Xerox
DocuShare®

White Paper

April 2007

The Xerox DocuShare CPX Extensible Database—Real Time Connection of XML Content

XEROX®

Technology | Document Management | Consulting Services

The Xerox DocuShare CPX Extensible Database– Real Time Connection of XML Content

With the majority of today's business content passing through Internet-based networks, XML has become the de facto language for transferring structured information among many business applications and processes. XML is now the informational fluid for managing the movement of data across an organization in ways not previously possible. Its inherent flexibility enables myriad options to format and structure that information.

But as a consequence, the structural variety of files passed makes them impossible to connect without either a precise matching of the XML structure (schema) or the use of some intermediary translation. Effective integration requires advanced knowledge of the detailed application or process schema, so that connection points can be pre-determined and accommodated in the XML code. Without that "prior alignment," conversion steps are often needed, involving costly and time-consuming human intervention.

Harnessing the true potential of XML as the conduit of informational connectivity requires a seamless mapping of structured content regardless of the source file's XML construction. Previously, this capability did not exist. Now, Xerox DocuShare CPX offers an extensible database (XDB) that enables simple, direct XML-to-XML connection, quickly and automatically linking diverse organizational content to accelerate business processes and productivity.

Over 80% of data within enterprises is estimated to be in unstructured formats like Microsoft Word and Excel as well as Adobe PDF file formats. There are 300 million Excel installations worldwide, 200 million PDF documents on the Web, and 100 million new Microsoft Office documents created every day.

—Informatica, Inc., November 2006

DocuShare CPX Takes XML to a New Level

Unlike many XML information passing systems, the DocuShare CPX extensible database retains an original document (such as a Microsoft Word, Microsoft Excel, Adobe PDF, or Adobe FrameMaker file) while also providing direct access to the information contained within the document. The XDB summarizes that information into XML and then uses the converted XML to extract and share relevant data for other organizational needs, such as quickly creating reports that pull from multiple source documents.

This capability not only applies to new documents created after business processes are defined, but also extends to archived or legacy documents which are already associated with a process. CPX XDB spans all structured information to identify common process touch points, eliminating manual intervention in mapping source document structures. To ensure adherence to established security policies, once information is brought into DocuShare CPX, its access permissions are enforced, whether the information is accessed in XML or in the original source format.

