

Laborator 1- XML

Limbajul XML, acronimul de la eXtensible Markup Language, a fost definit de consorțiul W3C în anul 1996.

Definiția completă a limbajului XML dată de W3C o găsiți la adresa: <http://www.w3.org/TR/REC-xml>.

Prezentarea în câteva cuvinte a XML-ului:

- XML, ca și HTML se bazează pe tag-uri.
- XML poate fi folosit de autorii de pagini Web ale căror nevoi depășesc limitele HTML-ului.
- XML a fost creat pentru descrierea structurilor de date.
- Tag-urile XML nu sunt predefinite; altfel spus, XML permite crearea tag-urilor personalizate, ceea ce este nemaipomenit.
- Pentru descrierea structurilor de date, documentele XML pot utiliza un DTD (Document Type Definition).
- Un document care conține un DTD este autodescriptiv

XML va înlocui HTML?

La ora actuală, răspunsul la această întrebare este NU. HTML rămâne primul limbaj utilizat pentru a indica browser-elor cum să afișeze informațiile pe Web.

XML a fost conceput pentru a stabili și a menține o interoperabilitate cu HTML.

Cele 10 obiective oficiale ale limbajului XML au fost publicate pe site-ul consorțiului W3C, la adresa <http://www.w3.org/TR/REC-xml>.

Aplicații XML:

Pe bună dreptate, vă puteți întreba ce faceți cu XML în practică? Teoria fără practică.

Prezentăm în continuare, o scurtă listă de aplicații XML:

- structurarea datelor în baze de date;
- structurarea documentelor;
- grafică vectorială (VML, sau Vector Markup Language);
- prezentări multimedia (SMIL – Synchronized Multimedia Integration Language; HTML + TIME, sau HTML Timed Interactive Multimedia Extensions);
- comunicarea deschisă între aplicații, via Web cu ajutorul mesajelor bazate pe XML (SOAP, sau Simple Object Access Protocol);
- schimbul de informații financiare (OFX, sau Open Financial Exchange);
- tranzacții comerciale pe Internet (XFDL, sau eXtensible Forms Description Language);
- resurse umane (HRMML, sau Human Resource Management Markup Language);
- formatarea formulelor matematice pe Web (MathML, sau Mathematical Markup Language);
- descrierea structurilor moleculare (CML, sau Chemical Markup Language);
- scrierea partiturilor muzicale (MusicML, sau Music Markup Language);
- buletine meteo (OMF, sau Weather Observation Markup Format);
- tranzacții imobiliare (RETS, sau Real Estate Transaction Standard).

XML - XML, un metalimbaj utilizat pentru descrierea datelor

Dacă sunteți documentat în XML, ați întâlnit cu siguranță această frază: „XML este un metalimbaj utilizat pentru descrierea datelor”.

Această afirmație nu ne oferă nici un detaliu privind metalimbajul în sine. În realitate, noțiunea de metalimbaj vă este bine cunoscută din primii ani de școală.

Un metalimbaj este un limbaj utilizat pentru descrierea unui alt limbaj.

Structura unei fraze din limbajul natural este un exemplu fundamental de metalimbaj.

Ex:

Iată cum descompunem în elemente de metalimbaj fraza:

Ultimul tren a plecat.

Fraza poate fi descompusă după cum urmează:

- adjectiv (ultimul);
- substantiv (tren);
- verb (a plecat);
- sfârșit de frază (.).

Această structură poate fi utilizată ca model în metalimbajul folosit pentru descrierea unei familii infinite de fraze de forma:

<adjectiv><substantiv><verb><sfârșit>

în care, elementele de metalimbaj sunt încadrate între „<” și „>”.

Pentru a delimita componentele unei fraze putem adopta sintaxa utilizată în (X)HTML, după cum urmează:

<adjectiv> Ultimul </adjectiv>

<substantiv> tren </substantiv>

<verb> a plecat </verb>

<sfârșit> . </sfârșit>

Putem scrie structura fundamentală, ca model pentru acest tip de fraze.

<adjectiv> </adjectiv>

<substantiv> </substantiv>

<verb> </verb>

<sfârșit> </sfârșit>

XML - Regulile de aur ale limbajului XML

Înainte de a începe crearea unui document XML este important să cunoașteți câteva din regulile XML pe care este bine să le respectați.

Regulile pe care trebuie să le respecte documentele XML facilitează atât crearea instrumentelor de analiză a documentelor cât și lucrul efectiv cu XML. Regulile sunt simple, multe

dintre ele fiindu-vă cunoscute de la HTML.

Regulile XML pot fi clasificate în două categorii:

- reguli de sintaxă XML;
- reguli de validitate a documentelor XML.

