**XML**

Bách khoa toàn thư mở Wikipedia

Bước tới: [menu](http://vi.wikipedia.org/wiki/XML#column-one), [tìm kiếm](http://vi.wikipedia.org/wiki/XML#searchInput)

**XML** (viết tắt từ [tiếng Anh](http://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh) *eXtensible Markup Language*, "Ngôn ngữ Đánh dấu Mở rộng") là [ngôn ngữ đánh dấu](http://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_%C4%91%C3%A1nh_d%E1%BA%A5u) với mục đích chung do [W3C](http://vi.wikipedia.org/wiki/W3C) đề nghị, để tạo ra các ngôn ngữ đánh dấu khác. Đây là một tập con đơn giản của [SGML](http://vi.wikipedia.org/w/index.php?title=SGML&action=edit&redlink=1), có khả năng mô tả nhiều loại [dữ liệu](http://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u) khác nhau. Mục đích chính của XML là đơn giản hóa việc chia sẻ dữ liệu giữa các hệ thống khác nhau, đặc biệt là các hệ thống được kết nối với [Internet](http://vi.wikipedia.org/wiki/Internet). Các ngôn ngữ dựa trên XML (thí dụ: [RDF](http://vi.wikipedia.org/w/index.php?title=RDF&action=edit&redlink=1), [RSS](http://vi.wikipedia.org/wiki/RSS), [MathML](http://vi.wikipedia.org/wiki/MathML), [XHTML](http://vi.wikipedia.org/wiki/XHTML), [SVG](http://vi.wikipedia.org/wiki/SVG), [GML](http://vi.wikipedia.org/w/index.php?title=GML&action=edit&redlink=1) và [cXML](http://vi.wikipedia.org/w/index.php?title=CXML&action=edit&redlink=1)) được định nghĩa theo cách thông thường, cho phép các chương trình sửa đổi và kiểm tra hợp lệ bằng các ngôn ngữ này mà không cần có hiểu biết trước về hình thức của chúng.

|  |
| --- |
| **Mục lục**  [[ẩn](javascript:toggleToc())]   * [1 Lịch sử](http://vi.wikipedia.org/wiki/XML#L.E1.BB.8Bch_s.E1.BB.AD) * [2 Đặc điểm](http://vi.wikipedia.org/wiki/XML#.C4.90.E1.BA.B7c_.C4.91i.E1.BB.83m) * [3 Sơ lược về cú pháp](http://vi.wikipedia.org/wiki/XML#S.C6.A1_l.C6.B0.E1.BB.A3c_v.E1.BB.81_c.C3.BA_ph.C3.A1p) * [4 Liên kết ngoài](http://vi.wikipedia.org/wiki/XML#Li.C3.AAn_k.E1.BA.BFt_ngo.C3.A0i)   + [4.1 Tiếng Việt](http://vi.wikipedia.org/wiki/XML#Ti.E1.BA.BFng_Vi.E1.BB.87t)   + [4.2 Tiếng Anh](http://vi.wikipedia.org/wiki/XML#Ti.E1.BA.BFng_Anh)     - [4.2.1 Đặc tả](http://vi.wikipedia.org/wiki/XML#.C4.90.E1.BA.B7c_t.E1.BA.A3)     - [4.2.2 Tài liệu cơ bản](http://vi.wikipedia.org/wiki/XML#T.C3.A0i_li.E1.BB.87u_c.C6.A1_b.E1.BA.A3n)   + [4.3 Trình soạn thảo XML](http://vi.wikipedia.org/wiki/XML#Tr.C3.ACnh_so.E1.BA.A1n_th.E1.BA.A3o_XML)   + [4.4 XML certification](http://vi.wikipedia.org/wiki/XML#XML_certification)   + [4.5 Trình phân tích cú pháp XML](http://vi.wikipedia.org/wiki/XML#Tr.C3.ACnh_ph.C3.A2n_t.C3.ADch_c.C3.BA_ph.C3.A1p_XML)   + [4.6 Danh sách thư XML](http://vi.wikipedia.org/wiki/XML#Danh_s.C3.A1ch_th.C6.B0_XML) |

**[**[**sửa**](http://vi.wikipedia.org/w/index.php?title=XML&action=edit&section=1)**] Lịch sử**

Vào giữa những năm 1990, các chuyên gia SGML đã có kinh nghiệm với [World Wide Web](http://vi.wikipedia.org/wiki/WWW) (vẫn còn khá mới vào thời đó). Họ tin tưởng rằng SGML có thể cung cấp giải pháp cho các vấn đề mà Web đang gặp phải. Jon Bosak đưa ra ý kiến W3C nên tài trợ một chương trình mang tên "SGML trên Web".