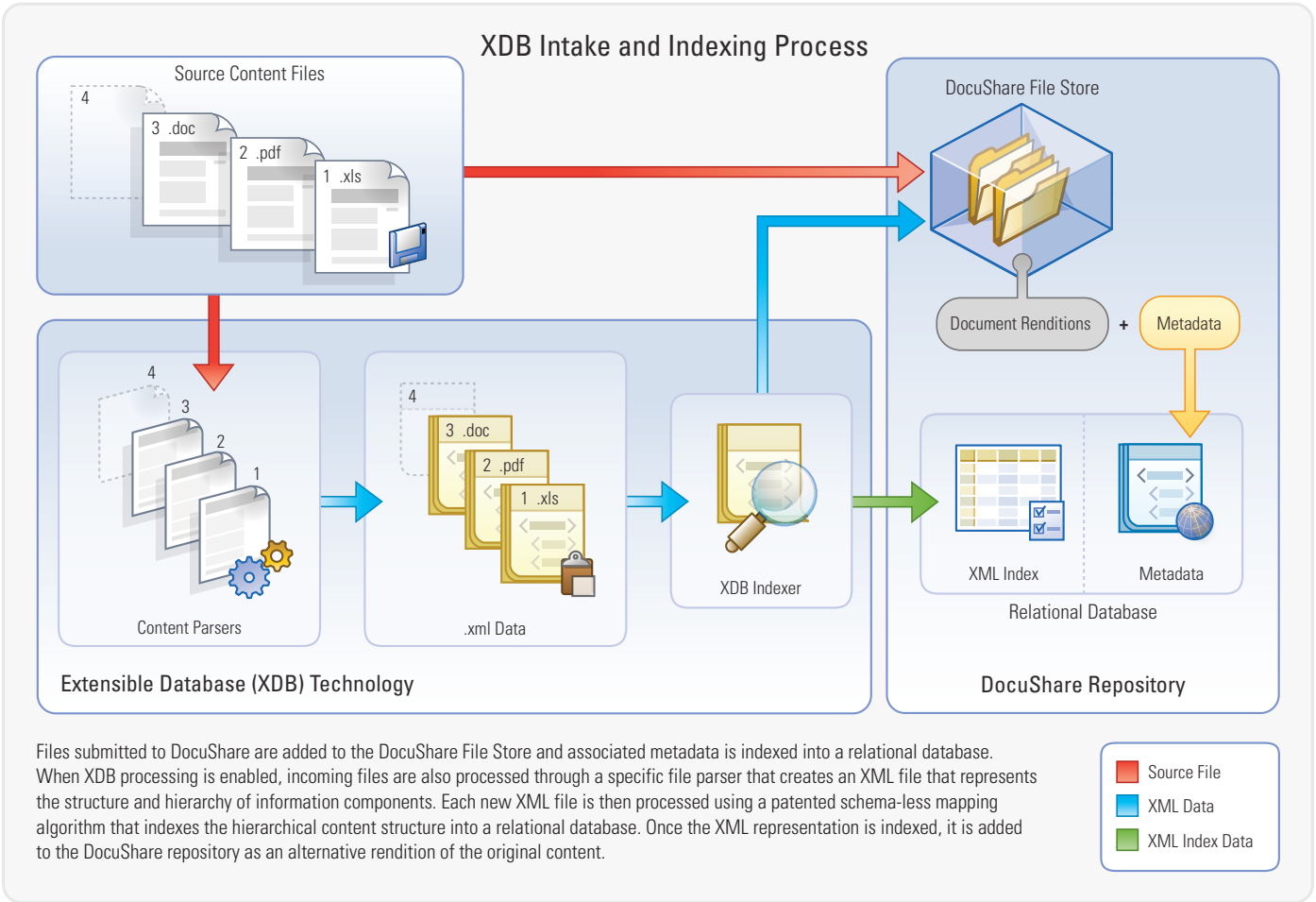
Intake and Indexing—with a Twist

DocuShare's extensible database accomplishes these connectivity goals through its unique intake and indexing process.

The process begins with the source content files, including today's most common formats. With the standard DocuShare CPX content management system, source files are added to the DocuShare repository where they are stored and where metadata is added to facilitate content management.

However, when the XDB is enabled, an additional process on the incoming content is performed in tandem. The original source file is passed through a content parser that creates an associated XML file. The XML file is stored in the DocuShare repository as a second rendition of the original document.

The XML rendition is then passed through the CPX XDB Indexer, a technology used by DocuShare that indexes the content into a relational database management system (either Oracle or Microsoft SQL Server). The resulting XDB index in the database co-exists along with the metadata attached in the standard DocuShare CPX process, becoming part of a flexible DocuShare knowledge network through which users can easily search for and retrieve stored content.



Files submitted to DocuShare are added to the DocuShare File Store and associated metadata is indexed into a relational database. When XDB processing is enabled, incoming files are also processed through a specific file parser that creates an XML file that represents the structure and hierarchy of information components. Each new XML file is then processed using a patented schema-less mapping algorithm that indexes the hierarchical content structure into a relational database. Once the XML representation is indexed, it is added to the DocuShare repository as an alternative rendition of the original content.