Regulile de sintaxă XML

Cea mai mare parte a regulilor de sintaxă XML sunt mult mai stricte decât cele aplicate unui document HTML.

Din acest motiv vom proceda la o prezentare comparativă a regulilor care se aplică unui document HTML, XHTML respectiv XML.

Ex: Iată cum definim paragraful: „Învăță să privești dincolo de aparențe!” în toate cele trei limbaje (figura 1.1).

Limbaaj	Codul sursă
HTML	<code><P>Învăță să privești dincolo de aparențe!</P></code>
XHTML	<code><p>Învăță să privești dincolo de aparențe!</p></code>
XML	<code><paragraf>Învăță să prvești dincolo de aparențe!</paragraf></code>

Figura 1

În exemplul precedent sintaxa XHTML este judecată corect atunci când exemplul HTML nu a fost considerat ca un document XHTML bine format, și nu trebuie să fie analizată de un parser. Aceasta nu înseamnă că un navigator nu va face analiza documentului, nu-l va interpreta și nu va identifica tag-ul care nu este bine format.

În cea mai mare parte a cazurilor, navigatoarele nu vor analiza documentul XHTML ca pe un document XHTML ci mai degrabă ca un document HTML; eventual, ele vor analiza documentul XHTML ca un document XML. Pentru moment puteți folosi un singur parser pentru a verifica erorile conținute de documentele dumneavoastră, înainte ca acestea să fie găzduite de un server.

Închideți toate elementele

Toate elementele XML trebuie să conțină, obligatoriu un tag de deschidere și unul de închidere. Acesta nu este cazul limbajului HTML, în care mai multe elemente sunt definite cu tag-urile de închidere facultative.

-În figura 1.1 (limbajul XHTML), elementul `<p>` conține tag-ul de început (`<p>`), conținutul (date de tip caracter) și tag-ul de sfârșit (`</p>`).

-În figura 1.1 (limbajul HTML), codul sursă conține tag-ul de început (`<P>`), conținutul propriu-zis (date de tip caracter) și opțional tag-ul de sfârșit `</P>`.

-În figura 1.1 (limbajul XML), elementul `<paragraf>` conține tag-ul de început `<paragraf>`, conținutul (date de tip caracter) și tag-ul de sfârșit `</paragraf>`.

-În limbajul HTML, singurul mod care-i permite navigatorului să deducă situația în care un paragraf s-a închis este prezența unui alt element, `<P>`, de exemplu. Deschiderea unui alt element sugerează că elementul precedent trebuie mai întâi să fie închis.

-Spre deosebire de HTML, XHTML impune ca toate elementele să fie închise.

Închideți elementele vide

Vă mai amintiți de cunoștințele noastre mai vechi: și
 pe care le-ați utilizat în HTML? În XML, spre deosebire de HTML aceste două elemente trebuie închise, dar fiind elemente vide (lipsite de conținut), ele urmează o sintaxă specială.

Iată cum inserăm o imagine, sigla.jpg în toate cele trei limbaje (figura 1.2).

Limbaaj	Codul sursă
HTML	
XHTML	
XML	<image source="sigla.jpg" />

Conform regulilor de sintaxă XML, elementele vide trebuie să se închidă, folosind una din metodele prezentate mai jos:

Metoda 1 - Închideți tag-ul cu un spațiu și o bară oblică (vezi figura 1.2);

Metoda 2- Adăugați un tag de închidere (Exemplu:
 </br>).

Utilizați ghilimelele pentru a delimita valorile atributelor

Limbajul HTML permite definirea fără ghilimele a valorilor atributelor. În XML, XHTML regula este alta: toate valorile atributelor trebuie să fie delimitate de ghilimele.

Iată cum delimităm valorile a trei atribute în limbajele: HTML, XHTML, XML (figura 1.3).

Limbaaj	Codul sursă
HTML	<P align=center>Învată să privești dincolo de aparențe!
XHTML	<p style="text-align:center">Învată să privești dincolo de aparențe!</p>
XML	<mesaj prioritate="urgent">Să nu-mi înșeli niciodată credința!</mesaj>

Toate atributele trebuie să conțină valori.

Elementele XML sunt case-sensitive.

În XML caracterele mici sunt interpretate diferit de cele mari (case-sensitive), ceea ce înseamnă că va trebui să utilizați elementele și atributele așa cum au fost ele scrise în DTD-ul (Document Type Definition) asociat.

Dacă decideți să nu utilizați un DTD, limbajul XML impune întotdeauna ca unui tag de deschidere să-i corespundă un tag de închidere. De exemplu, tag-ul <P> nu este același cu tag-ul <p>.

