Bijlage 27
Vergelijking van sgml en xml

Simon Pepping
Elsevier Science
E-mail: s.pepping@elsevier.nl

XML (eXtensible Markup Language) is één van de buzz words van het afgelopen jaar op het Internet. Het wordt begroet met kreten als 'Een tweede kans voor SGML', 'Web auteurs bevrijd uit het keurslijf van HTML', 'de nieuwe Web standaard voor informatieuitwisseling'. Zowel in SGML kringen als bij de softwareindustrie roept het enthousiaste reakties en grootse verwachtingen op.

XML komt voort uit twee bewegingen. 1. In de SGML wereld beseft men dat de SGML wel mooi maar moeilijk implementeerbaar is. Daardoor zijn er maar weinig toepassingen van ontstaan. 2. In de software industrie beseft men dat HTML te beperkt is om allerlei soorten informatie over het Internet uit te wisselen. XML is de ontmoeting van die twee: Het is een vorm van SGML ontdaan van zijn moeilijkste kanten, zodat het eenvoudiger wordt om er toepassingen voor te maken. Doordat het nog steeds SGML is, is het flexibel en krachtig genoeg om velerlei typen informatie te kunnen beschrijven.

Nu al zijn er verscheidene parsers voor XML beschikbaar, en hun grootte bedraagt maar rond een vijfde tot een tiende van de James Clark's `nsgmls` parser voor volledig SGML (op mijn Linux machine 1.2MB).

De meeste informatie over XML komt uit de SGML hoek en is vaak technisch van aard. Bovendien benadert het XML vanuit het SGML gezichtspunt en benadrukt vooral de verschillen met SGML. Daar heb je niet zoveel aan als je geen SGML kent. Een informatief dokument over XML voor belangstellenden die minder van SGML af weten is Peter Flynn's 'Frequently Asked Questions about the Extensible Markup Language' (`http://www.ucc.ie/cml/faq.sgml` en `http://www.ucc.ie/xml`). Het beantwoordt zelfs de vraag: 'What is SGML?'

Op de volgende bladzijden drukken we een notitie af van één van de voormannen van SGML, James Clark. Hierin geeft hij een precieze opsomming van de verschillen tussen SGML en XML. In de volgende alinea's zal ik enkele punten naar voren halen.

Het valt me op dat een in het oog springend verschil tussen SGML en XML niet genoemd wordt: XML documenten hoeven geen DTD te hebben. Zulke documenten kunnen weliswaar niet door een parser op hun geldigheid gekontroleerd worden (ze zijn niet 'valid'), maar als ze aan redelijke eisen van markup voldoen, zoals volledig getagged, geen slechte nesting, zijn ze wel 'well-formed'.

In XML zijn veel moeilijk implementeerbare kenmerken van SGML geschrapt in het belang van een grotere praktische haalbaarheid. Zulke vereenvoudigingen zijn o.a.:

- XML heeft een vaste SGML deklaratie; dit sluit al heel veel bijzondere mogelijkheden en moeilijkheden uit.
- Alle openings- en sluittags moeten aanwezig zijn (OMITTAG NO). Ongesloten openings- en sluittags zijn niet toegestaan. Attribuutwaarden moeten altijd tussen aanhalingstekens gegeven worden ('entered as literals'). Dit maakt de markup regelmatiger en dus gemakkelijker te parsen.
- < en & mogen niet als tekstkarakter voorkomen. In SGML mag dit onder bepaalde voorwaarden wel. Weer zo'n beperking die parsen gemakkelijker maakt.
- Een aantal verfijningen in SGML zijn overboord gezet ten gunste van praktische haalbaarheid, b.v. `NUTOKEN(S)`, `NUMBER(S)` en `NAME(S)` zijn niet toegestaan als gedeklareerde waarden voor attributen.
- Zelfs vaker gebruikte en dus zeker nuttig gebleken mogelijkheden van SGML zijn verwijderd ten gunste van de haalbaarheid, b.v. `INCLUDE/IGNORE` marked sections zijn niet in een document toegestaan.

XML heeft ook enkele zaken aan SGML toegevoegd, zoals:

- Er mogen meerdere attribuut deklaraties voor één element voorkomen.
- Er is geen onderscheid meer tussen als `EMPTY` gedeklareerde elementen en elementen zonder inhoud. Beide mogen gekodeerd worden als `<fig file="fig1.eps"/>`.

Deze zaken hebben een aanpassing van de SGML specifikatie vereist, de 'Web SGML Adaptations Annex' die in het document genoemd wordt.

Toepassingen van XML staan nog in de kinderschoenen, maar er is veel aktiviteit aan het front. Er zit een XML parser in MS Internet Explorer 4 en in Netscape Navigator 5 (James Clark's `expat` parser). Er zijn verscheidene standalone XML parsers, vooral in Java geschreven. Er zijn XML modules in Perl. En er komt dagelijks wat bij. Voor informatie zie de SGML/XML web page op `http://www.sil.org` en de nieuwsgroep `comp.text.sgml`.