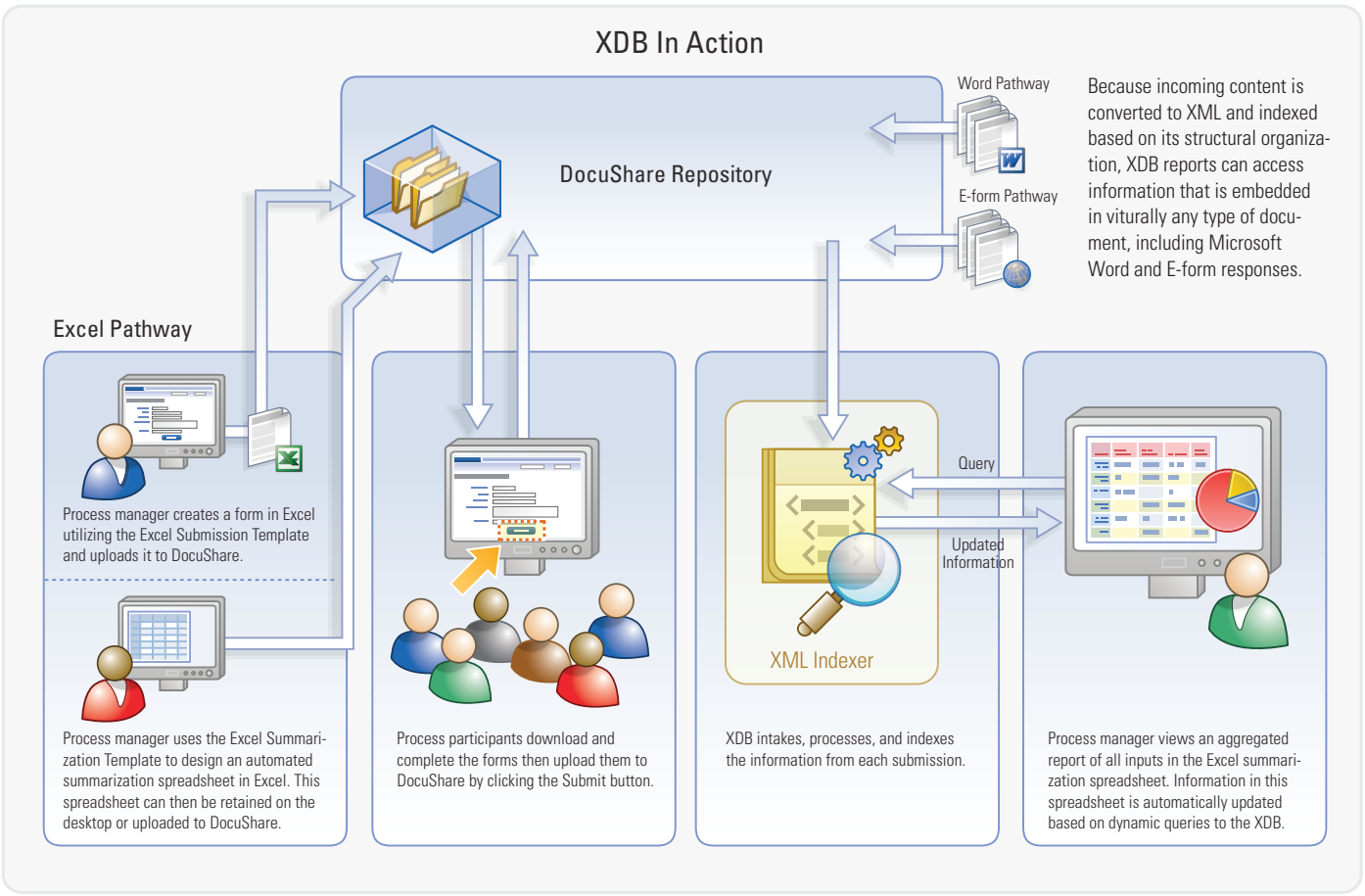
Because the information is indexed based on contextual identifiers, the XDB can easily access information represented in the content and summarize it across documents whenever needed. These optional processes are easily enabled by the DocuShare CPX administrator who specifies what types of content should be subject to XML conversion and XDB indexing.

Find and Repurpose Specific Content Strings

One of greatest strengths of the XDB technology is its ability to find and assimilate specific content components from similarly structured business documents like contracts, presentations, or spreadsheets. The individual components can be retrieved and re-assembled by the XDB to create concise summaries of relevant information across multiple source documents.

The components are found based on the XML context that is associated with them. For instance, a company may use standard contracts created in Microsoft Word as part of a specific business process. Each contract has a unique termination clause as part of its content, which is identified based on textual markers (headlines, bolding, underlining, page position, etc.) and tagged as context during the XML conversion. The XDB has a simple-to-use search function that queries the context information to find all occurrences of the termination heading. It then returns the text in the

> "Enterprises must recognize that the thousands of uncontrolled spreadsheets their employees use every day represent a significant risk. Poorly managed spreadsheets may–through negligence, incompetence or deliberate criminal conduct–result in significant business losses, exposure to legal liability, damage to reputation and unwelcome regulatory attention."
>
> —From Gartner, Symposium/ITXpo 2006, "The Information Explosion and What to Do About It," Toby Bell, October 2006



paragraphs of just that clause for each contract file within a designated DocuShare collection. The retrieved content summary can be either saved to a report format or repurposed, wholly or in part, into another document through a simple cut and paste.

