

WWW – Grundlagen und Technologie

Grundlegende Komponenten und ihr Zusammenhang



Erik Wilde
TIK – ETH Zürich
Sommersemester 2001

Übersicht

- verbreitete Sichtweisen des WWW
- Hypermedia-Grundkomponenten
 - Hypertext Markup Language (HTML)
 - Uniform Resource Locators (URL)
 - Hypertext Transfer Protocol (HTTP)
- zusätzliche wichtige Architekturbestandteile
 - Identifikation von Medientypen (MIME)
 - Präsentationssteuerung (Style Sheets, CSS/XSL)
 - Extensible Markup Language (XML)
 - Dynamic HTML (DHTML)
- Zusammenfassung

Sichtweisen des WWW

- **Hypermediasystem**
 - Schwerpunkt sind Hypermedia-Dokumente
 - HTML als Sprache für Inhalte
 - wenig Interesse an der Präsentation
 - wenig Interesse an detaillierter Kontrolle der Schnittstelle
 - weniger Programm, mehr Spezifikation
- **Programmierplattform**
 - Schwerpunkt sind APIs
 - Java(-Script) als System zur verteilten Programmierung
 - grosse Kontrolle über funktionale Aspekte der Präsentation
 - wenig Interesse an gutem Hypermedia
 - wenig Interesse an maschineller Auswertbarkeit

WWW als Hypermediasystem

- **Entstehung als Informationssystem**
 - Dokumentation von Forschungsergebnissen
 - wichtig sind die Inhalte, nicht die Präsentation
- **im ersten Ansatz sehr einfache Mittel**
 - HTML für sehr einfache Strukturierung
- **problematische Entscheidungen des 1. Entwurfs**
 - die HTML DTD ist alles andere als optimal
 - kein Austausch von Metadaten
 - Server weiss nichts über Client und vice versa
 - keinerlei Kontrolle der Präsentation
- **seltsames Zwitterwesen Forschung/Pragmatik**

Evolution des WWW

- erste kommerzielle Nutzungen
 - (X)Mosaic als erster graphischer Browser
 - Integration von Funktionen im Browser (GIF/JPEG)
- kommerzielle Nutzer identifizieren Probleme
 - kein Transaktionskonzept (Sessions/Tracking)
 - keine Präsentationssteuerung (Corporate Identity)
- Netscape beginnt die Entwicklung zu steuern
 - HTML "Standardisierung" als Verfolgungsjagd
- Microsoft entdeckt das Internet 1996
 - Konkurrenz führt zu vielen Divergenzen
- (schnelle!) Standardisierung wird wichtig

Hauptkomponenten des WWW

- 1989: erster Entwurf
- 1990: erste Implementierung
- Hypertext Markup Language (HTML) 4.01
 - *"the document format for hypertext"*
- Uniform Resource Locator (URL)
 - *"how to name a document"*
 - nur minimal verändert seit dem ersten Entwurf
- Hypertext Transfer Protocol (HTTP) 1.1
 - *"how to get a document"*
- XML erste wirklich neue Komponente seit 1990
 - Entwurf ab 06/96, Standard 02/98

Inhalt und Präsentation (I)

<ul style="list-style-type: none"> • Vorlesungs-Titel • Vorlesungs-Beschreibung • Kapitel, Titel • Absatz • Unterabschnitt, Titel • Absatz • Unterabschnitt, Titel • Absatz 	<p>World Wide Web</p> <p>Das Web (WWW) ist ein weltweites, hypermediales Dokument.</p> <p>In der Beschreibung werden das Semantische Web und allgemeine Grundlagen zu Internet und Hypermedia besprochen.</p> <p>Das Programm der Vorlesung teilt sich in drei Termine auf, die einen jeweils abgeschlossenen Aspekt des WWW besprechen.</p> <p>Das Internet bildet die technische Infrastruktur des WWW auf. Es bietet unter anderem Dienste zum Datentransport und zur</p>	<ul style="list-style-type: none"> • Fett, zentriert, 24pt • Abstand, 16pt • Abstand, Nummer, 20pt • Abstand, 16pt • Abstand, Nummer, 18pt • Abstand, 16pt • Abstand, Nummer, 18pt • Abstand, 16pt
---	---	--