Imbricați corect elementele

În general, tag-urile sunt imbricate, ceea ce înseamnă că un element poate conține alte elemente. Tag-urile XML funcționează în același mod ca și parantezele în matematică.

Regulile de imbricare sunt următoarele:

- Fiecărui tag de deschidere trebuie să-i corespundă un tag de închidere. Numele celor două tag-uri trebuie să fie identice.
- Tag-ul de închidere al unui element fiu trebuie să preceadă tag-ul de închidere al părintelui său. Altfel spus, un fiu trebuie să fie închis înaintea părintelui său.
- Elementele imbricate sunt „fii ai elementului părinte”.

Exemple de elemente HTML, XHTML, XML corect imbricate :

- HTML <P> Acesta este venitul meu !
- XHTML <p> Acesta este venitul meu </p>!
- XML <întreprindere> TIMPURI NOI <cod> 2000 </cod> </întreprindere>

Exemple de elemente XML incorect imbricate :

- XML <întreprindere> TIMPURI NOI <cod> 2000</întreprindere></cod>
- XML <rezultate> <victorii> 6 </rezultate> </victorii>

Structura generala a unui document XML

O definitie simplificata a XML-ului este ca acesta ar fi un text structurat cu ajutorul unor etichete care sunt definite de catre utilizator. Un exemplu de fisier XML este dat mai jos:

```
<?xml version="1.0"?>
<Employees>
  <Employee EmployeeID="1">
    <FirstName>John</FirstName>
    <LastName>Smith</LastName>
    <Wage>40000</Wage>
  </Employee>
  <Employee EmployeeID="2">
    <FirstName>Mary</FirstName>
    <LastName>Pooh</LastName>
    <Wage>50000</Wage>
  </Employee>
</Employees>
```

Cu toate ca se remarca o mare asemanare intre un fisier XML si unul HTML, ele sunt clar deosebite ca scop: HTML este un limbaj care specifica datele si modul in care trebuie sa arate ele pe ecran, pe cand XML descrie datele si modul in care sunt ele structurate.

Datele dintr-un fisier XML sunt descrise de elemente si attribute create de catre utilizator (vocabularul nu este predefinit, ca in cazul HTML-ului).

Elementele au un tag (o eticheta) de inceput (ex: <Angajat>) si unul de sfarsit (</Angajat>).

Primul element (la noi: <Angajati>) este numit element radacina. Un element poate de asemenea sa aibe atribute (la noi: EmployeeID) carora le corespund valori ("2").

De asemenea, elementele pot contine subelemente. In exemplul nostru, elementul (radacina) <Angajati> contine doua subelemente <Angajat>; acesta este un aspect fundamental al documentelor XML: organizarea ierarhica. Un document XML poate fi "bine format" daca are un element radacina, orice tag de deschidere are si pe cel echivalent de inchidere, elementele sunt corect imbricate, iar valorile pentru atribute sunt continute intre ghilimele.

Reguli de validitate a documentelor XML

Regulile de sintaxă se învață și se aplică ușor deoarece sunt foarte simple dar și foarte stricte. Un document XML este bine format dacă respectă toate notațiile și regulile structurale XML.

Dacă un document XML bine format respectă regulile unui DTD asociat el se numește document valid. Atenție! Nu întotdeauna un document bine format are o structură validă!

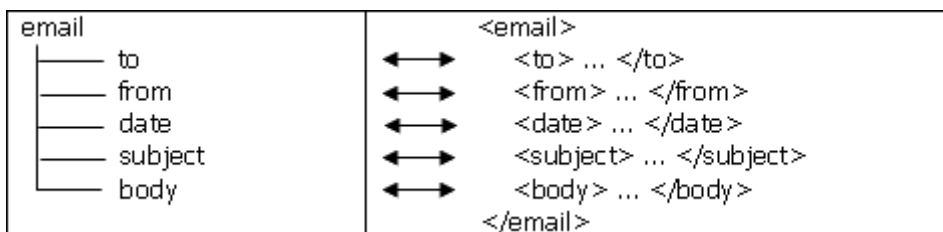
Cele mai multe restricții de validitate sunt facultative; ele merită să fie cunoscute chiar dacă nu sunt necesare!

Utilizați un element rădăcină!

Toate documentele XML trebuie să conțină cel puțin un element rădăcină. Această regulă este foarte simplă. Elementul rădăcină trebuie să conțină toate celelalte elemente ale paginii.

Un element rădăcină este delimitat printr-un tag de deschidere plasat la începutul documentului și un tag de închidere plasat la finele documentului.

