Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares

Maurício Barcellos Almeida
Universidade Federal de Minas Gerais. Mestrando em ciência da informação
E-mail: mba@eb.ufmg.br

Resumo

O HTML – Hypertext Markup Language – é uma linguagem de marcação, inicialmente concebida como uma solução para a publicação de documentos científicos em meios eletrônicos, que ganhou popularidade e se tornou padrão para a Internet. Diversos tipos de aplicações, como navegadores, editores, programas de e-mail, bancos de dados etc., tornam possível atualmente o uso intensivo do HTML. Ao longo dos anos, recursos têm sido adicionados ao HTML para que ele possa atender às expectativas de usuários e sistemas computadorizados, aumentando sua complexidade. Estima-se que a versão 4.0 do HTML possua aproximadamente cem diferentes marcações fixas (conhecidas como tags), sem contar aquelas específicas para cada tipo de navegador da Internet. É comum se encontrarem páginas HTML que possuem mais marcações do que conteúdo.*
Uma possível solução para novas demandas nessa área é a utilização do Extended Markup Language (XML), uma linguagem de marcação que pode introduzir novas possibilidades e trazer melhor integração entre dados e usuários. Este artigo se propõe a abordar, de forma introdutória, o XML, sua utilização na Internet, alguns conceitos complementares necessários ao entendimento do assunto em apresentar vantagens no uso do XML, em relação ao HTML. Além disso, pretende apresentar o assunto como um campo fértil para discussões, proposições e estudo por profissionais da ciência da informação.

Palavras-chave

XML; HTML; Linguagens de marcação; Internet; Intranet.

An introduction to XML, its use on the Internet and some complementary concepts

Abstract

Initially conceived as a solution to publish scientific documents in electronic midia, the HTML language soon gained popularity and became standart for the Internet. Several kinds of applications as browsers, editors, e-mail programs, databases, etc, have nowadays made possible the intensive use of HTML. Along the years, resources have been added to HTML so that it can support users' expectations and systems' prerrogatives, making it more and more complex. Estimates say that the HTML version 4.0 has about one hundred different tags, not considering those which are specific to each browser. It's common find HTML pages that have more tags than content.
A possible solution to new demands in this area is the use of XML-Extended Markup Language, that can introduce new possibilities and provide better integration between data and users. This article is an introductory approach to the XML language, to its use on the Internet and to some complementary concepts necessary to understand the matter. It also presents advantages of XML use as opposed to HTML, demonstrating the matter as a fertile field to discussions, propositions and studies to Science Information professionals.

Keywords

XML; HTML; markup languages; Internet; Intranet.

INTRODUÇÃO

Neste trabalho, apresenta-se uma visão geral das chamadas "linguagens de marcação", utilizadas para a transferência e representação de dados na Internet. Serão abordados aspectos gerais de linguagens de marcação XML, sua utilização na Internet e conceitos complementares necessários ao entendimento do assunto. É também apresentado o relacionamento do XML com outras linguagens, como o HTML, atual padrão para a Internet, e com o SGML – Standard Generalized Markup Language.

Em dezembro de 1997, o W3 Consortium publicou a versão 1.0 do XML, uma simplificação do SGML, que possibilita a páginas da Internet apresentar estrutura semântica. A aplicação do XML no tratamento de dados bibliográficos é estudada atualmente por diversas empresas e instituições em todo o mundo, com o objetivo de se obterem abordagens mais eficientes para a manipulação de dados na Internet. A ciência da informação, como campo dedicado às pesquisas científicas voltadas para os problemas da efetiva comunicação do conhecimento e seus registros entre os seres humanos (Saracevic, 1996), encontra, nesse assunto, boas oportunidades para estudo e discussão, no contexto das tecnologias informacionais.

O artigo está organizado da seguinte forma: a seção 1 apresenta histórico e definição de linguagens de marcação, além da motivação para estudo do XML; na seção 2 é apresentado breve histórico do XML e conceitos básicos relativos às linguagens de marcação; a seção 3 introduz o conceito de dados semi-estruturados, como se aplicam na Internet e as possíveis representações utilizadas; o item 4 estuda em maior profundidade o XML, abordando sua sintaxe, sua relação com os dados semi-estruturados, gramáticas utilizadas para geração de arquivos XML e exemplos; na seção 5 discutem-se algumas considerações finais e apresentam-se conclusões; finalmente, na seção 6, apresentam-se possibilidades de trabalhos futuros.

Seção 1
LINGUAGENS DE MARCAÇÃO

Alguns conceitos sobre marcas e linguagens de marcação se fazem necessários para melhor entendimento do assunto e são apresentados a seguir.