**[**[**sửa**](http://vi.wikipedia.org/w/index.php?title=XML&action=edit&section=2)**] Đặc điểm**

XML cung cấp một phương tiện dùng văn bản (*text*) để mô tả thông tin và áp dụng một cấu trúc kiểu cây cho thông tin đó. Tại mức căn bản, mọi thông tin đều thể hiện dưới dạng text, chen giữa là các [thẻ đánh dấu](http://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_%C4%91%C3%A1nh_d%E1%BA%A5u) (*markup*) với nhiệm vụ ký hiệu sự phân chia thông tin thành một cấu trúc có thứ bậc của các *dữ liệu ký tự*, các *phần tử* dùng để chứa dữ liệu, và các *thuộc tính* của các phần tử đó. Về mặt đó, XML tương tự với các [biểu thức S](http://vi.wikipedia.org/w/index.php?title=Bi%E1%BB%83u_th%E1%BB%A9c_S&action=edit&redlink=1) (*S-expression*) của ngôn ngữ lập trình [LISP](http://vi.wikipedia.org/w/index.php?title=LISP&action=edit&redlink=1) ở chỗ chúng đều mô tả các cấu trúc cây mà trong đó mỗi nút có thể có một danh sách tính chất của riêng mình.

Đơn vị cơ sở của XML là các [*ký tự*](http://vi.wikipedia.org/w/index.php?title=K%C3%BD_t%E1%BB%B1&action=edit&redlink=1) theo định nghĩa của [Universal Character Set](http://vi.wikipedia.org/w/index.php?title=Universal_Character_Set&action=edit&redlink=1) (*Bộ ký tự toàn cầu*). Các ký tự được kết hợp theo các tổ hợp chuỗi hợp lệ để tạo thành một *tài liệu* XML. Tài liệu này gồm một hoặc nhiều [thực thể](http://vi.wikipedia.org/w/index.php?title=Th%E1%BB%B1c_th%E1%BB%83_SGML&action=edit&redlink=1), mỗi thực thể thường là một phần nào đó của các ký tự thuộc tài liệu, được mã hóa dưới dạng một chuỗi các [bit](http://vi.wikipedia.org/wiki/Bit) và lưu trữ trong một [tệp văn bản](http://vi.wikipedia.org/w/index.php?title=T%E1%BB%87p_v%C4%83n_b%E1%BA%A3n&action=edit&redlink=1) (*text file*).

Các tệp XML có thể dùng cho nhiều loại [dữ liệu đa phương tiện](http://vi.wikipedia.org/wiki/MIME). [RFC3023](http://www.ietf.org/rfc/rfc3023.txt) định nghĩa các loại "application/xml" và "text/xml", với ý rằng dữ liệu được biểu diễn bằng XML mà không nói gì đến ngữ nghĩa của dữ liệu.

Sự phổ biến của các [phần mềm soạn thảo văn bản](http://vi.wikipedia.org/w/index.php?title=Ph%E1%BA%A7n_m%E1%BB%81m_so%E1%BA%A1n_th%E1%BA%A3o_v%C4%83n_b%E1%BA%A3n&action=edit&redlink=1) (*word processor*) đã hỗ trợ việc soạn thảo và bảo trì tài liệu XML một cách nhanh chóng. Trước XML, có rất ít ngôn ngữ mô tả dữ liệu với các đặc điểm đa năng, thân thiện với giao thức Internet, dễ học và dễ tạo. Thực tế, đa số các định dạng trao đổi dữ liệu thời đó đều chuyện dụng, có tính độc quyền, và có định dạng nhị phân (chuỗi bit thay vì chuỗi ký tự) khó dùng chung giữa các ứng dụng phần mềm khác nhau hay giữa các hệ nền (platform) khác nhau. Việc tạo và bảo trì trên các trình soạn thảo thông dụng lại càng khó khăn.

