

An Introduction to XML

Péter Jeszenszky

September 11, 2025

What is XML? (1)

- A general purpose markup language.
- Was born in the second half of the 1990's.
- A key technology that is widely used in the IT industry.

What is XML? (2)

- **In the strict sense:** A syntax for representing structured documents that enables the automatic processing of these documents (i.e., an electronic document format).
- **In the broadest sense:** A set of related specifications that are also called collectively as **the XML family**.

Predecessor

Its predecessor is SGML:

- An ISO standard for electronic documents (1980's):
 - ISO 8879:1986 Information processing – Text and office systems – Standard Generalized Markup Language (SGML)
- Is very complex and is not suitable to be used widely on the web.
- XML is based on SGML, it is compatible with it (is a subset of it), but is far simpler to use.

Markup Languages

- Markup languages are computer languages for annotating text.
- They allow the association of metadata with parts of text in a clearly distinguishable way.
- Examples:
 - AsciiDoc <https://asciidoc.org/>
 - TeX, LaTeX <https://www.latex-project.org/>
 - Markdown <https://daringfireball.net/projects/markdown/>
 - troff (man pages) <https://www.gnu.org/software/groff/>
 - XML <https://www.w3.org/XML/>
 - Wikitext <https://en.wikipedia.org/wiki/Help:Wikitext>

Structured Documents

- Are composed of different structural components, such as titles, chapters, sections, paragraphs, comments, and tables.
- They must be processable automatically, to allow it, markup languages provide means to identify structural components.
- They are often not documents in the conventional sense.

XML as a Markup Language

Structural components in documents are identified by tags, such as:

```
<author>Sir Arthur Conan Doyle</author>
```

```
<title xml:lang="en">The Hound of the Baskervilles</title>
```

Naming

- XML stands for Extensible Markup Language.
- It is extensible, because it does not use a predefined set of tags for identifying structural components, instead, it provides a mechanism for defining such sets of tags.

Meta Markup Language

- Provides a means for defining markup languages, thus, it can be considered as a meta markup language.

Comparison of XML and HTML

- XML:
 - No predefined set of tags
 - Its purpose is to describe data
 - Is used as a data exchange format
- HTML:
 - Uses a predefined set of tags
 - Its purpose is to present information
 - Is a presentation language
 - Can be considered as a special application of XML (XHTML)

Advantages

- Simplicity
 - XML documents are ordinary text files that can be created and manipulated with any text editor.
- Openness
- Vendor independence
- Platform independence
- A universal data exchange format
- Extensive infrastructure
- A de-facto standard in the industry

Disadvantages

Nevertheless, it is important and has to be lived with.

- Verbose and cumbersome to use syntax
- Highly inefficient storage
- Complexity
 - There are several XML-related specifications.

Document-Centric XML

- Documents are composed of text intermingled with markup constructs.
- Highly varied document structure.
- The ordering of elements is important.
- Such documents are primarily intended to be consumed by humans.
- For example, XHTML is such an XML format.

Data-Centric XML

- Documents are composed of a large number of data elements.
- Less random document structure.
- The ordering of elements is less important.
- Such documents are primarily intended to be processed by computers.
- For example, SVG is such an XML format.

Alternative

JSON (JavaScript Object Notation) <https://www.json.org/>

- Lightweight, textual, and platform-independent data exchange format.
- Used for representing structured data.
- For this purpose, it is an alternative to XML.
- Provides the same advantages as XML but without its disadvantages.

Standard

- [Extensible Markup Language \(XML\) 1.0 \(Fifth Edition\)](#). W3C Recommendation, 26 November 2008.
 - This is the widely used standard, although there exists XML 1.1.
- [Extensible Markup Language \(XML\) 1.1 \(Second Edition\)](#). W3C Recommendation, 16 August 2006.
 - It is not widely used.

File Properties

- File extension: `.xml`
- IANA media types: `application/xml`, `text/xml`
- Many XML formats have their own file extension and media type.
 - The structured syntax suffix `+xml` at the end of a subtype indicates that the format is based on XML.
 - Examples: `application/xhtml+xml`, `image/svg+xml`, `model/x3d+xml`

The XML Family (1)

- **Specifications directly related to XML itself:**
 - Extend the capabilities of XML.
 - Provide means to express constraints on the structure and content of XML documents (XML schema languages).
 - Provide means to extract information from XML documents (query languages).
 - Provide means to transform XML documents to other forms (transformation languages).