Historicamente, usa-se a palavra "marcação" para descrever anotações ou marcas em um texto, que tem por objetivo dar instruções ao desenhista ou datilógrafo sobre a maneira como uma parte do texto deveria ser representada. Como exemplos, pode-se citar um sublinhado ondulado que indicaria negrito, símbolos especiais para passagens a serem omitidas ou impressas com uma fonte especial, dentre outras. Como a formatação e a impressão de textos se tornaram automatizadas, o termo foi estendido para todos os tipos de códigos de marcação em textos eletrônicos. Todos os textos impressos são codificados com sinais de pontuação, uso de letras maiúsculas e minúsculas, regras para a disposição do texto na página, espaço entre as palavras etc. Estes elementos são um tipo de "marcação", cujo objetivo é ajudar o leitor na determinação de onde uma palavra termina e onde outra começa, ou identificar características estruturais (por exemplo, cabeçalhos) ou simples unidades sintáticas (por exemplo, parágrafos e sentenças). Codificar ou "marcar" um texto para processamento por computadores é também um processo de tornar explícito o que é conjetural. Indica como o conteúdo do texto deve ser interpretado.

Dessa forma, por "linguagem de marcação", entende-se um conjunto de convenções utilizadas para a codificação de textos. Uma linguagem de marcação deve especificar que marcas são permitidas, quais são exigidas, como se deve fazer distinção entre as marcas e o texto e qual o significado da marcação.

As linguagens de marcação encontram atualmente sua melhor aplicação nos arquivos HTML, mais conhecidos como "páginas da Internet", os quais são interpretados por *softwares* populares (navegadores ou *browsers*).

Por que o XML?

Diversas áreas do conhecimento discutem atualmente sobre a possibilidade de melhor aproveitar a massa de informações disponível na Internet, transformando-a em algo mais gerenciável e útil. Algumas propostas em estudo contemplam a adoção da linguagem de marcação XML em conjunto com procedimentos complementares (como, por exemplo, padrões de metadados em formato eletrônico) que permitam conferir elementos semânticos à Internet.

Em uma rede como a Internet, que conecta diferentes tipos de computadores e plataformas, a informação deve ser acessível, sem restrições impostas por modelos ou formatos de dados proprietários, os quais possibilitam às empresas que detêm seus direitos alterá-los arbitrariamente.

A linguagem de marcação SGML é um padrão internacional, não proprietário e de código aberto, utilizado já há bastante tempo para troca eletrônica de dados e que pode ser utilizada por diferentes sistemas informatizados. Um dos objetivos do SGML é garantir que documentos codificados de acordo com suas regras possam ser transportados de um ambiente de *hardware* e *software* para outro, sem perda de informação. Tanto o HTML (atual padrão em uso na Internet) quanto o XML derivam do SGML e, portanto, apresentam características similares.

O XML tem uma importante característica adicional: permite ao autor do documento a definição de suas próprias marcas. Esta característica confere à linguagem XML "habilidades" semânticas, que possibilitam melhorias significativas em processos de recuperação e disseminação da informação. As possibilidades e os benefícios reais em processos de recuperação da informação não são tratados neste trabalho.

O XML é uma arquitetura que não possui elementos* e marcas predefinidas. Não especifica como os autores vão utilizar metadados, sendo que existe total liberdade para utilizar qualquer método disponível, desde simples atributos**, até a implementação de padrões mais complexos. Como exemplos de padrões de metadados para meios eletrônicos pode-se citar o *ISO11179*, *Dublin Core*, *Warwick Framework*, *RDF – Resource Description Framework*, e *PICS – Platform for Internet Content Selection*.***

Marchal (2000) cita algumas áreas em que o XML pode ser útil no curto prazo, com vantagens significativas em relação ao HTML: na manutenção de grandes *sites*, no intercâmbio da informação entre organizações, no gerenciamento de conteúdo de *sites*, nas aplicações de

* Ver seção Sintaxe do XML - Elementos

** Ver seção Sintaxe do XML - Atributos

*** Sobre estes padrões consultar Internet via WWW ISO URL http://ww.iso.org / Dublin Core URL http://purl.oclc.org / Warwick Framework URL http://cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593 / Platform for Internet Content Selection URL http://www.w3.org/PICS

comércio eletrônico, nas aplicações científicas com o uso de novas linguagens de marcação para fórmulas matemáticas e químicas, em dispositivos computacionais alternativos (como os *palmtops*, *handhelds* etc.), entre outros.

Seção 2
CONCEITOS BÁSICOS

O XML não é uma linguagem de marcação predefinida (como o HTML) e possibilita ao autor do documento projetar sua própria marcação. A especificação do XML define um dialeto simples do SGML, permitindo o processamento dos documentos na Internet e utilizando-se de recursos inexistentes no HTML. Torna simples a transmissão e compartilhamento desses documentos via Internet.

Breve histórico

Até alguns anos, a publicação de dados eletrônicos estava limitada a poucas áreas científicas e técnicas, mas, atualmente, trata-se de uma atividade universal. O uso do HTML na Internet possibilitou que os dados fossem apresentados em uma estrutura simples e de fácil leitura. Entretanto, o HTML apresenta limitações fundamentadas em sua própria concepção, baseada em marcações fixas. A emergência do XML como um padrão para a representação de dados na Internet pode facilitar a publicação em meios eletrônicos, por prover uma sintaxe simples, legível para computadores e seres humanos.

O XML é uma linguagem derivada da SGML e foi idealizada por Jon Bosak, engenheiro da Sun Microsystems. O autor era conhecedor e usuário da SGML e apresentou ao W3 Consortium* sua idéia de explorar o SGML em aplicações voltadas para Internet. Em 1996, foi criado o XML, inicialmente como uma versão simplificada do SGML, e, em fevereiro de 1998, o XML tornou-se uma especificação formal, reconhecida pelo W3 Consortium.