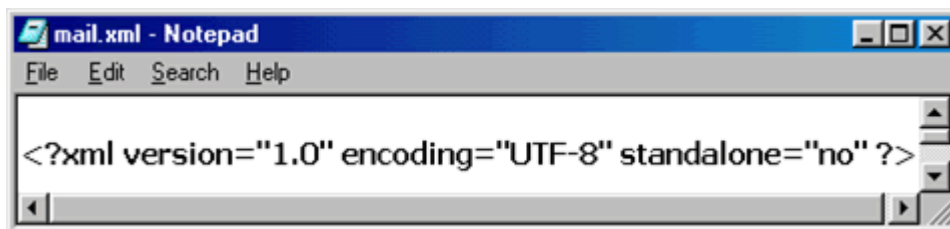
În figura de mai jos este prezentat tipul de date email (element rădăcină) sub formă de structură arborescentă.



În figura de mai jos este prezentat documentul XML, care descrie datele e-mail (vezi figura de sus) de o manieră structurată.

Declarația XML este facultativă

Deși declarația xml este facultativă, consorțiul W3C recomandă folosirea acesteia. Declarația xml semnalează procesorului că documentul pe care îl prelucrează este un document XML. Declarația XML poate accepta trei atribute: version, encoding și standalone



Observatie

- Declarația xml utilizează caractere mici cu excepția valorii atributului encoding (UTF – 8).
- Întrucât nu există decât o singură versiune XML și cum declarația XML este facultativă, ea poate fi abandonată. Totuși, este puțin probabil că specificația XML nu se va dezvolta. Cum interpretați declarația `<?xml version="2.0">?` Dacă utilizați declarația xml la începutul documentului, procesorul curent sau cel viitor vor ști, fără nici-o ambiguitate cum să prelucreze documentul dumneavoastră.
- O a doua versiune a limbajului XML a fost publicată dar ea nu corectează decât erorile minore ale specificației 1.0; ea nu reprezintă modificări de conținut pentru standard. Titlul oficial al acestei versiuni este XML 1.0 (A doua versiune).

Declarația DOCTYPE este facultativă

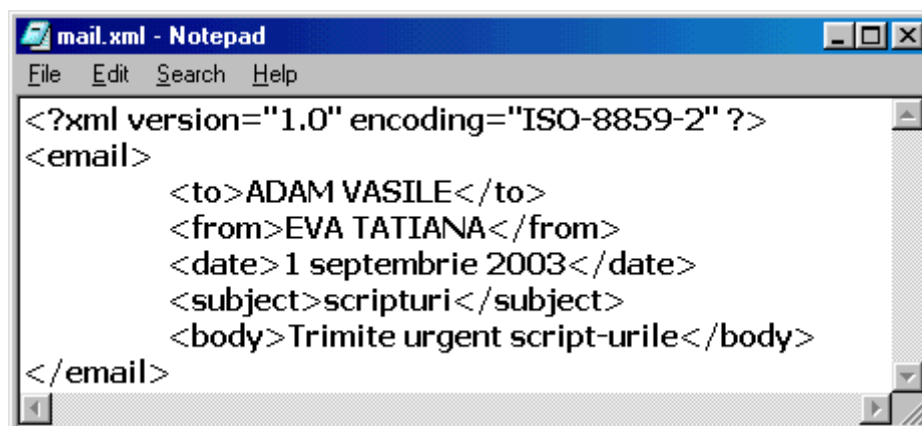
Un document XML poate fi scris în două maniere:

- documente bine formate;
- documente valide.

Documente XML bine formate

În figura 1 este prezentat un document bine format (o mai veche cunoștință a noastră!), care respectă regulile de gramatică și de sintaxă XML.

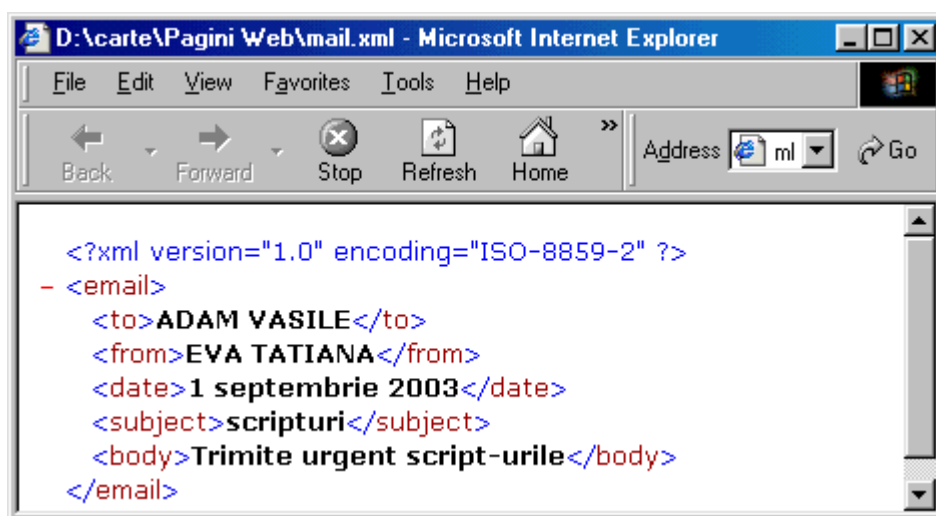
Figura 1



Dacă deschidem acest document în Internet Explorer (versiunea 5.0 sau superioară) se

obține rezultatul afișat în figura 2.