Bằng cách cho phép các tên dữ liệu, cấu trúc thứ bậc được phép, và ý nghĩa của các phần tử và thuộc tính có tính chất mở và có thể được định nghĩa bởi một [giản đồ](http://vi.wikipedia.org/w/index.php?title=Gi%E1%BA%A3n_%C4%91%E1%BB%93_XML&action=edit&redlink=1) tùy biến được, XML cung cấp một cơ sở cú pháp cho việc tạo lập các ngôn ngữ đánh dấu dựa XML theo yêu cầu. Cú pháp chung của các ngôn ngữ đó là cố định — các tài liệu phải tuân theo các quy tắc chung của XML, bảo đảm rằng tất cả các phần mềm hiểu XML ít ra cũng phải có khả năng đọc (*phân tích cú pháp* - *parse*) và hiểu bố cục tương đối của thông tin trong các tài liệu đó. Giản đồ chỉ bổ sung một tập các ràng buộc cho các quy tắc cú pháp. Các giản đồ thường hạn chế tên của phần tử và thuộc tính và các cấu trúc thứ bậc được phép, ví dụ, chỉ cho phép một phần tử tên 'ngày sinh' chứa một phần tử tên 'ngày' và một phần tử có tên 'tháng', mỗi phần tử phải chứa đúng một ký tự. Đây là điểm khác biệt giữa XML và [HTML](http://vi.wikipedia.org/wiki/HTML). HTML có một bộ các phần tử và thuộc tính không mềm dẻo, chỉ có một tác dụng và nói chung là không thể dùng cho mục đích khác.

XML không hạn chế về việc nó được sử dụng như thế nào. Mặc dù XML về cơ bản là dạng text, các phần mềm với chức năng [trừu tượng hóa](http://vi.wikipedia.org/w/index.php?title=Tr%E1%BB%ABu_t%C6%B0%E1%BB%A3ng_h%C3%B3a_(khoa_h%E1%BB%8Dc_m%C3%A1y_t%C3%ADnh)&action=edit&redlink=1) nó thành các định dạng khác giàu thông tin hơn đã nhanh chóng xuất hiện, quá trình trừu tượng hóa này được thực hiện chủ yếu qua việc sử dụng các giản đồ định hướng kiểu dữ liệu (*datatype-oriented schema*) và khuôn mẫu [lập trình hướng đối tượng](http://vi.wikipedia.org/wiki/L%E1%BA%ADp_tr%C3%ACnh_h%C6%B0%E1%BB%9Bng_%C4%91%E1%BB%91i_t%C6%B0%E1%BB%A3ng) (mà trong đó, mỗi tài liệu XML được thao tác như là một đối tượng). Những phần mềm như vậy có thể coi XML như là dạng text đã được tuần tự hóa chỉ khi nó cần truyền dữ liệu qua mạng.

**[**[**sửa**](http://vi.wikipedia.org/w/index.php?title=XML&action=edit&section=3)**] Sơ lược về cú pháp**

Cú pháp XML cơ bản cho một phần tử là

**<t**ên thuộc\_tính="giá trị"**>**nội dung**</t**ên**>**

Dưới đây là ví dụ về một công thức nấu ăn viết bằng XML:

1. **<?xml** version="1.0" encoding="UTF-8"**?>**
2. **<c**ông\_thức\_nấu\_ăn tên="bánh mì" thời\_gian\_chuẩn\_bị="5 phút" thời\_gian\_nấu="3 tiếng"**>**
3. **<title>**Bánh mì cơ bản**</title>**
4. **<nguy**ên\_liệu lượng="3" đơn\_vị="ca"**>**Bột mì**</nguy**ên\_liệu**>**
5. **<nguy**ên\_liệu lượng="7" đơn\_vị="gram"**>**Men**</nguy**ên\_liệu**>**
6. **<nguy**ên\_liệu lượng="1.5" đơn\_vị="ca" trạng\_thái="ấm"**>**Nước**</nguy**ên\_liệu**>**
7. **<nguy**ên\_liệu lượng="1" đơn\_vị="thìa cà phê"**>**Muối**</nguy**ên\_liệu**>**
8. **<ch**ỉ\_dẫn**>**
9. **<b**ước**>**Trộn tất cả các nguyên liệu với nhau và nhào kĩ**</b**ước**>**
10. **<b**ước**>**Phủ một mảnh vải, ủ một tiếng đồng hồ trong phòng ấm.**</b**ước**>**
11. **<b**ước**>**Nhào lại, đổ vào khuôn, cho vào lò nướng.**</b**ước**>**
12. **</ch**ỉ\_dẫn**>**
13. **</c**ông\_thức\_nấu\_ăn**>**

Dòng đầu tiên là **Khai báo XML** (*XML declaration*): đó là một dòng không bắt buộc, với nhiệm vụ thông báo phiên bản XML đang được sử dụng (thường là phiên bản 1.0), và còn có thể chứa thông tin về mã hóa ký tự và các phụ thuộc bên ngoài.