Visão geral das linguagens de marcação – SGML, HTML e XML

Uma linguagem de marcação é diferente das linguagens de programação tradicionais, como o C, Basic, Java etc. Essas últimas especificam uma maneira para se calcular, executar ações e tomar decisões. Já o SGML e o XML permitem maneiras de descrever o dado para armazenamento, transmissão ou processamento por um programa.

XML, SGML e HTML não têm os mesmos propósitos , conforme é mostrado a seguir:

• **SGML** é o *Standard Generalized Markup Language* definido pela ISO 8879* e representa um padrão internacional para definição de estrutura e conteúdo de diferentes tipos de documentos eletrônicos. A SGML pode ser chamada de "língua mãe" e é usada para descrever tipos diferentes de documentos em muitas áreas da atividade humana, desde transcrições de antigos manuscritos irlandeses até documentação técnica para aviões de guerra; de registros de pacientes em unidades médicas até notação musical.

• **HTML** é o *Hypertext Markup Language* definido pela IETF-RFC1866** e consiste de uma aplicação específica do SGML utilizada na Internet. O HTML define um tipo de documento simples, com marcações fixas projetadas para uma classe de relatórios técnicos de uso comum em escritórios, como, por exemplo, cabeçalhos, parágrafos, listas, ilustrações e algumas possibilidades para hipertexto e multimídia. É o padrão atualmente em uso na Internet.

• **XML** ou *Extended Markup Language* é uma versão abreviada do SGML, que possibilita ao autor especificar a forma dos dados no documento, além de permitir definições semânticas. Um arquivo eletrônico XML pode conter, simultaneamente, dados e a descrição da estrutura do documento, através do DTD-*Data Type Definitions* (gramáticas que conferem estrutura ao documento XML). O XML obtém benefícios omitindo as partes mais complexas e menos utilizadas do SGML.

De acordo com o W3 Consortium, entre os objetivos estabelecidos na especificação da linguagem XML, estão as seguintes características: ser diretamente utilizável na Internet; ser legível por humanos; possibilitar um meio independente para publicação eletrônica; permitir a definição de protocolos para troca de dados pelas empresas (independentemente da plataforma de *hardware* e *software*); facilitar às pessoas o processamento de dados

* World Wide Internet Consortium: grupo baseado no MIT – Massachussets Institute of Technology – responsável pelo projeto de padrões para a Internet.

* Consultar ISO – International Standard Organization. Disponível na Internet via WWW URL http://ww.iso.org

** Consultar IETF – Internet Engineering Task Force. Disponível na Internet via WWW. URL http://ww.ietf.org

pelo uso de *softwares* de baixo custo; facilitar a utilização de metadados que auxiliam na busca de informações; aproximar "produtores" e "consumidores" de informação.

Seção 3
CLASSIFICAÇÃO DE DADOS A PARTIR DE SUA ESTRUTURA

Ao se trabalhar com dados em bases eletrônicas, pode-se distinguir formas em que é possível representá-las. As páginas da Internet são consideradas dados "semi-estruturados", de caráter intermediário, ou seja, que apresentam "alguma estrutura". Têm-se ainda os dados "estruturados", como, por exemplo, aqueles presentes nos bancos de dados relacionais, e os "não-estruturados", como, por exemplo, o texto livre. Os documentos da Internet apresentam variações, mas possuem alguma regularidade, ou seja, alguma estrutura. Têm sido estudados novos modelos de dados semi-estruturados e linguagens de consulta adaptadas a eles, visto que correspondem a uma representação mais flexível e mais adaptada ao ambiente da Internet.

De acordo com a figura 1, os dados estão dispostos na Internet e em outras fontes em três categorias, em relação à maneira como estão estruturados.
Os dados semi-estruturados representam hoje grande parte dos dados disponíveis em ambientes como Internet e intranets corporativas, nem sempre podem ser integrados à base de conhecimento organizacional pela ausência de processos adequados para sua manipulação.

Os dados semi-estruturados não estão em forma de um texto livre, que requer processamento pesado, mas também não estão sujeitos às restrições impostas por uma estrutura rígida como a dos bancos de dados relacionais. Os dados estruturados organizam suas instâncias em regras bem definidas, de forma a possibilitar, através da aplicação de filtros e consultas, agrupamento e extração de dados relevantes para os usuários. Os dados semi-estruturados possuem a habilidade de aceitar variações na estrutura, de forma que possam se adequar melhor a situações reais. Por exemplo, uma página da Internet de uma livraria virtual, em que se faça uma consulta solicitando-se todos os livros disponíveis cujo assunto é "ciência da informação", retorna uma lista dos nomes dos livros e mais alguns campos, como, por exemplo, preço, título, autor etc. Porém, alguns itens, mas não todos, podem apresentar o campo "desconto". De uma livraria para outra, a variação na estrutura pode ser ainda maior.