The XML Family (2)

- **Applications:** application-domain specific XML formats
 - Serving digital content (e.g., Atom, DocBook, MathML, OSM XML, RSS, SVG, X3D, XHTML)
 - Communication (e.g., XMPP)
 - Storing configuration data (e.g., Apache Maven, FXML)
 - Semantic Web (e.g., OWL, RDF, XMPP)
- **Application programming interfaces (APIs):** provide means to process XML documents in programming languages (e.g., DOM, JAXB, JAXP, JDOM, SAX, StAX)

Specifications Extending the Capabilities of XML

- [Associating Style Sheets with XML documents 1.0 \(Second Edition\)](#). W3C Recommendation, 28 October 2010.
- [Namespaces in XML 1.0 \(Third Edition\)](#). W3C Recommendation, 8 December 2009.
- [XML Base \(Second Edition\)](#). W3C Recommendation, 28 January 2009.
- [XML Inclusions \(XInclude\) Version 1.0 \(Second Edition\)](#). W3C Recommendation, 15 November 2006.
- [XML Linking Language \(XLink\) Version 1.1](#). W3C Recommendation, 6 May 2010.

XML Schema Languages (1)

- Provide means to express constraints on the structure and content of XML documents.
- An XML schema defines a set of documents that are also called instances.
- An XML document that conforms to a schema is said to be valid.
- The process of conformance checking is called validation.

XML Schema Languages (2)

Contemporary XML schema languages:

- Document Type Definition (DTD): is part of the XML specification
- W3C XML Schema <https://www.w3.org/XML/Schema>
- RELAX NG <https://relaxng.org/>
- Schematron <https://schematron.com/>

Query Languages

- [XML Path Language \(XPath\) Version 1.0](#). W3C Recommendation, 16 November 1999.
- [XML Path Language \(XPath\) 3.1](#). W3C Recommendation, 21 March 2017.
- [XQuery 3.1: An XML Query Language](#). W3C Recommendation, 21 March 2017.

Transformation Languages

- [XSL Transformations \(XSLT\) Version 1.0](#). W3C Recommendation, 16 November 1999.
- [XSL Transformations \(XSLT\) Version 3.0](#). W3C Recommendation, 8 June 2017.
- [XQuery 3.1: An XML Query Language](#). W3C Recommendation, 21 March 2017.

Invisible XML (1)

- Invisible XML (ixml) is a method for treating non-XML documents as if they were XML, enabling authors to write documents and data in a format they prefer while providing XML for processes that are more effective with XML content.
 - It can be used for context-free languages.
- Webhely: <https://invisiblexml.org/>
- Tároló: <https://github.com/invisibleXML/ixml>

Invisible XML (2)

- Specifications:
 - [Invisible XML Specification](#). Community Group Editorial Draft, 1 September 2025.
 - [Invisible XML Specification \(W3C\)](#). Final Community Group Report, 12 December 2023.
- Tutorial:
 - Steven Pemberton. [Invisible XML \(ixml\) Tutorial](#).
- Examples:
 - <https://github.com/invisibleXML/ixml/tree/master/samples>

Invisible XML (3)

Implementations:

- Aparecium (license: GPLv3) <https://github.com/cmsmcq/Aparecium>
- CoffeePot (license: MIT License) <https://coffeepot.nineml.org/>
<https://github.com/nineml/coffeepot>
- Markup Blitz (license: Apache License 2.0)
<https://github.com/GuntherRademacher/markup-blitz>

Invisible XML (4)

- ixml support was introduced in XPath 4.0 and XQuery 4.0.
 - See the `fn:invisible-xml` built-in function.
- BaseX support: https://docs.basex.org/main/Invisible_XML

Editors (1)

Free and open source software:

- Visual Studio Code (platform: Linux, macOS, Windows; license: MIT License) <https://code.visualstudio.com/>
<https://github.com/Microsoft/vscode>
 - Recommended extension: XML <https://marketplace.visualstudio.com/items?itemName=redhat.vscode-xml>
<https://github.com/redhat-developer/vscode-xml>
 - Documentation:
<https://github.com/redhat-developer/vscode-xml/tree/main/docs>

Editors (2)

Non-free software:

- `<oXygen/>` XML Editor (platform: Linux, macOS, Windows)
<https://www.oxygenxml.com/>
 - See: https://www.oxygenxml.com/xml_editor/xml_editing.html
- IntelliJ IDEA (platform: Linux, macOS, Windows)
<https://www.jetbrains.com/idea/>
 - See: <https://www.jetbrains.com/help/idea/working-with-xml.html>
- XMLSpy XML Editor (platform: Windows)
<https://www.altova.com/xmlspy-xml-editor>