WWW (SS2001) - Grundkomponenten

7

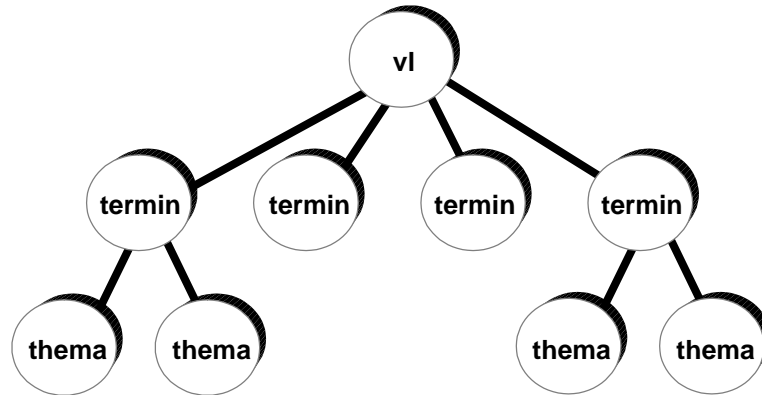
Inhalt und Präsentation (II)

- Inhalt eines Dokumentes
 - Menge von logischen Einheiten
 - unabhängig von der Präsentation
 - Konzepte wie Absatz, Überschrift, Kapitel, Teil, ...
- Präsentation eines Dokumentes
 - Menge von Layout-Einheiten
 - Resultat einer Formatierung des Dokumentes
 - viele mögliche Präsentationen eines Dokumentes
 - Konzepte wie Block, Seite, Abstand, Schrifttyp, ...
- abhängig von Umgebung und Anforderungen
 - *Festlegung* der Präsentation durch den Autor
 - *Freiheiten* für die Präsentation, Autor liefert Inhalt

WWW (SS2001) - Grundkomponenten

8

Document Tree



WWW (SS2001) - Grundkomponenten

9

SGML

- *Standard Generalized Markup Language (SGML)*
- definiert in ISO Standard 8879 (1986)
- Markup Language trennt Struktur und Text
 - Struktur ist syntaktisch identifizierbar
 - Text ist an definierten Stellen der Struktur erlaubt
- Vorteile einer Markup Language
 - einfache Repräsentation (Erzeugung "von Hand")
 - Austauschbarkeit als rein textbasierte Dokumente
- Definition beliebiger Dokumenttypen
 - anwendungsspezifische Dokumenttypen
 - gemeinsamer zugrundeliegender Mechanismus
 - gemeinsam verwendbare Software