FIGURA 1
Dados conforme sua estrutura

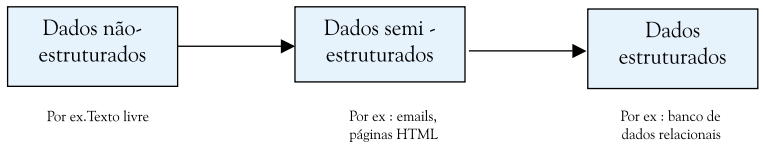


FIGURA 2
Exemplo de texto e seu correspondente XML

```
Catálogo de endereços                    <?xml version="1.0"?>
João Silva                               <catálogo de endereços>
Rua Carijós, 135                           <entrada>
Belo Horizonte, MG 30.000                    <nome> João Silva </nome>
Brasil                                       <endereço>
31 3335-5556 (preferido)                       <rua> Carijós, 135</rua>
31 3549-4446                                   <estado> MG </estado>
joaosilva@net.com.br                           < cep> 30.000 </cep>
José Almeida                                   <país> Brasil </país>
jalmeida@net.com.br                        <endereço>
                                           <telefone preferido="true">31 3335-4456</telefone>
                                             <telefone> 31 3594-4446 </telefone>
                                             <email> joaosilva@net.com.br </email>
                                         </entrada>
                                         <entrada>
                                            <nome><prim>José</prim>
                                                  <sobren>Almeida</sobren>
                                            <email> jalmeida@net.com.br </email>
                                         </entrada>
                                       </catalogo de endereço>
```

Os dados semi-estruturados representam hoje um componente importante de ambientes heterogêneos como a Internet, e o padrão XML, por permitir a criação de uma marcação flexível, aceita bem variações na estrutura característica desse tipo de dado. Alguns conceitos básicos sobre dados semi-estruturados são apresentados na seção seguinte.

Dados semi-estruturados e XML

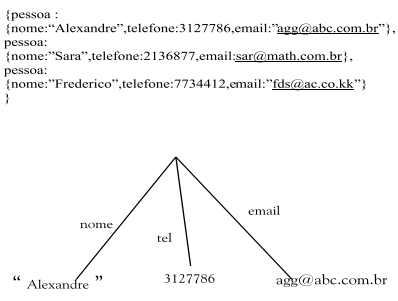
Ao se falar em "dados estruturados", é comum imaginar utilitários como planilhas, catálogos de endereço, transações financeiras, desenhos técnicos, entre outros (W3 Consortium, 2001). São dados dispostos em representações rígidas, sujeitas a regras e a restrições impostas pelo esquema que os criou. Programas que produzem tais dados os armazenam em disco, para que possam ser utilizados em formato binário ou texto.

Abiteboul (2000) refere-se a "dados semi-estruturados" como representações "sem esquema" ou "autodescritivas", termos que indicam que não há descrição em separado para tipo ou estrutura de dados (como acontece em um banco relacional). Em geral, quando se armazena ou se trabalha com dados, primeiro se descreve sua estrutura (tipo, esquema) e então se criam instâncias daquele tipo. Ao usar dados semi-estruturados, os dados são descritos diretamente, usando-se uma sintaxe simples. O exemplo na figura 2 apresenta um arquivo em texto livre e seu correspondente em XML.

Em uma abordagem mais próxima da ciência da computação (por sua familiaridade com linguagens de programação), os dados podem ser representados usando-se listas associadas* ou graficamente (por grafos), conforme exemplo a seguir:

{pessoa :
{nome:“Alexandre”,telefone:3127786,email:”agg@abc.com.br”},
pessoa:
{nome:”Sara”,telefone:2136877,email:sar@math.com.br},
pessoa:
{nome:”Frederico”,telefone:7734412,email:”fds@ac.co.kk”}

FIGURA 3
Representações de dados semi-estruturados simples em uma lista associada e em um grafo



Os dados semi-estruturados apresentam concepção mais próxima do que se encontra no mundo real, ou seja, os dados geralmente estão fora dos esquemas de bancos de dados, dispersos e não integrados a uma estrutura rígida. Conforme já citado, uma das maiores vantagens dos dados semi-estruturados é a sua habilidade de acomodar variações na estrutura.

Representação de dados semi-estruturados e XML por grafos

A complexidade de um modelo de dados depende da complexidade estrutural da informação que se quer representar. Um modelo que consiste apenas de um conjunto de pares (nome-atributo, valor-atributo) não possibilita a representação dos tipos de relacionamentos encontrados usualmente em sistemas de arquivos, principalmente em um ambiente heterogêneo como a Internet. Neste caso, um modelo de dados complexo (baseado em grafos e, seu caso particular, árvores) é mais apropriado. Além disso, algoritmos para busca e “navegação” em grafos são amplamente conhecidos e

FIGURA 4
Alguns exemplos de tipos de grafos

estudados, e sua adoção poderá facilitar o desenvolvimento de sistemas automatizados. Apresentam-se, a seguir, aspectos básicos da teoria dos grafos que possibilitam representação de dados complexos, a qual permitirá entender o modelo utilizado em um sistema informatizado com arquivos XML.