Phần còn lại của tài liệu này chứa các *phần tử* lồng nhau, một số phần tử trong đó có các *thuộc tính* và *nội dung*. Một **phần tử** thường bao gồm hai thẻ (*tag*), một *thẻ bắt đầu* và một *thẻ kết thúc*, có thể bao quanh văn bản và các phần tử khác. **Thẻ bắt đầu** bao gồm một cái tên đặt trong một cặp ngoặc nhọn, như "<bước>"; **thẻ kết thúc** bao gồm chính cái tên đó đặt trong một cặp ngoặc nhọn, với một dấu gạch chéo đứng trước, như "</bước>". **Nội dung** của phần tử là tất cả những gì nằm giữa thẻ bắt đầu và thẻ kết thúc, bao gồm văn bản và các phần tử (con) khác. Dưới đây là một phần tử XML hoàn chỉnh, với thẻ bắt đầu, nội dung văn bản, và thẻ kết thúc:

1. **<b**ước**>**Nhào lại, đổ vào khuôn, c dgdsho vào lò nướng.**</b**ước**>**

Bên cạnh nội dung, một phần tử có thể chứa các **thuộc tính** — các cặp tên - giá trị được đặt trong thẻ bắt đầu, ngay sau tên phần tử. Giá trị của thuộc tính phải được đặt trong cặp nháy đơn hoặc nháy kép, mỗi tên thuộc tính chỉ được xuất hiện một lần trong mỗi phần tử.

1. **<nguy**ên\_liệu lượng="3" đơn\_vị="ca"**>**Bột mì**</nguy**ên\_liệu**>**

Trong ví dụ này, phần tử *nguyên\_liệu* có hai thuộc tính: *lượng* với giá trị "3", và *đơn vị* với giá trị "ca". Trong cả hai trường hợp, cũng như tên và nội dung của các phần tử, tại cấp độ đánh dấu, tên và giá trị của các thuộc tính cũng chỉ là dữ liệu text — các giá trị "3" và "ca" không phải một số lượng và một đơn vị đo lường mà chỉ là các chuỗi ký tự mà tác giả tài liệu có thể dùng để *biểu diễn* những thứ đó.

Ngoài văn bản, các phần tử còn có thể chứa các phần tử khác:

1. **<ch**ỉ\_dẫn**>**
2. **<b**ước**>**Trộn tất cả các nguyên liệu với nhau và nhào kĩ**</b**ước**>**
3. **<b**ước**>**Phủ một mảnh vải, ủ một tiếng đồng hồ trong phòng ấm.**</b**ước**>**
4. **<b**ước**>**Nhào lại, đổ vào khuôn, cho vào lò nướng.**</b**ước**>**
5. **</ch**ỉ\_dẫn**>**

Trong đó, phần tử *chỉ\_dẫn* chứa ba phần tử *bước*. XML đòi hỏi rằng các phần tử phải được lồng nhau một cách đúng đắn — các phần tử không được có phần xen vào nhau. Ví dụ, đoạn dưới đây không phải XML định dạng đúng (*well-formed XML*) vì các phần từ *em* và *strong* xen vào nhau:

1. *<!-- SAI! ĐỊNH DẠNG KHÔNG ĐÚNG! -->*
2. **<p>**Normal **<em>**emphasized **<strong>**strong emphasized**</em>** strong**</strong></p>**

Mỗi tài liệu XML phải có đúng một **phần tử gốc** tại bậc trên cùng (còn gọi là *phần tử văn bản*), do đó đoạn sau cũng sẽ là một tài liệu XML định dạng sai:

1. **<?xml** version="1.0" encoding="UTF-8"**?>**
2. *<!-- SAI! ĐỊNH DẠNG KHÔNG ĐÚNG! -->*
3. **<**đồ vật**>**Đồ vật thứ nhất**</**đồ vật**>**
4. **<**đồ vật**>**Đồ vật thứ hai**</**đồ vật**>**