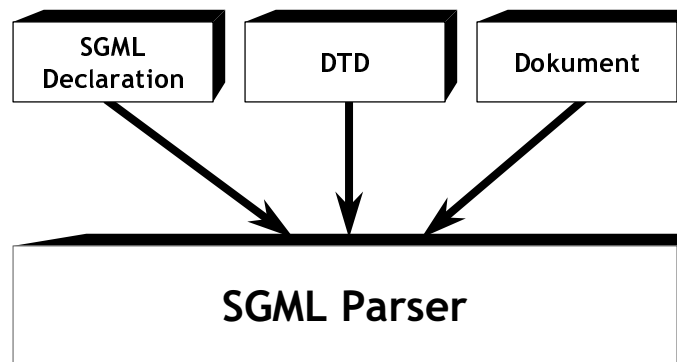
WWW (SS2001) - Grundkomponenten

10

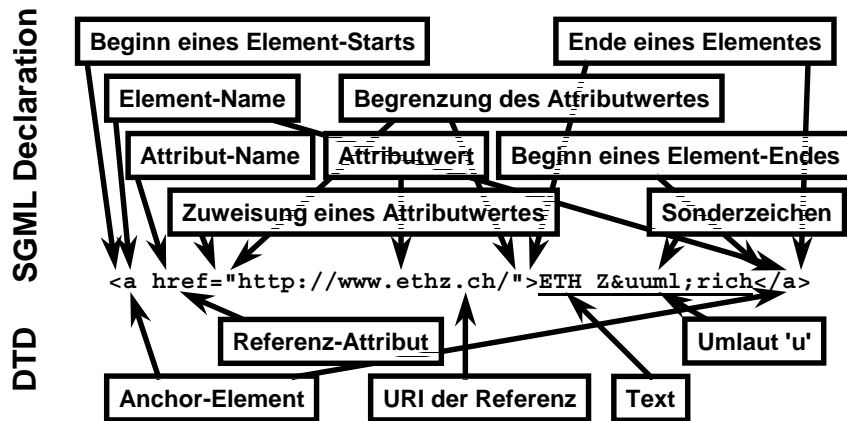
Hypertext Markup Language

- Anwendung (*Application*) von SGML
- HTML SGML *Declaration*
 - bestimmt die konkrete Syntax von HTML
 - bestimmt die SGML Features von HTML
- HTML SGML *Document Type Definition*
 - drei verschiedene Varianten der DTD
 - Definition der Elemente und Attribute
- zusätzliche Definitionen (kein SGML!)
 - Einschränkungen von Attributwerten
 - Bedeutungen von Elementen und Attributen

SGML Parser



SGML-Teile des HTML Standards



WWW (SS2001) - Grundkomponenten

13

HTML und Verwandte

- *Hypertext Markup Language (HTML)*
 - festgelegter Dokumententyp
 - basierend auf allgemeiner Sprache
- *Standard Generalized Markup Language (SGML)*
 - Mechanismus zur Definition von HTML
 - erlaubt Definition beliebiger Dokumententypen
 - HTML ist eine Anwendung von SGML
- *Extensible Markup Language (XML)*
 - Web-spezifisches Profile von SGML
 - leicht vereinfachte Version (weniger kompliziert)
 - keine Einschränkung der Allgemeinheit
 - optimiert auf Anwendbarkeit auf dem Web