A terminologia utilizada em geral para descrever dados semi-estruturados é baseada na teoria de grafos. Segundo Szwarcfiter (1986), um grafo *G(V,E)* é um conjunto finito não vazio V e um conjunto E de pares não ordenados de elementos distintos de V. Os elementos V são os vértices, e E, as arestas do grafo G, respectivamente. Cada aresta “e“ pertencente à E será representada pelo par de vértices ou nós $e = (v,w)$ pela qual é formada. Os vértices v, w são os extremos ou nós da aresta “e“, conforme figura 4.

Um modelo possível para dados semi-estruturados que é utilizado para o padrão XML é conhecido como “grafo de arestas rotuladas” (*edge-labeled graph*), onde os nós representam objetos que estão conectados por arestas que representam valores (ver figura 3).

Dados semi-estruturados na Internet

Como exemplo da aplicação de um esquema de dados semi-estruturados baseado em XML, apresenta-se a seguir modelo XML parcial, de uma consulta realizada a uma página da DBLP-*Database Library Publications* (LEY, 2000). Este *site* apresenta dados de artigos, publicações e bibliografia de autores e foi convertido, no exemplo a seguir de sua representação na Internet em HTML, para o XML. É importante observar as variações de estrutura que ocorrem (por exemplo, nem todos os autores terão disponíveis os mesmos campos de dados) e como o XML é capaz de acomodar estas variações.

* Listas asssociadas são uma possibilidade para a representação de dados, utilizada por programadores de linguagem Lisp, que nada mais são que pares rótulo (por exemplo, na figura 1, “nome”) - valor (na figura 1, “Alexandre”).

Na figura 5, apresenta-se o conteúdo da página capturado da Internet, e, em seguida, seu correspondente em XML.

Uma possível representação para os dados do *site* ao lado, em XML, é vista na figura 6.

Existem diferentes abordagens para se tratar das variações na estrutura dos dados. Existem estudos para diversas linguagens de consulta específicas para dados semi-estruturados e XML (por exemplo, a *X-Query*, proposta pelo W3 Consortium). Outros trabalhos procuram conferir estrutura aos dados semi-estruturados para que estes possam ser explorados por meios mais conhecidos. Por exemplo, no sistema Stored, proposto por Deutsch, Fernandez & Suciu (1999), um arquivo XML é percorrido automaticamente por um algoritmo de *data-mining*, o qual reconhece estruturas similares que são transformadas em instâncias de um banco de dados relacional. Nesse sistema, os dados que não se adaptam às estruturas reconhecidas (ou seja, representam a estrutura variável) são armazenados em um grafo auxiliar, de onde podem ser recuperados.

Seção 4
SOBRE O XML

Conforme já citado, a linguagem XML compreende um padrão adotado pelo W3 Consortium, que possibilita a troca de dados na Internet, além de representar dados semi-estruturados. Uma grande quantidade de dados é atualmente publicada em páginas HTML. A figura 7, a seguir, mostra como dispor dados por meio de listas em página HTML.

O resultado apresentado no navegador seria algo como a figura 8, a seguir.

As marcações HTML, utilizadas acima, são *<h1>* para cabeçalho (o texto que será impresso em fonte maior), *<p>* para parágrafos (inicia uma nova linha), *<b>* para negrito e *<i>* para itálico. Enquanto a apresentação é de fácil leitura para humanos, não existe nada no texto em HTML que facilite a outros programas (*softwares*) entender a estrutura e o conteúdo dos dados. O HTML foi desenhado especificamente para descrever apresentação.

Já o XML foi projetado para descrever conteúdo e difere do HTML em três características: novas marcações podem ser definidas, estruturas de dados podem ser agrupadas em profundidade ilimitada, e um documento XML pode conter uma descrição opcional de sua gramática (DTD).

FIGURA 5
Dados capturados de página HTML

2000

- Gabriel M. Kuper, Leonid Libkin, Jan Paredaens (Eds.): Constraint Databases. Springer Verlag, 2000, 428pp., ISBN 3-540-66151-4
  Contents

1999

- Norman W. Paton (Ed.): Active Rules in Database Systems. Springer-Verlag, New York, 1999, ISBN 0-387-98529-8
  Contents
- Asuman Dogac, M. Tamer Özsu, Özgür Ulusoy (Eds.): Current Trends in Data Management Technology. Idea Group Publishing, January 1999, ISBN 1-878289-51-9
  Internet site of this book

1998

- Jan Chomicki, Gunter Saake (Eds.): Logics for Databases and Information Systems. Kluwer Academic Publishers, 1998, the book grow out of the Dagstuhl Seminar 9529: Role of Logics in Information Systems, 1995
  Contents
- Michael Stonebraker, Joseph M. Hellerstein (Eds.): Readings in Database Systems, Third Edition. Morgan Kaufmann 1998, ISBN 1-55860-523-1
  Internet site of this book

