to spend their time on more strategic editorial responsibilities.

Editors simply don’t scale well – you can only get so much quality work out of a single human. With Inxight’s e-Content Publishing Platform, scalability is built in; the engine is more than capable of sifting through billions of entries in minutes and extracting relevant, personalized information. Also, human editors often miss related content.

Take, for example, a business analyst looking to correlate internal sales data with external market research and consumer trend documents. In essence, she is looking for very closely related information that, on the surface, may seem to have little in common. A human editor may miss the connections entirely. Because editorial decisions are highly interpretive, it’s predictable that even experienced editors will miss or inaccurately categorize and tag relevant information.

Inxight’s e-Content Publishing Platform doesn’t miss a thing. It uses the industry’s most advanced linguistic algorithms, capable of making valid connections between seemingly disparate items. It’s also transnational, capable of culling information in 12 western and four Asian languages.

But productivity savings don’t stop at the publisher or aggregator. Rather, the chain accelerates in value as consumers interact with the information.

**Increased value of information**

Buried under an ever-increasing volume of unfiltered data, readers don’t have the time to search and read every piece of news in hopes of finding that rare, relevant kernel of useful information. Using Inxight’s e-Content Publishing Platform, however, e-Publishers can boost revenues by increasing overall readership by offering a menu of time saving features such as auto-summarization.

In addition, publishers can add enormous value by personalizing information to meet the demographic profiles or preference lists of their subscribers. Studies show that, even though the Web is viewed as an essentially free medium, people are more than willing to pay subscriber fees to get precise, targeted information that saves them significant search time. E-Publishers wanting to stave off the effects of commoditization on their bottom lines can automatically increase the value of their information through personalization.

**Inxight Technology Focus**

Inxight’s architecture model focuses on the four key requirements involved in identifying, sorting and delivering targeted information from the mass of data residing on both public Web sites and private intranets – creation, collection/aggregation, normalization and distribution.

- **Organizing** – The process of automatically classifying content into both topics/subjects and entities (companies, people’s names, etc.)

- **Enriching** – The automation of enriching content by applying metatags into the document that embed the characterization of the document’s topics, key entities, hyperlinks to related information, and summaries using XML technology.

- **Collection/Aggregation** – The process of integrating content from multiple, disparate sources, both internal and external, and organizing it into a body of useful information.
Normalization – The process of processing and refining aggregated information into cohesive search results. Different infomediary sources use different naming conventions and categorizations. For example, a search for auto racing may turn up articles on individual drivers, NASCAR safety regulations, and repaving the Indianapolis Speedway. Normalizing metadata from content means both an intelligent search that recognizes the relationships and contexts of seemingly unrelated articles, as well as rejecting articles that seem to fit search criteria but are only tangentially related.

Data personalization – The process of sending the right information to the right people, in the right format, according to both search criteria and the format preferences of the user – abstracts and summaries for downloading to mobile devices such as PDAs and Internet appliances; full article with graphics for computer users.

Paying close attention to these core requirements yields a system for searching, categorizing and retrieving information that encompasses the 10 keys that allow users to take full advantage of dynamic content.

The Inxight e-Content Publishing Platform

Inxight delivers a best-of-breed content infrastructure known as the Inxight e-Content Publishing Platform that enables content businesses to fulfill delivery promises for quality over quality, faster access to pertinent information (horizontal and vertical information), and personalization. The Inxight e-Content Publishing Platform enables content businesses to:

- Automatically classify and index content, such as news feeds and web sites, into pre-defined subject categories.
- Automatically create executive summaries from the context of each article.
- Automatically create and embed hyperlinks of key concepts.
- Provide dynamic annotation on hyperlinks, allowing users to actively see key 'live' information (such as a company’s current stock price or company overview), or jump to related Web sites such as the company home page, SEC (Edgar) database, a financial news page, or a list of related news articles.
- Actively find similar news content to support the “find more like this” function.
- Create a user-definable thesaurus that translates acronyms, industry terms, company names, name aliases, abbreviations and stock ticker symbols into full titles and names.
- Increase accuracy using patented, linguistic pre-processing engines.
- Store category indexes, summaries and key entities via XML output to an industry standard format, allowing for ease of integration with other enterprise applications.
## Inxight’s Linguistic Technology