Figura 2



Reguli:

-Prima linie a acestui exemplu
<?xml version="1.0" encoding="ISO – 8859 – 2"?>
declară că documentul este scris în XML.

Valoarea atributului version nu poate fi în acest moment decât 1.0. Atributul encoding se referă la definiția setului de caractere utilizat în documentul XML.

– Această linie trebuie scrisă cu caractere mici. Nu este admisă nici-o majusculă! În caz contrar, programul de verificare (parser) XML al navigatorului va afișa o eroare – în acest caz documentul XML nu mai este bine format.

– Celelalte elemente XML au fost scrise cu tag-uri ... inventate de noi: (<email>...</email>; <to>...</to>; <from>...</from>; <date>...</date>; <subject>...</subject>; <body>...</body>.

– Toate elementele XML trebuie să conțină un tag de deschidere și un tag de închidere.

– Elementele vide XML trebuie să se termine cu un spațiu urmat de o bară oblică.

– Pentru a delimita valorile atributelor XML utilizați ghilimelele.

– Toate atributele XML trebuie să aibă valori.

– Toate documentele XML trebuie să aibă cel puțin un element rădăcină.

– Declarația XML este facultativă.

– Tag-urile XML trebuie să fie imbricate în mod corect.

– Numele tag-urilor pot conține: caractere alfanumerice (litere și cifre), liniuță (-), liniuță de

subliniere () și trebuie să conțină cel puțin o literă.

- Numele atributelor se scriu cu minuscule.
- Spre deosebire de HTML, sfârșitul unui document XML nu este marcat cu `</xml>`.
- Comentariile XML au aceeași sintaxă ca și comentariile XHTML.
- Un document bine format nu necesită un DTD (Document Type Definition). Declarația DOCTYPE este facultativă.

XML - Structura unui document XML bine format

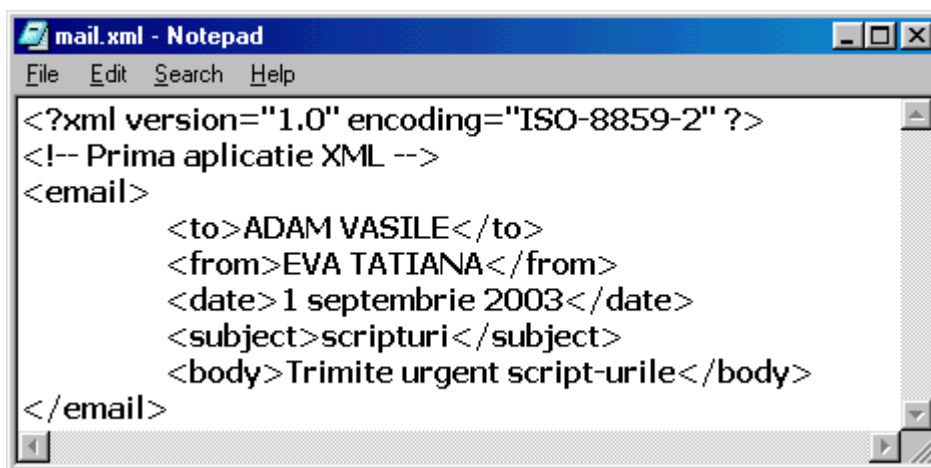
Un document bine format satisface un minimum de criterii de conformitate XML. Atunci când creați un document XML bine format, puteți începe direct prin a adăuga elementele de care aveți nevoie și prin a introduce datele documentului. Într-un document XML utilizatorii sunteți aceia care ... inventează propriile elemente în loc de a utiliza elementele predefinite. Vă asigurăm că nu veți avea nici o problemă de prelucrare sau de afișare a unui document XML bine format în Microsoft Internet Explorer.

După cum ați putut constata deja citind informațiile de mai sus, un document XML conține două părți principale:

- prologul;
- elementul rădăcină (numit și elementul document).

