
Migrating to XML DB2 Native XML with Exegenix conversion technology

Ryan Germann

Gary Robinson

Abstract

Document and content management in industry sectors such as insurance, financial services, healthcare, and government, are ripe for the combination of relational and structured information. IBM is breaking new ground with the Native XML version of its trusted DB2 repository, taking a more holistic approach than its competitors to combine XML and relational systems. XML support is incorporated far deeper in the DB2 data engine than can be achieved via content shredding or BLOB storage, which ensures superior information retrieval with shorter development time and lower overhead. Data storage is centralized for applications that support both structured content modules and tabular data, removing the requirement to develop middleware between distinct data sources.

DB2 XML users will need to populate their databases with high quality XML content in order to realize the full potential of such applications. Organizations with structured XML content already available have a head start. But the vast majority of material exists in unstructured formats such as PDF, Word, and WordPerfect, and the applications used to create new material generate structure that is inadequate for today's discerning XML content applications. Documents created in word processors or output as PDF from other applications need their structural components (not just appearance) to be made explicit before being imported and used effectively.

Almost all methodologies that convert unstructured documents or content into XML rely on "mapping" - the manual designation of combinations of formatting codes as specific XML constructs. But these rules-based conversion maps can change from document to document, depending on the formatting discipline of each document's author, and this approach is time- and resource-intensive both in map development, and in QA and cleanup of the resultant XML.

Exegenix takes a radically different approach. Exegenix's unique intelligent conversion technology uses visual cues to uncover each document's structure automatically, much the same way that humans do. People rarely have problems determining the hierarchical structure of any document they encounter, because they look at a document as a whole, taking into consideration each graphical object's format, position, and context. Exegenix technology does the same thing - it interprets a document's logical structure based on the appearance and position of its components, with no dependency on consistently formatted input. This rules-free XML construction process requires no mapping, no scripting, and no programming, yet produces the most consistently structured and highest quality XML available.

To streamline the DB2 XML deployment process, IBM and Exegenix have incorporated this best of breed conversion technology into the DB2 XML Document Migration Toolkit, which emulates the "extract, transform, load" methodology commonly employed for structured data migration.

This session previews the native XML capabilities of IBM DB2, the legacy document migration process based on Exegenix technology, and illustrates the difference that appropriately-marked up material can make to applications built on DB2 Native XML.

Table of Contents

1. Product Presentation Paper	4
-------------------------------------	---

1. Product Presentation Paper

The author did not prepare a summary of this product presentation for the proceedings.

Biography

Ryan **Germann**

Product Manager

[Exegenix](http://www.exegenix.com) [<http://www.exegenix.com>]

Toronto

Ontario

Canada

Ryan is Product Manager and a founding employee of Exegenix Inc., a company providing content access and conversion technologies. He is involved in market research and the strategic aspects of developing practical applications of Exegenix technology. His professional focus is on facilitating broad adoption of XML content and XML-enabled applications across organizations of all sizes. Since 1995, Ryan has been involved in SGML and XML projects, lending his expertise to both client-side and server-side components. Ryan was previously employed at SoftQuad Software where he was involved in Product Management, Marketing, Web site development and user interface design.

Gary **Robinson**

Senior Software Engineer, IBM Software Solutions, Silicon Valley Lab

[IBM Corporation](http://www.ibm.com) [<http://www.ibm.com>]

San Jose

California

United States of America

Gary Robinson is an Executive Consultant in the Information Management division of IBM's Software Group. He is a member of the team developing native XML support in DB2 and specializes in working with customers and partners using this exciting new technology. Gary has been with IBM for 17 years, both in the UK and the USA, providing technical leadership in Business Intelligence, Information Integration and now XML. Prior to joining IBM, Gary worked in semiconductor manufacturing research in the Netherlands and space research in Oxford UK.