

## Why XML?

*In order to appreciate XML, it is important to understand why it was created. XML was created so that richly structured documents could be used over the web.*

# Why XML?



### What is XML?

XML is a markup language for documents containing structured information.

Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something differ-

ent than content in a figure caption or content in a database table, etc). Almost all documents have some structure.

A markup language is a mechanism to identify structures in a document. The XML specification defines a standard way to add markup to documents.

### Key Benefits of XML

The benefits of using XML as a document representation format are great and apply across all areas of industry. Let's look at what you gain when you adopt XML:

**Content Identification** Perhaps the most important aspect of XML is that text elements are identified, not on the basis of what they look like, but on the basis of what they are—that is, of their significance in the context of a document.

**Databasing** An XML tagged document can be viewed as fielded text. The fielding makes it possible to break documents down to their component parts to any degree of granularity for storage in a document management system. The documents can then be re-assembled in different ways,

and for different audiences, without the need to track multiple document versions.

**Enforced Structure** XML documents are composed in accordance with a DTD, or Schema, which defines the legal tag set for that document type.

**Merging Materials** The uniform structure and lack of internal formatting makes it easy to merge documents into seamless document sets—even if they are coming in from different facilities.

**International Standard** XML is an international standard that is maintained by an independent standards' committee, which means it enjoys widespread support across industry boundaries and gets extensive support from vendors.

## XML Benefits

**Industry Standardization** Many industries have adopted standardized XML DTDs to allow documents to be easily exchanged across different areas of industry.

**Platform Independent** Because “raw” XML consists only of ASCII and Unicode approved characters (the tags themselves are represented in ASCII), XML data can be moved freely between all hardware and operating system platforms that support these character sets.

**Software Independent** There are a wide variety of XML-compliant tools available from many vendors. Because XML is an independent standard, tool sets can be upgraded or changed without fear of data incompatibility.

**Endurance** Appearance-based text representations are constantly changing—making conversion costly when migrating from one software package to another or even when upgrading an existing software package. There is also potential for data loss when performing such conversions. XML, however, is a “permanent” representation. Even as the standard evolves, there is no problem upgrading data.

**Repurpose Data** With XML, formatting is done on a “just in time” basis. As noted, tags identify content, not appearance. Appearance decisions are therefore left until documents are actually published, which means they can easily be modified based on the publication platform.

**XML (eXtensible Markup Language)** is a less complex, more concise dialect of the larger more complex SGML (Standard Generalized Markup Language). In the simplest terms XML uses syntax tags to identify various types of data in a file. For example:

```
<Client>
  <name>Quality Associates Inc</name>
  <street>9017 Red Branch Road</street>
  <city>Columbia</city>
  <state>MD</state>
  <zip>21045</zip>
  <phone>410-884-9100</phone>
</Client>
```

XML makes it very easy for various programs to extract data because the tags conform to particular models.

### Past Choices Not the Best Choices

In the past, organizations that needed their content in XML have only had two choices: manual conversion or error prone automated solutions that try to map a word-processor’s formatting styles to XML. Manual conversion is a slow, expensive, and labor-intensive process. The re-keying and hand-tagging of content not only requires extensive knowledge of XML, but can also introduce both typographic and syntactic errors into the conversion process.

Scripting solutions are no better, and rely on rigid conformance to formatting styles in the original word processor document. However, most authors generally don’t create documents in such a consistent manner, and low-quality XML output can result. Scripting software can also involve costly and lengthy custom programming for each type of document, and for each proprietary application that is used to create those documents.

## QAI's Unique Conversion Workflow

QAI utilizes a newer solution designed to minimize the costs of XML content conversion via a groundbreaking approach that can convert into XML any file that can be printed to PostScript or PDF. We use an automated process that eliminates human involvement for most documents, yet offers opportunities for user-intervention if desired. The results are a flexible framework to support unique conversion workflows, and offers dramatic cost savings over traditional conversion methodologies.