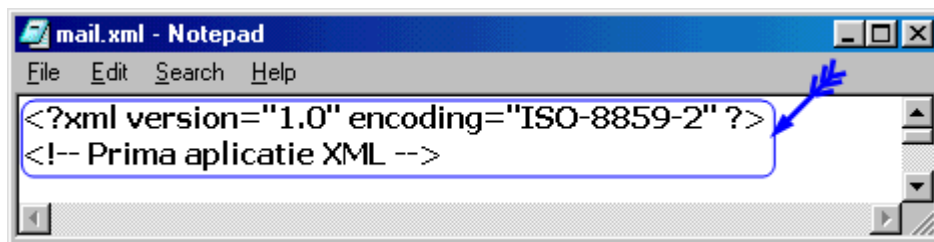
Pentru a înțelege mai bine anatomia unui document XML, vom face referiri la documentul mail.xml.

mail.xml



Să începem cu ... prologul!

Prologul cuprinde primele două linii



Prima linie este declarația XML (opțională). Dacă vă hotărâți să includeți totuși o declarație XML, ea trebuie să apară la începutul documentului. Cea de-a doua linie a prologului este un comentariu. Comentariile sunt recomandate în măsura în care acestea facilitează citirea și înțelegerea codului sursă.

Observatie:

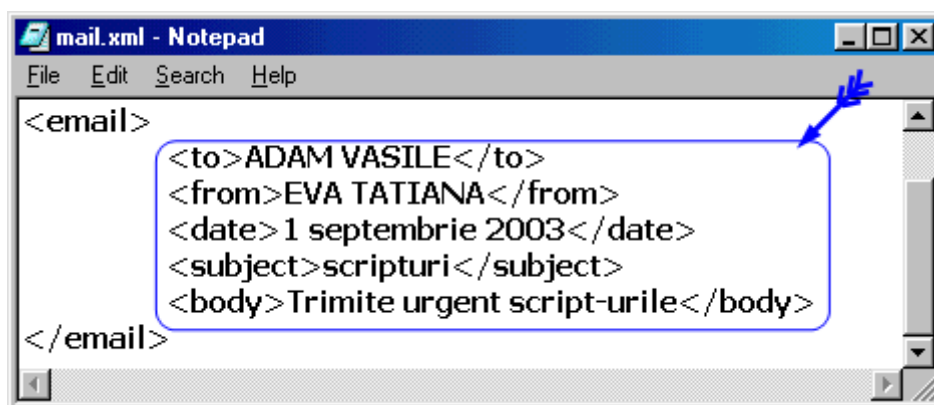
Prologul poate conține de asemenea și următoarele elemente opționale:

- un DTD (Document Type Definition), care definește tipul și structura documentului. Atunci când este utilizat, DTD-ul trebuie să fie plasat după declarația XML;
- o linie albă de spații;
- una sau mai multe instrucțiuni de prelucrare.

Elementul rădăcină. Cea de-a doua parte a unui document XML corespunde unui element unic desemnat ca element rădăcină sau element document, care poate conține alte elemente. Într-un document XML, elementele indică structura logică a documentului și conțin informațiile corespunzătoare. Un element tip cuprinde un tag de deschidere, conținutul elementului și tag-ul de închidere. Conținutul elementului poate fi constituit din date de tip caracter, alte elemente imbricate (nested), sau combinații ale celor două.

În exemplul nostru (documentul mail.xml), elementul rădăcină este email. Tag-ul său de deschidere este <email>, tag-ul de închidere este </email>, iar conținutul său este alcătuit din 5 elemente: <to>, <from>, <date>, <subject>, <body> (figura 3).

Figura 3



Observatie:

- Numele care apare în tag-ul de deschidere și în tag-ul de închidere ale unui element se mai numește și tipul de element.
- Fiecare din elementele inserate în elementul email (to, from, date, subject, body) nu conțin decât

date de tip caracter.

- Elementul rădăcină din cadrul unui document XML este similar elementului BODY din cadrul unei pagini (X)HTML, cu singura deosebire că puteți să-i atribuiți orice nume legal

Elemente vide

Puteți introduce de asemenea în documentul dumneavoastră un element vid, altfel spus un element fără conținut.

Puteți crea un element vid plasând tag-ul de închidere imediat după tag-ul de deschidere (figura 4).

Figura 4

```
<hr> </hr>
```

Sau, puteți utiliza tag-ul specific elementului vid (figura 5).

Figura 5

```
<hr />
```

Observatie:

- Cele două notații au aceeași semnificație.
- Puteți utiliza un element vid pentru a indica aplicației XML efectuarea unei acțiuni sau afișarea unui obiect.
- Un element vid poate stoca informație utilizând atribute XML.

Adăugați atribute elementelor XML

Puteți insera unul sau mai multe specificații de atribut în tag-ul de deschidere al unui element sau în tag-ul elementului vid. O specificație de atribut este o pereche nume-valoare care este asociată unui element.