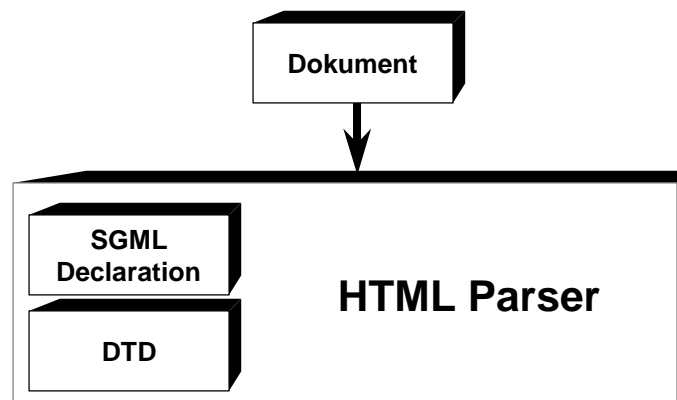
WWW (SS2001) - Grundkomponenten

14

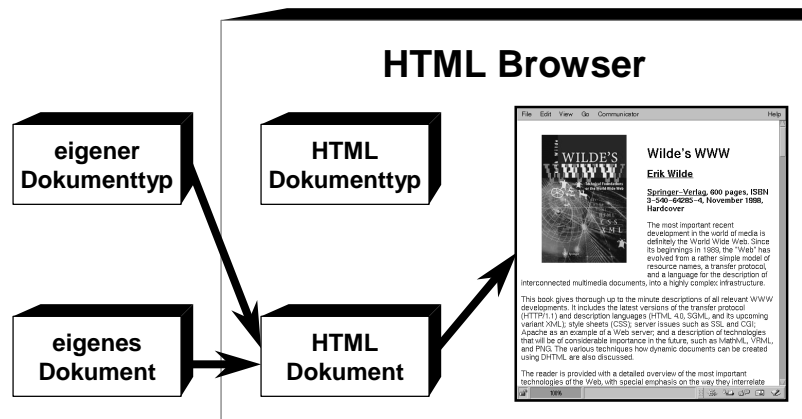
HTML als SGML Anwendung

- jeder Browser implementiert HTML Parser
 - mehr als 90% aller HTML Seiten sind kein SGML
 - Browser implementieren fehlertolerante Parser
 - kein Browser ist eine formal korrekte SGML Implementierung
 - HTML ist mehr als formal korrektes SGML
- Verbreitung generierter HTML Seiten
 - HTML Editoren (oftmals "eigenwilliges" HTML)
 - nicht unbedingt sinnvolle Wahl der Elemente
 - Verwendung eigener HTML-Erweiterungen
 - Generierung durch Skripte oder Programme
 - on-line bei der Abfrage
 - off-line bei der Generierung von Web-Sites

HTML Parser



Publishing mit HTML



WWW (SS2001) - Grundkomponenten

17

Identifizier vs. Links

- *Identifizier*
 - Identifikation unabhängig von Verbindungen
 - Identifizier existieren auch ohne Referenzen

`ETH`

- *Links*
 - enthalten (u.U. mehrere) Identifizier
 - zusätzliche Semantik (z.B. Link-Typ)
 - HTML definiert sehr einfache Links (→ *XLink*)

WWW (SS2001) - Grundkomponenten

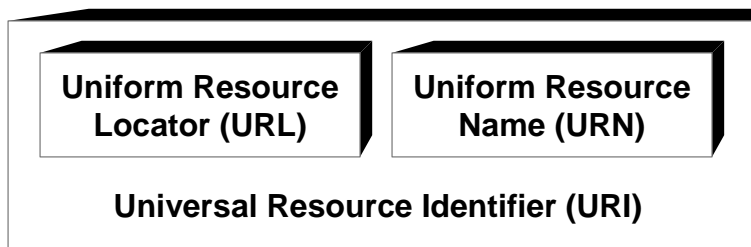
18

Universal Resource Identifier (URI)

uri = scheme ":" scheme-specific-part

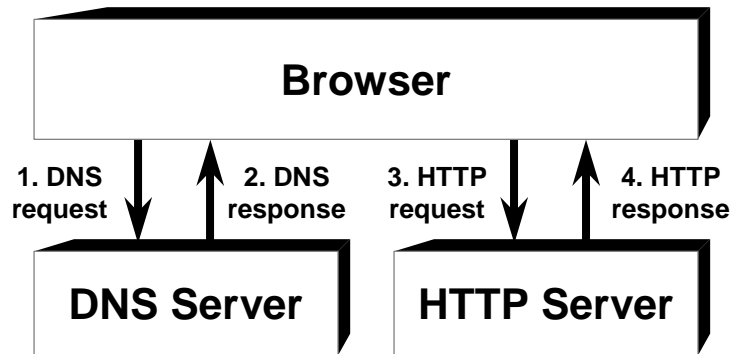
- definiert in Internet RFC 2396
- URI *schemes*
 - definieren eine Art der Identifizierung
 - definieren Identifizierung innerhalb des schemes
- URI *scheme specific parts*
 - Bedeutung hängt vom scheme ab
 - einige allgemeine Festlegungen zur Syntax
- URI sind *Locators (URL)* oder *Names (URN)*

Locators and Names



"The name of a resource indicates what we seek,
and an address indicates where it is."

