The City University of New York uses Intelligent Content technology for publishing its policies on the Web

Ictect Case Study Series

Summary

Country or Region: United States

Industry: Education

Customer Profile
The City University of New York (CUNY) is the largest urban public university in the United States. The University serves more than 450,000 students in its degree programs and in its adult, continuing, and professional education programs.

Business Situation
The City University of New York needed a system to disseminate its policy documents and streamline policy content management.

Solution
The business solution involved continued use of Microsoft® Word by the Board of Trustees, legal department, and others. The Information Management department, using Ictect's intelligent content technology, XML-based Open Source server, and ISYS software, is now able to provide rich, accessible content on the Web.

Benefits
CUNY estimates that it implemented a streamlined and scalable policy content management system for a fraction of the cost of a commercial system.

Technology Enablers
- Microsoft Word
- Intelligent Content Plug-In (icPlugIn)
- ISYS:Web
- Apache Cocoon

Rich, accessible policy content on the Web

The City University of New York (CUNY) needed a system to disseminate its policy documents and streamline policy content management.

By combining the best commercial off-the-shelf products and open-source software, a team of CUNY staff was able to create a best-of-breed solution. A critical component of this project was providing for ongoing conversion of policy documents from Microsoft Word DOC format to XML.

CUNY met this need using Ictect's patented intelligent content technology. The icPlugIn product provides an MS Word-based interface that could be easily used by nonprogrammers.

Benefits of the new policy document management solution include a streamlined and scalable system at a lower cost than commercially available end-to-end solutions.

The Challenge

As the largest urban public university in the United States, The City University of New York is home to an academic community of stunning diversity. In addition to eleven senior colleges and six community colleges, the University is home to the CUNY Honors College, and the Graduate School and University Center. The University is also home to the Graduate School of Journalism, the CUNY School of Law, and the Sophie Davis School of Biomedical Education. All together, the University serves more than 450,000 students in its degree programs and in its adult, continuing, and professional education programs.

Given the size and complexity of the University, it is imperative that the University community be aware of, and involved in, decision-making processes and their outcomes. As part of the continuing efforts to further integrate the City University of New York, the University undertook the creation of a Web site to disseminate University-wide policy documents. The Policy Documents Site (PDS) was envisioned as a central source for the minutes of the University's Board of Trustees, secondary consolidated policy sources, and interpreting documents.

The Answer

In designing the Policy Documents Site, the PDS team recognized that what was needed was not just a website, but an entire information management system.
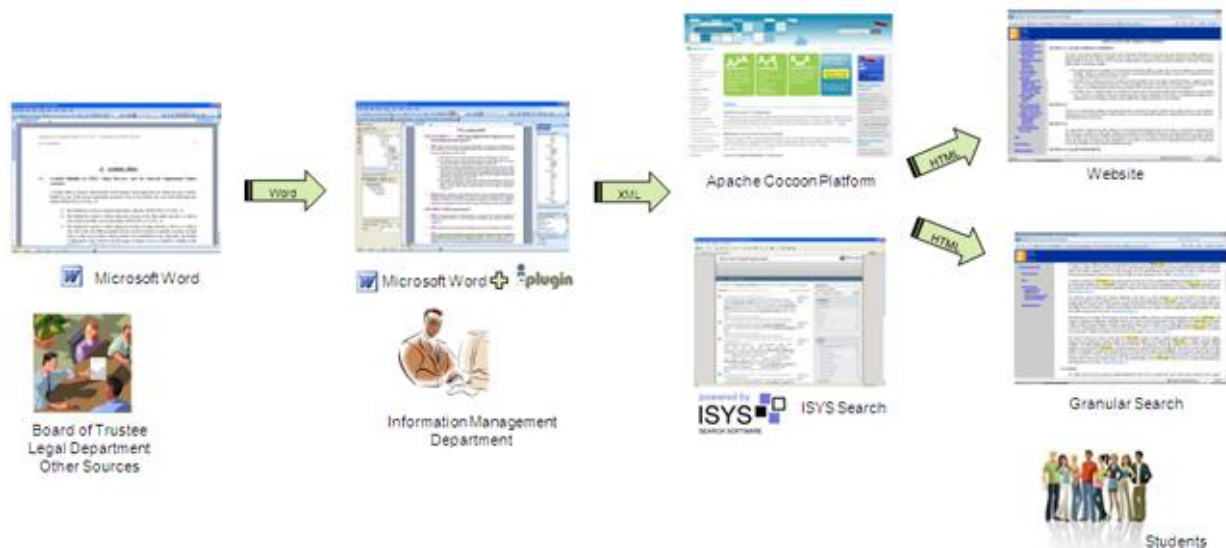
The ultimate expression of this PDS system would be a dynamically generated online interface providing users with multiple modes of access to policy documents updated in real time. The approach eventually chosen for realizing this model was the design and construction of a system combining multiple open-source and commercial software components with human systems. This approach allowed the PDS team to weave together the best purpose-built software available—capitalizing on the stability of established products, while achieving a more agile system than possible with available end-to-end solutions.

Laying the Foundation

The first step in developing the PDS System was the choice of a storage format for the policy documents that would form its foundation. Policy documents from the early 1990s back to 1940 were only available in hard copy and were scanned as searchable image PDFs. Moving forward, however, policy documents were available in Microsoft Word DOC format—allowing for the selection of a format suitable for archiving and dynamic content management. Chief among *the PDS team's concerns were that it should be an open format, platform independent, human* readable, hierarchically structured, and either Web ready or readily made so. XML was the clear choice for such a format and the PDS team began looking for DOC to XML conversion software.