În figura 6 este prezentat un exemplu de atribut (vânzare) căruia i s-a atribuit valoarea RODIPET.

Figura 6

```
<pret vanzare="RODIPET"> 30 euro </pret>
```

În figura 7 atributul sursa indică numele fișierului care conține imaginea pe care dorim s-o afișăm.

Figura 7

```
<sigla sursa="Droopy.gif" />
```

Observatie:

- Atributele reprezintă o alternativă la inserarea informațiilor într-un document.
- Utilizarea atributelor permite obținerea unui mai mare grad de flexibilitate în procesul prelucrărilor.
- În figura 8 sunt prezentate câteva exemple de specificații de attribute ilegale. Fără comentarii!

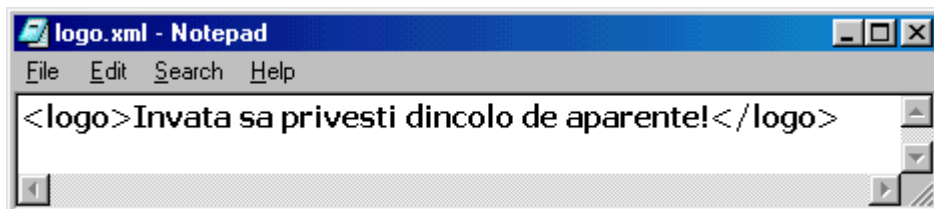
Figura 8

```
<curs tip=" "seral" ">  
<album tip="<CD>">  
<magazin tip="C&A">
```

Un document XML cu structură minimală

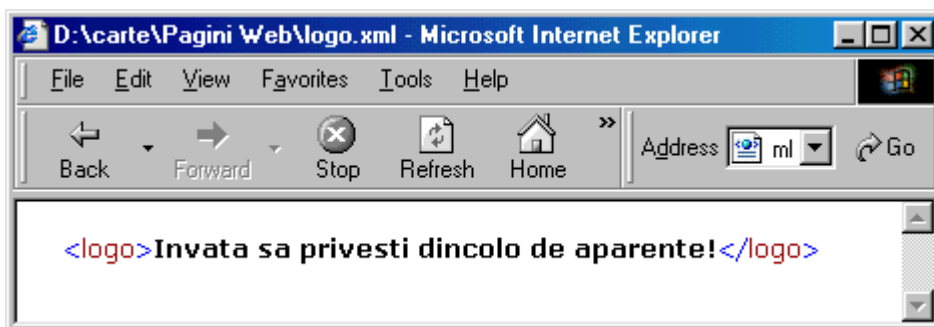
După cum am precizat, elementele din prologul unui document XML bine format sunt toate opționale. În consecință, prologul este el însuși opțional, iar documentul din figura 9 (document cu structură minimală) conține un singur element <logo>, care respectă norma XML pentru ca acesta să fie un document bine format.

Figura 9



În figura 10 se prezintă rezultatul vizualizării documentului în Internet Explorer.