This capability is especially useful for highly complex content structures such as those found in Excel spreadsheets. Excel content can vary from basic names of columns or rows to detailed cell ranges. The XDB content parser identifies spreadsheet content based on range names, attaches XML context data, and then passes it on to the XDB Indexing process and into the RDBMS. The content is then readily retrieved and shared on demand. Because Excel information so frequently drives corporate business processes, the XDB can be a particularly powerful tool for integrating systems around Excel spreadsheets or quickly accessing summarizations of Excel data from disparate sources.

Even further, the extensible database is impartial to the original source format of stored data. Once it is passed through the XDB Indexing process, identically named data from varying source documents and formats, such as from Word and Excel, can be retrieved to the same report. For example, a column labeled 'location of travel' from Excel-based expense reports can be combined with 'location of travel' information contained in standard Word-based sales trip reports.

> Accumulated studies by audit firms since 1998 show that as many as 94% of corporate spreadsheets may have some form of error, ranging from negligible to extremely serious.
>
> —Results of research by R. Panko, "What We Know About Spreadsheet Errors," University of Hawaii, January 2005

Faster, More Accurate Business Intelligence with XML Submission and Summary: Universities Space Research Association

The Universities Space Research Association (USRA), a non-profit research organization chartered to foster cooperative research, development, and education associated with space science and technology, helps the National Aeronautics and Space Administration (NASA) manage its business intelligently. With billions of dollars worth of research and development projects currently underway, certain centers within NASA were facing efforts required to manage information resources for its financial and project performance reports. Managers were required to manually copy and paste detailed financial and project information from many disparate sources into numerous reports. This resulted in valuable time being spent consolidating data rather than analyzing it. For example, one report alone took up to 360 person hours each month to capture and collate the necessary information into useful reports. This manual process also generated a high number of transcription errors—an audit of one $700M program with over 500 mile-stones revealed a 40% discrepancy rate.

USRA addressed this growing problem by creating a performance management tool with NASA that leveraged the XML submission and rendering capabilities built into the DocuShare CPX extensible database. USRA uses XDB as an XML-hub for managing, storing, and synchronizing project data among source documents, including integration with the organization's core systems from Oracle and SAP. The solution enables project managers to automate submission of content through XDB-enabled source documents, such as Excel spreadsheets. XDB then reassembles the XML content as required by each manager into accurate summary documents. $2.2B of internal activity is now managed using the tool.

The resulting time and labor efficiencies have made project performance information available to managers in a much more timely, accurate, and effective manner. By automating the process, report creation time was significantly reduced, from 360 person hours down to 52 for example, and discrepancies were virtually eliminated. Now managers and analysts can spend time actually analyzing and using data rather than consolidating it.

For more information, contact USRA's Research Institute for Advanced Computer Science (**www.riacs.edu**) **info@riacs.edu**.

Would You Like to Learn More?

For more information on DocuShare CPX XDB, please call **1.800.735.7749** or visit **docushare.xerox.com**.

About DocuShare CPX

Xerox DocuShare CPX, a highly intuitive and secure Enterprise Content Management (ECM) application, enables document intensive organizations to dynamically capture, manage, retrieve and distribute information easily, regardless of skill level or location. Part of the Xerox DocuShare family of ECM products, DocuShare CPX customers can significantly improve productivity, streamline business processes, and reduce the time and cost of managing routine business documents and information. Leading the industry in speed of deployment and ease of administration and use, DocuShare CPX significantly reduces installation and complexity, and flexibly extends into an existing infrastructure, resulting in lower total cost of ownership and faster return on investments. Tightly integrated with Xerox Document Centre and WorkCentre Pro, DocuShare CPX can manage both hard copy and electronic content with unsurpassed ease and convenience.

Xerox DocuShare Business Unit
A Division of Xerox Global Services
3400 Hillview Avenue
Palo Alto, California 94304
U.S.A.
1.800.735.7749

© 2007 Xerox Corporation. All rights reserved. Copyright protection claimed includes all forms and matters of copyrightable material and information now allowed by statutory or judicial law or hereinafter granted. Xerox, DocuShare, and WorkCentre are registered trademarks of Xerox Corporation. All other trademarks are the property of their respective companies and are recognized as such.