XML cung cấp cú pháp đặc biệt để biểu diễn một phần tử với nội dung rỗng. Thay vì viết một thẻ bắt đầu và một thẻ kết thúc ngay sau đó, tài liệu có thể chứa **thẻ phần tử rỗng** mà trong đó dấu gạch chéo đứng ngay sau tên phần tử. Hai ví dụ sau là tương đương về chức năng:

**<foo></foo>**  
**</div>**  
:**<source** lang="xml" enclose="div"**>**  
**<foo** **/>**

XML cung cấp hai phương pháp biểu diễn các ký tự đặc biệt: các *tham chiếu thực thể* (*entity reference*) và các [*tham chiếu ký tự số*](http://vi.wikipedia.org/w/index.php?title=Tham_chi%E1%BA%BFu_k%C3%BD_t%E1%BB%B1_s%E1%BB%91&action=edit&redlink=1) (*numeric character reference*).

Trong XML, một **thực thể** (*entity*) là một thân dữ liệu được đặt tên với dữ liệu thường là text, chẳng hạn một ký tự đặc biệt.

Một **tham chiếu thực thể** là một ký hiệu đại diện cho thực thể đó. Nó bao gồm tên của thực thể với dấu ("&") đứng trước và một [dấu chấm phảy](http://vi.wikipedia.org/w/index.php?title=D%E1%BA%A5u_ch%E1%BA%A5m_ph%E1%BA%A3y&action=edit&redlink=1) (";") đứng sau. XML có năm thực thể đã được khai báo trước:

* &amp; (&)
* &lt; (<)
* &gt; (>)
* &apos; (')
* &quot; (")

Dưới đây là một ví dụ sử dụng một thực thể XML khai báo trước để biểu diễn dấu & trong tên "AT&T":

**<t**ên-công-ty**>**AT&amp;T**</t**ên-công-ty**>**

Nếu cần khai báo thêm các thực thể khác, việc đó được thực hiện tại [DTD](http://vi.wikipedia.org/wiki/XML#DTD) của tài liệu. Sau đây là một ví dụ cơ bản về khai báo thực thể tại một DTD nhỏ nội bộ. Các thực thể được khai báo có thể mô tả các ký tự đơn hay các đoạn văn bản, và có thể tham chiếu lẫn nhau.

1. **<?xml** version="1.0" encoding="UTF-8"**?>**
2. <!DOCTYPE example [
3. <!ENTITY copy "&#xA9;">
4. **<**!ENTITY copyright-notice "Copyright &copy; 2006, XYZ Enterprises"**>**
5. ]>
6. **<root>**
7. &copyright-notice;
8. **</root>**

Khi xem tại một trình duyệt thích hợp, tài liệu XML trên sẽ hiện ra như sau:

**<root>** Copyright © 2006, XYZ Enterprises **</root>**

Các **tham chiếu ký tự số** trông giống như các thực thể. Nhưng thay cho một cái tên, chúng gồm một ký tự "[#](http://vi.wikipedia.org/w/index.php?title=Number_sign&action=edit&redlink=1)" và theo sau là một con số. Con số (theo [hệ thập phân](http://vi.wikipedia.org/wiki/H%E1%BB%87_th%E1%BA%ADp_ph%C3%A2n) hoặc [hệ cơ số 16](http://vi.wikipedia.org/w/index.php?title=H%E1%BB%87_c%C6%A1_s%E1%BB%91_16&action=edit&redlink=1) với tiền tố "x") đại diện cho một mã hiệu Unicode (*Unicode code point*), và thường được dùng để đại diện cho các ký tự không dễ gõ trên máy tính, chẳng hạn một chữ cái Ả-rập trong một tài liệu được soạn trên một máy tính châu Âu. Dấu & trong ví dụ "AT&T" có thể được biểu diễn như sau (số 38 thập phân và 26 trong hệ cơ số 16 đều đại diện cho giá trị Unicode của dấu &):

1. **<t**ên-công-ty**>**AT&#38;T**</t**ên-công-ty**>**
2. **<t**ên-công-ty**>**AT&#x26;T**</t**ên-công-ty**>**

Còn có nhiều quy tắc khác cần thiết cho việc viết các tài liệu XML định dạng đúng, chẳng hạn một tên XML có thể chứa các ký tự nào, nhưng phần giới thiệu ngắn này chỉ cung cấp các kiến thức căn bản để đọc và hiểu được nhiều tài liệu XML.