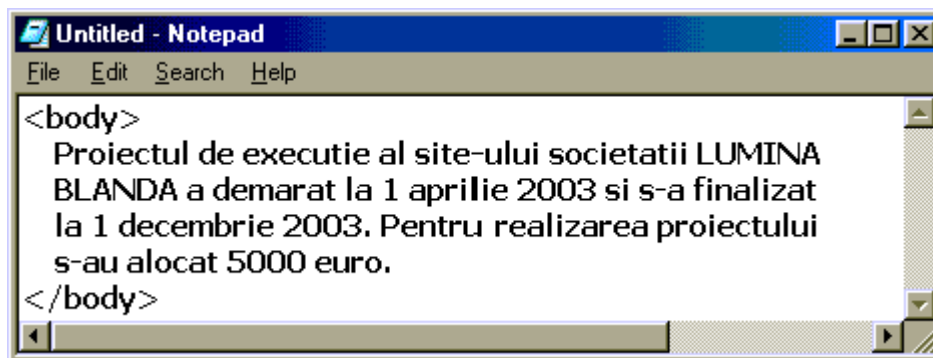
Figura 10



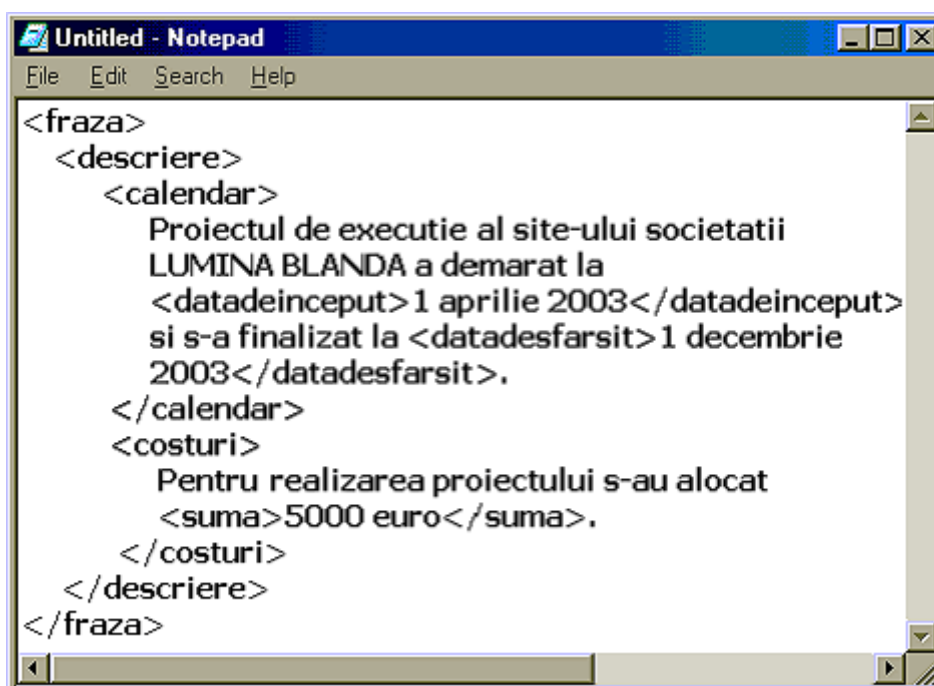
Aplicatie:

Creați un document XML cu structură minimală, pornind de la secvența de cod (X)HTML din figura 11.

Figura 11



În figura 12 se prezintă codul sursă XML, în care tag-urile ... imaginate de noi sunt semnificative, nu creează confuzii și descriu conținutul lor



Observatii:

- Acest document simplu XML poate fi citit și înțeles fără a utiliza un program specializat!
- Tag-urile <calendar>, <datadeinceput>, <datadesfârșit>, <costuri>, <suma> imaginate de noi furnizează informații privind datele (de tip caracter).
- Numele tag-urilor trebuie să respecte următoarele reguli:

- numele unui tag trebuie să înceapă cu o literă (majusculă sau minusculă) sau cu o liniuță de subliniere;
 - numele unui tag poate să conțină: litere, cifre, liniuțe de subliniere, linii și puncte;
 - numele unui tag nu trebuie să conțină spații;
 - numele unui tag nu trebuie să înceapă cu XML, xml, Xml, xML, xMl, xmL, Xml;
 - numele tag-urilor sunt limitate numai la caractere latine;
 - numele unui tag poate conține caracterul (:) numai în construcția specială, spațiu de nume.
- În XML, un document bine format trebuie să respecte regula: „fiecărui tag de început îi corespunde un tag de sfârșit”.
 - XML permite utilizarea elementelor fără conținut (vide).
 - Pentru a defini relațiile între elemente, XML folosește următoarea terminologie:
 - elementul rădăcină;
 - elementul părinte;
 - elementul fiu;
 - elementul frate.
 - În exemplul nostru (figura 12), <frază> este elementul rădăcină; <descriere> este elementul părinte al elementelor <calendar> și <costuri>; <calendar> și <costuri> sunt elemente fiu ale elementului <descriere>; <calendar> și <costuri> sunt elemente frați.
 - În structura unui document XML numai elementul rădăcină nu are un element părinte. Un element fiu aparține unui singur element părinte.

Regulile de aur ale documentelor XML bine formate

Documentul mail.xml pe care l-am analizat anterior este un exemplu de document XML bine format care respectă regulile de aur ale documentelor XML bine formate. Prezentăm în continuare câteva din regulile care stau la baza creării unui document XML bine format.

- Documentul trebuie să conțină un element de cel mai înalt nivel (elementul rădăcină sau elementul document). Toate celelalte elemente trebuie să se găsească în interiorul elementului rădăcină.
- Elementele trebuie să fie corect imbricate.
- Fiecare element trebuie să conțină un tag de deschidere și un tag de închidere, care trebuie să poarte același nume.
- Numele tag-urilor nu trebuie să conțină spații, ci numai litere, cifre (nu în prima poziție) și liniuțe de subliniere (_).
- Elementele XML sunt case-sensitive.
- Numele tag-urilor elementelor vide trebuie să se termine cu o bară oblică (/).
- Valorile atributelor trebuie să fie plasate între ghilimele sau apostrofuri.

Dacă utilizatorii respecta aceste reguli, documentele XML pe care le veți realiza vor fi cu siguranță bine formate și vor fi corect interpretate de browser-e cât și de alte programe.