Microsoft Word — Board of Trustee Legal Department Other Sources → Word → Microsoft Word + plugin — Information Management Department → XML → Apache Cocoon Platform / ISYS Search → HTML → Website / Granular Search — Students

the intelligent content company
ictect

Selecting XML Conversion Software

In selecting conversion software, the PDS team's options were limited by the diverse complex hierarchical structures of the policy documents to be converted. To preserve these structures a product had to allow the team full flexibility in defining the structure of XML output files, and mapping elements in input files to elements in output files.

In addition to the requirements defined by the nature of the documents to be converted, the PDS team had to take human factors into consideration. University policy documents are authored and updated by policy experts not by programmers. While the PDS team would set up the conversion process for policy documents, ongoing conversion of updated documents would be performed by these document administrators. This arrangement avoided the need for a document conversion staff, but meant that one of the minimum criteria for any conversion software was a graphical interface. This interface would need to have human readable error reports and make provision for correcting errors in both the input and output files without editing code.

Unfortunately, the PDS team found that a limited range of DOC to XML conversion products was available. A review of publications in the industry generated fewer than a dozen products and most were quickly rejected. The rejected products used simplistic, built-in output templates and conversion mappings without the flexibility needed to handle complex policy documents. Evaluation copies of the remaining four products were acquired and the PDS team compared their performance converting sample policy documents. Based on these tests the PDS team found that only one product, icPlugIn, met the University's needs.

icPlugIn uses XML schema definitions (XSDs) to define the structure of XML output files and is capable of recognizing and adhering to any structure set forth in this flexible standard. The hierarchical elements of the input files are mapped to the elements of the output files using conversion definitions that draw upon the wide range of style and formatting options in Microsoft Word. Because icPlugIn is designed for use with Word, users can work within the same familiar interface that they use to create their documents. On the strength of its testing performance and its features, the PDS team chose icPlugIn to serve as one of the cornerstones of the PDS system. These four programs operate on the underlying policy documents that are the foundation of the PDS system.

Accessibility beyond Search

The ability to dynamically generate individual sections of policy documents for spidering and searching is also used to provide a back-of-the-book-style index.

Linked locaters allow users to follow entries directly to occurrences in the text, but when a document is frequently updated the cost of maintaining them becomes prohibitive. The PDS Team solved this problem using its ability to spider and search individual sections of policy documents.

Policy document indices compiled by human indexers are converted from DOC to XML using icPlugIn, and requested sections of these indices dynamically generated using Cocoon. As part of this dynamic generation, linked locators are generated from policy/item/paragraph-number locators in the XML index files. When a user clicks on a linked locater for a term entry in an index, Cocoon interprets the requested URL and generates a search query. These queries restrict the search to that portion of the document identified by the policy/item/paragraph number in the linked locator. The user is taken directly to the text of the locator-referenced policy/item/paragraph at the first occurrence of the index term with all occurrences of the term highlighted—regardless of pluralization. Index subentries are handled by searching for subentries within a set number of words of the entries—and any higher-level subentries—under which they appear in the index.

the intelligent content company

The Policy Documents Site System

When a new DOC-formatted policy document is to be added to the PDS System, the PDS team creates an XSD and a set of conversion definitions based on a structural analysis. After running test conversions and making any adjustments, the team distributes the XSD and conversion definitions to document administrators for converting updated documents. If the formatting or structure for a DOC-formatted policy document changes on the input end, the PDS team simply modifies its XSD and conversion definitions accordingly.

When the PDS team receives new or updated policy documents in XML format it deploys the files—along with documents in PDF format and processing files—in Apache Tomcat for on-the-fly updates to the Policy Documents Site. Tomcat is the front-end server for the Policy Documents Site and the container for an Apache Cocoon servlet to which incoming URL requests are forwarded. Cocoon is the development framework within which the dynamic content management for the site is performed. Cocoon responds to URL requests by using XML-encoded processing pipelines to return PDFs, redirect requests, and dynamically generate documents containing requested information. When generating dynamic documents Cocoon can extract content from multiple XML documents, transform it using multiple XSL files, and aggregate it with other data feeds. The results can be serialized to multiple formats including JSP forms and HTML policy Web pages for easy online access using any Web browser.

As the final step in the information management process, the Policy Documents Site is spidered using ISYS® Search Software's ISYS: web search server. Rather than spider documents locally on the PDS server, the ISYS spidering agent is pointed at URLs managed dynamically by pipelines inside the Cocoon framework. This setup allows documents to be spidered and searched identically, whether they are static PDF files or HTML web pages dynamically generated from XML files. As a result, large policy documents can be converted to XML as single files while spidering and searching is performed on sections of these documents dynamically generated by Cocoon. Instead of returning one large document of several hundred or more pages, user searches return only those sections that are relevant. The end result is easier updating of policy documents, and the flexibility to provide these documents to users in smaller, more searchable units.

Conclusions

By building an information management system capable of dynamic online content generation, the PDS Team was able to provide the University community with an up-to-date source for policy documents. Using icPlugIn allows the PDS Team to shape the document conversion integral to this system to fit University policy documents and staffing, rather than reshaping documents and staffing to fit software. One month after deployment the Policy Documents Site has become an important resource in policy and legal matters. The University looks forward to long-term savings from this alternative to costly annual editions and monthly updates of policy documents.

For More Information

For more information about Ictect products and services, call 262-898-7568 or visit the Web site at: **www.ictect.com**

the intelligent content company

ictect