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-categorization</td>
<td>Automatically categorizes and organizes documents, such as news feeds, word processing files and email, into pre-defined subject categories.</td>
<td>Speeds search by organizing information into logical directories and folders.</td>
</tr>
<tr>
<td>Auto-index tagging</td>
<td>Automatically embeds category indexes as metadata into each article or file, or into an index repository.</td>
<td>Creates lasting, reusable value to documents.</td>
</tr>
<tr>
<td>Intelligent summaries</td>
<td>Automatically creates intelligent summaries based on the context of an article or file. Configurable – summaries can be defined to have a specific length or genre preference.</td>
<td>Users save time by ‘previewing’ documents while searching.</td>
</tr>
<tr>
<td>Hyperlinks</td>
<td>Automatically create and embed hyperlinks that identifies and highlights 27 key entities including people, company names and ticker symbols, product names and places.</td>
<td>Users save time by previewing key information (e.g. stock price) or by jumping to a related URL via a hyperlink.</td>
</tr>
<tr>
<td>Hyperlink annotation/menus</td>
<td>Extends the hyperlink function to include dynamic 'annotative’ menus, allowing users to “roll-over” a hyperlink and view information or menus related to the hyperlink.</td>
<td>Creates a user-friendly dimension to hyperlinks where the user can decide, using a menu, which place he/she wants to jump to, as well as previewing “live” dynamic data, like a stock price, without having to “jump” around.</td>
</tr>
<tr>
<td>Article similarity</td>
<td>Compares document with others in the knowledge base to discover related or “similar” documents.</td>
<td>Makes discovery of related articles fast and easy.</td>
</tr>
<tr>
<td>Customizable Thesaurus module</td>
<td>Translates acronyms, abbreviations, name aliases and ticker symbols into normalized concepts. Examples: SEC=Securities Exchange Commission, CTO=Chief Technology Officer, President Bill Clinton = William J. Clinton, ORCL=Oracle Corp.</td>
<td>Enhances search and hyperlinking functionalities by taking abstract concepts and normalizing them to a common term for easier navigation and search.</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>Benefit</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Linguistics-based drivers</td>
<td>Uses natural language processing that includes language identification, word stemming, compound word analysis, word and phrase tokenization, noun-phrase concept identification, and part-of-speech tagging. Examples: <strong>Language ID</strong> &quot;Ich bin ein Berliner&quot; = German (Deutsch) &quot;I am a citizen of Berlin&quot; = English <strong>Word Stemming</strong> &quot;selling&quot; = sell; &quot;bought&quot; = buy <strong>Compound Word Analysis</strong> &quot;homeowner&quot; = home owner <strong>Tokenization:</strong> &quot;Mr. Kim, an investor's representative, said, 'The stock is undervalued'.” [mr kim an investor representative said the stock is undervalued] <strong>Noun-Phrase</strong> &quot;The financial analyst reports are listed on the finance Internet website.&quot; [financial analyst report] and [financial Internet website] <strong>Part-of-Speech Tagging</strong> &quot;The merger initiative folded,&quot; [merger:adjective][initiative:noun][folded:verb]</td>
<td>When systems know &quot;what you mean&quot;, it enhances the quality of categorization, hyperlinking, summarization and finding similar/related documents. Moreover, the linguistics aspects helps process text in a way that helps a computer better manage information by knowing more than words – but extracting concepts and content.</td>
</tr>
<tr>
<td>XML output</td>
<td>Indexes and summaries are created using XML.</td>
<td>For IT administrators, developers and architects, they will benefit by using an standard XML encoding to integrate Inxight applications and output into their system infrastructure.</td>
</tr>
<tr>
<td>Java and C/C++ API’s</td>
<td>APIs coded in Java and C/C++ exposes key features in Inxight products.</td>
<td>APIs expose Inxight’s best-of-breed technologies and core functions. The APIs are designed to allow for ease of integration for all primary features of Inxight products, whether you are building and application or adding functionality into an existing system infrastructure.</td>
</tr>
</tbody>
</table>

For more information...

To request more information on Inxight’s linguistic products, please visit [www.inxight.com/about/request_info.html](http://www.inxight.com/about/request_info.html), email Inxight Sales at sales@inxight.com or call 888-414-4949 (US), +44 (0) 1252 761314 (Europe, Africa and the Middle East) or 408-969-7200 (worldwide).