FIGURA 6
Possível representação XML para os dados da figura 5

```
< collection >
    < year-of-collection > 2000 < / year-of-collection >
        < publication >
            < author > Gabriel M. Kuper, Leonid Libkin, Jan Paredaens < / author >
            < title > Constraint Databases < / title >
            < publisher > Springer Verlag < / pusblisher >
            < date-of-publication > 2000 < / date-of-publication >
            < number-of-pages > 428 < number-of-pages >
            < ISBN > 3-540-66151-4 < / ISBN >
            < city-of-publication > - < / city-of-publication >
            < comments > - < / comments >
        < / publication >
    < year-of-collection > 1999 < / year-of-collection >
        < publication >
            < author > Norman W. Paton < / author >
            < title > Active Rules in Database Systems < / title >
            < publisher > Springer Verlag < / pusblisher >
            < date-of-publication > 1999 < / date-of-publication >
            < number-of-pages > - < number-of-pages >
            < ISBN > 0-387-98529-8 < / ISBN >
            < city-of-publication > New York < / city-of-publication >
            < comments > - < / comments >
        < / publication >
        < publication >
            < author > Asuman Dogac, M. Tamer Özsu, Özgür Ulusoy < / author >
            < title > Current Trends in Data Management Technology < / title >
            < publisher > Idea Group Publishing < / pusblisher >
            < date-of-publication > January 1999 < / date-of-publication >
            < number-of-pages > - < number-of-pages >
            < ISBN > 1-878289-51-9 < / ISBN >
            < city-of-publication > - < / city-of-publication >
            < comments > - < / comments >
        < / publication >
    < year-of-collection > 1998 < / year-of-collection >
        < publication >
            < author > Jan Chomicki, Gunter Saake < / author >
            < title > Logics for Databases and Information Systems < / title >
            < publisher > Kluwer Academic Publishers < / pusblisher >
            < date-of-publication > 1998 < / date-of-publication >
            < number-of-pages > - < number-of-pages >
            < ISBN > - < / ISBN >
            < city-of-publication > - < / city-of-publication >
            < comments > the book grow out of the Dagstuhl Seminar 9529: Role of
                                Logics in Information Systems, 1995 < / comments >
        < / publication >
        < publication >
            < author > Michael Stonebraker, Joseph M. Hellerstein < / author >
            < title > Readings in Database Systems, Third Edition < / title >
            < publisher > Morgan Kaufmann < / publisher >
            < date-of-publication > 1998 < / date-of-publication >
            < number-of-pages > - < number-of-pages >
            < ISBN > 1-55860-523-1 < / ISBN >
            < city-of-publication > - < / city-of-publication >
            < comments > - < / comments >
        < / publication >
                                < / collection >
```

O usuário de XML pode definir novas marcações para indicar estrutura. Por exemplo, a estrutura *<pessoa> .... </pessoa>* pode ser utilizada para descrever os dados referentes a uma pessoa. Ao contrário do HTML, o XML não fornece instruções de como o conteúdo deve ser apresentado. Existem várias maneiras de se apresentar o conteúdo de arquivos XML, como CSS-*Cascade Stylesheets*, XSL-*XML Stylesheet Language*, as quais não são abordadas neste trabalho.

Em sua forma básica, o XML é uma sintaxe simples para a transferência de dados, na concepção de dados semi-estruturados apresentada anteriormente. Como tal, é possível e mesmo provável que se torne um padrão para troca de dados na Internet. Para uma organização ou grupo de usuários, o XML permite uma especificação que facilita o intercâmbio de dados e a sua reutilização por múltiplas aplicações.

Sintaxe básica do XML

A sintaxe do XML é adequada para descrever dados semi-estruturados, apesar da possibilidade de existência de ambigüidades introduzidas pela presença de atributos. Conforme citado, a linguagem XML pode ser descrita por grafos com nós e rótulos, possibilitando que algoritmos conhecidos possam ser utilizados para a recuperação ou localização de dados em um grafo de grandes proporções.

Os elementos

O XML é uma representação textual de dados. O componente básico no XML é o "elemento", que é um "pedaço" de texto intercalado pelos sinais "<" e ">", como, por exemplo, "<pessoa>" e "</pessoa>". Dentro do elemento tem-se texto bruto, outro elemento ou uma mistura dos dois. Considere-se o exemplo da figura 9.

Uma expressão *<pessoa>* é chamada marcação inicial, e *</pessoa>* é chamada marcação final. Estas marcações, conforme já citado, são definidas pelo usuário. O texto entre a marcação inicial e a final, inclusive as marcações, é chamado "*elemento*"; a estrutura entre as marcações é descrita como o "*conteúdo*". O termo "*subelemento*" é também usado para descrever a relação entre um elemento e os outros elementos que o compõem. No exemplo acima *<email> ... </email>* é subelemento de *<pessoa> ... </pessoa>*. Usam-se elementos repetidos com a mesma marcação para representar coleções de dados, conforme ilustra a figura 10.