- Unlike manual hand-tagging, our process is accurate, fast, and minimizes human resource requirements, enabling valuable expertise to be employed more effectively elsewhere.
- Unlike scripted conversion, there is no dependency on consistently-applied formatting styles, and no programming expertise required to develop/maintain configuration scripts.

Our conversion solution uses visual cues to uncover a document's structure, much the same way that humans do. The valid XML output file not only maintains the original document's content and logical structure, but also retains all relevant formatting information.

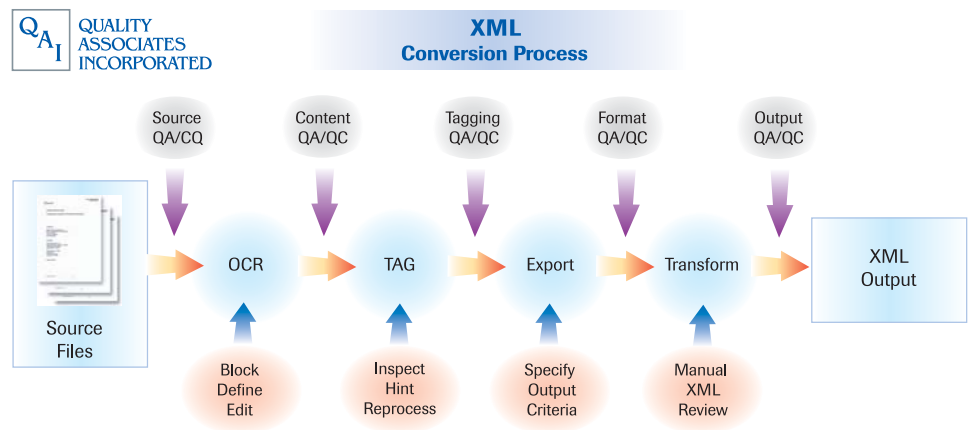
Once documents have been sorted/selected for conversion, they enter the OCR Process.

## QAI's Conversion Process

The initial step in the process is to "block" the document. This step involves drawing color-coded boxes around sections of each page to define text, tables, and images. Once "blocking" has been validated the results of OCRing the text areas will go through a clean-up process. The output of the OCR Process are files that have gone through a text recognition process, that were then "cleaned-up", and then are saved in the PDF "Formatted Text and Graphics" format (also called PDF Normal). At this point the documents are ready for the XML conversion process.

All content goes through four processes within the XML Processing Engine in order to uncover the document's structure and generate valid XML output.

## Four Step Process



- 1** The first step in the process analyzes a document's PostScript or PDF representation to extract all information about the appearance of the document. This includes the characters in the document and their typography, and any other visual objects. Because the process extracts text directly from the input datastream, all content is accurately retained during conversion.
- 2** The next step in the process identifies the basic building blocks of document structure, including many important visual cues, and the large-scale layout areas of the page.
- 3** The third step places these now identified building blocks into a tree structure. This phase identifies sections, paragraphs, quotes, lists, tables, footnotes, and other graphical objects, and forms a complete, cohesive, internal representation of the structured document.
- 4** The final step in the process uses the internal representation of the document, from step three, to export an XML file that not only presents the document's content and logical structure but also retains all relevant formatting information.

This exported XML file is now ready to be transformed to meet the customer requirements.

*QAI's Mission: To utilize technology, science and our passion for quality to assure that all of our clients benefit from our experience.*

QAI



**Quality Associates, Inc.**

Quality Associates, Inc. (QAI) is a complete solutions provider. We present our clients with cutting-edge solutions that address document management issues in an ever-changing marketplace. QAI has extensive experience with federal and state agencies. We

have the ability to understand and address customer confidentiality/privacy issues. For clients in need of XML conversion services, QAI operates a state-of-the-art secured, onshore conversion operation located in the Washington, DC metropolitan area.



**QUALITY  
ASSOCIATES  
INCORPORATED**

The Chesapeake Building  
9017 Red Branch Road | Suite 102  
Columbia MD 21045

T 410.884.9100  
800.488.3547  
F 410.884.9122