FIGURA 7
Fragmento de uma página HTML

```
<h1> Pessoas que estudam na UFMG </h1>
<p> <b> João </b>, 30 anos, <i>joao@ufmg.br </i> </p>
<p> <b> Claudia </b>, 25 anos, <i>claudia@ufmg.br </i> </p>
<p> <b> Jose </b>, 27 anos, <i>jose@ufmg.br </i> </p>
```

FIGURA 8
Página equivalente ao código HTML apresentado acima, na figura 7

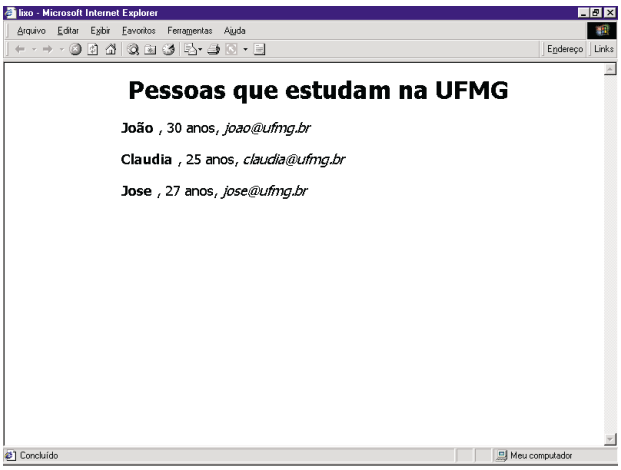


FIGURA 9
Fragmento de uma página XML

```
<pessoa>
        <nome> João </nome>
        <idade> 30 </idade>
        <email> joao@ufmg.br </email>
</pessoa>
```

FIGURA 10
Fragmento de uma página XML

```
<mestrado>
    <descrição> Pessoas que estudam na UFMG </descrição>
    <turma>
        <pessoa>
            <nome> João </nome>
            <idade> 30 </idade>
            <email> joao@ufmg.br </email>
        </pessoa>
        <pessoa>
            <nome> Claudia </nome>
            <idade> 25 </idade>
            <email> claudia@ufmg.br </email>
        </pessoa>
        <pessoa>
            <nome> Jose </nome>
            <idade> 25 </idade>
            <email> jose@ufmg.br </email>
        </pessoa>
    <turma>
<mestrado>
```

Os atributos

A linguagem XML permite associar "atributos" aos elementos. O termo "atributo" é utilizado no contexto do XML para especificar propriedades ou características do elemento. Atributos são definidos como pares (nome, valor). Na figura 11, atributos são utilizados para explicitar detalhes de um elemento.

Assim como no caso das marcações , o usuário pode definir também os atributos, como no exemplo anterior "língua", "moeda" e "formato". O valor do atributo é sempre um conjunto de caracteres que deve estar entre aspas.

O DTD – *Document Type Definitions*

Um DTD é uma gramática para o documento XML, e sua importância está relacionada à possibilidade que o próprio usuário defina suas marcações. Assim, é necessária uma gramática que apresente o significado da marcas criadas. Considere-se um documento XML consistindo de um número de elementos "pessoa" (figura 12).

Um DTD para este fragmento pode ser visualizado na figura 13.

A primeira linha diz que o elemento raiz (aquele que está situado no topo da árvore) é *<bd>*. As próximas cinco linhas são declarações de marcações, que mostram que *<bd>* pode conter um número arbitrário (representado pelo asterisco) de elementos *<pessoa>*, cada um contendo os elementos *<nome>*, *<idade>* e *<email>*, os quais contêm apenas caracteres "data" (não possuem mais elementos). A expressão "*pessoa**" é uma expressão regular, significando qualquer número de elementos pessoa. Outras expressões regulares são possíveis.

Uma característica importante do DTD é que ele pode se referir a dados externos usando uma URL-*Uniform Resource Locator*[*] . Tais referências externas podem ser úteis para processo de intercâmbio de dados. Encontram também aplicação prática na área de Tratamento da Informação, na definição de repositórios de autoridades e em metadados. Na figura 14, um exemplo de DTD, e na figura 15, do XML construído a partir desse DTD.

[*] Endereço utilizado nos navegadores da Internet para acesso a um conjunto específico de páginas.

FIGURA 11
Fragmento de uma página XML com atributos

```
<produto>
        <nome língua = "inglês" > book </nome>
        <preço moeda = "dólar" > 45,00 </preço>
        <fornecedor formato = "XLB56" língua = "inglês">
                <rua> Penbridge Square </rua>
                <número> 30 </número>
                <cep> 92310 </cep>
                <país> United Kingdom </país>
        <fornecedor>
<produto>
```

FIGURA 12
Fragmento de uma página XML

```
<bd>
        <pessoa>
                <nome> João </nome>
                <idade> 30 </idade>
                <email> joao@ufmg.br </email>
        </pessoa>
        <pessoa> ... </pessoa>
...
</bd>
```

FIGURA 13
Parte de um DTD para o fragmento XML da figura 12

```
<DOCTYPE bd [
        <!ELEMENT BD (pessoa*)>
        <!ELEMENT pessoa (nome, idade, email)>
        <!ELEMENT nome (#PCDATA)>
        <!ELEMENT idade (#PCDATA)>
        <!ELEMENT email (#PCDATA)>
]>
```

FIGURA 14
Fragmento DTD XML

```
<?xml version "1.0" ?>
<!DOCTYPE relatório {
<!ENTITY %abstract SYSTEM " www.eb.ufmg.br/mauricio/artigo1/abstract">
<!ENTITY %conteudo SYSTEM "" www.eb.ufmg.br/mauricio/artigo1 ">
]>
```

FIGURA 15
Fragmento de uma página XML, referente ao DTD da figura 14

```
<relatório>
        <meta keywords = "xml, www, Internet, semi-estruturado"
                autor = "Almeida"
                data = "25/05/2001"
        <titulo> Recuperação de informações em bases semi-estruturadas </título>
        %abstract;
        %conteudo;
</relatório>
```

Define-se *abstract* como sendo uma entidade que consiste de algum arquivo XML externo. O uso do *abstract* nos resultados do relatório consiste em inserir o documento *abstract* inteiro naquela posição. Além disso, o documento inclui explicitamente apenas elementos do tipo "meta" (alguma informação a respeito do relatório) e título; o *abstract* e o conteúdo estão em outro documento (externo).

Seção 5
CONSIDERAÇÕES FINAIS E CONCLUSÕES

Grande parte das referências ao assunto em estudos acadêmicos e de empresas é feita por pesquisadores de ciência da computação, que procuram estabelecer parâmetros, regras e padrões para utilização do XML no intercâmbio de dados entre sistemas automatizados. A Internet, com seu grande volume de dados, e os usuários, cada vez mais exigentes por buscas precisas e rápidas, têm levado pesquisadores a buscar ferramentas automáticas cada vez mais potentes. A construção destas ferramentas faz parte do escopo da ciência da computação, e, para a ciência da informação, o interesse está em estudar a informação e seus impactos, utilizando, conforme necessário, ferramental desenvolvido em outras áreas do conhecimento. Acredita-se, assim, que o assunto seja relevante para pesquisa por profissionais de ciência da informação.

Apresentou-se uma abordagem introdutória, que levanta características do XML, proporcionando uma visão geral, sem, entretanto, aprofundar-se. Novos trabalhos são necessários para que o tema seja estudado de forma adequada, visto que se trata de matéria extensa. O XML e estudos correlatos, como, por exemplo, o RDF-*Resource Description Framework*[*], estão ligados aos interesses de pesquisa da ciência da informação e se apresentam como um campo fértil para novas pesquisas. Um exemplo de utilização dessas tecnologias e padrões feito por profissionais da ciência da informação[**] é o trabalho "*TEI and XML In Digital Libraries*". A parte de maior interesse é intitulada "*XML and What It Will Mean for Libraries*" e apresenta uma abordagem básica das características do XML.

TRABALHOS FUTUROS

Como trabalhos futuros, propõem-se novos estudos que possam se aprofundar nas características do XML, aqui citadas de forma sucinta, possibilitando maior entendimento do universo de conceitos e acessórios relacionados. Além disso, são relevantes estudos que possam verificar os benefícios reais para processos de recuperação da informação, a partir das possibilidades semânticas do XML. Outros trabalhos pretendem relacionar o XML com padrões consagrados como o MARC e o Z239-50, além do estudo dos padrões de metadados associados ao XML (como o RDF). Estes estudos procurarão colaborar com iniciativas que visam à melhoria no atendimento às necessidades dos usuários de busca de dados e informações na Internet.

AGRADECIMENTOS

O autor gostaria de agradecer a contribuição para este trabalho da professora Beatriz Valadares Cendón, que fez diversas sugestões para melhorias no texto.

Artigo aceito para publicação em 07-02-2002

[*] RDF – padrão de metadados para a Internet, utilizado em conjunto com a liguagem XML

[**] Página de profissionais da Biblioteca do Congresso dos Estados Unidos e da Universidade de Michigan. Disponível na Internet via WWW. URL http://tigger.uic.edu/~cmsmcq/talks/teidlf1.html . Arquivo capturado em 03/06/01

REFERÊNCIAS

ABITEBOUL, S. Buneman, P. SUCIU, D. *Data on the web*. São Francisco : M. Kaufmamm, 2000.

BENDER, M. *Desenvolvendo sites com XML*. Florianópolis : Advanced Books, 2001.

DEUTSCH, A. Fernandez M. SUCIU D. Storing semistructured data with STORED. Disponível em: <www. http://citeseer.nj.nec.com/deutsch98storing.html>. Acesso em: 10 out. 2000.

LEY, M. Computer science bibliography. Disponível em: <http://www.informatik.uni-trier.de/~ley/db/books/collections/index.html>. Acesso em: 10 out. 2000.

MARCHAL, B. *XML by example*. Indianapolis : Que, 2000.

SARACEVIC, T. *Ciência da informação*: origem, evolução e relações. Perspectiva em *Ciência da Informação*, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SOUZA, T. B.; CATARINO, M. E.; SANTOS, P. C. Metadados: catalogando dados na Internet. *Transinformação*, v. 9, n. 2, p. 93-105, maio/ago. 1997.

SZWARCFITER, J. L. *Grafos e algoritmos computacionais*. São Paulo : Campus, 1986.

W3 CONSORTIUM. *Extensible Markup Language (XML)*: activity statement. Disponível em: <http://www.w3c.org.>. Acesso em: 25 set. 2001.

_____. *Extensible Markup Language (XML)*: W3C recommendation 6. 2. ed. versão 1.0. Oct. 2000. Disponível em: <http://www.w3c.org.>. Acesso em: 24 set. 2001.

_____. *XML in 10 points*. Disponível em: <http:www.w3c.org>. Acesso em: 20 ago. 2001.