Verwendung von HTTP



WWW (SS2001) - Grundkomponenten

21

Schema einer HTTP Interaktion

- *Verbindungsaufbau*
 - Aufbau TCP-Verbindung vom Client zum Server
 - normalerweise Port 80 des Servers
- *Request* vom Client zum Server
 - Auswahl einer Methode
 - zusätzliche Parameter zur Methode
- *Response* vom Server zum Client
 - Resultat in Form eines Statuscodes
 - zusätzliche Parameter zum Resultat
- *Verbindungsabbau*
 - normalerweise Abbau der Verbindung
 - neue Versionen (ab HTTP/1.1) können die Verbindung länger offen lassen

WWW (SS2001) - Grundkomponenten

22

Beispiel für einen HTTP Request

- TCP-IP Verbindung zu www.ethz.ch
- Erstellen der Verbindung "von Hand"

```
> telnet www.ethz.ch 80
GET / HTTP/1.1
Host: www.ethz.ch
```

Beispiel für einen HTTP Response

- Akzeptieren der Verbindung vom Client
- Interpretation des Requests

```
HTTP/1.1 200 OK
Server: Microsoft-IIS/4.0
Content-Location: http://www.ethz.ch/home_en.html
Date: Mon, 02 Nov 1998 09:24:53 GMT
Content-Type: text/html
Accept-Ranges: bytes
Last-Modified: Fri, 30 Oct 1998 07:55:15 GMT
ETag: "74335alda3be1:1b1a5"
Content-Length: 4229

<!DOCTYPE...
```

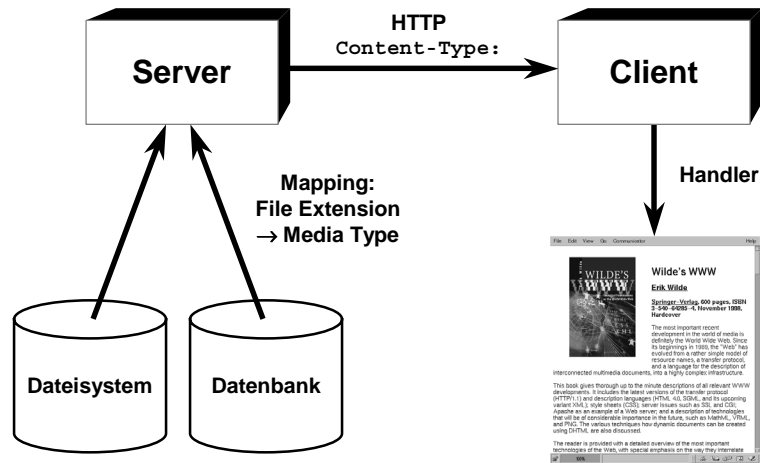
Multimedia-Aspekt des WWW

- ein Grossteil aller Inhalte ist HTML
- Verweis auf externe Ressourcen aus HTML
 - Bilder mit dem Element
 - Applets mit dem <APPLET> Element
 - allgemeine Objekte mit dem <OBJECT> Element
- Verweis auf externe Ressourcen durch URIs
 - <http://www.tik.ee.ethz/img/tiklogotitle.gif>
- dynamische Bestimmung des Inhaltstyps
 - Anforderung der URL beim Server (HTTP GET)
 - Interpretation des **Content-Type** Header-Feldes
 - entsprechende Darstellung beim Client

MIME-Types

- Multipurpose Internet Mail Extensions (MIME)
- ursprünglich standardisiert für Electronic Mail
 - Strukturierung und Typisierung von Mails
 - definiert E-Mail Header und Medientypen
- MIME-Types definieren einen Inhaltstypen
 - media type definiert das Medium
 - subtype definiert die verwendete Codierung
 - text/ascii, text/html, image/gif, image/jpeg
- Zuweisung von Inhaltstypen zu Programmen
 - werden von Browser oder Mail-Programm verwendet
 - z.B.: "image/*" wird von Photoshop verarbeitet"

Neue Inhaltstypen



WWW (SS2001) - Grundkomponenten

27

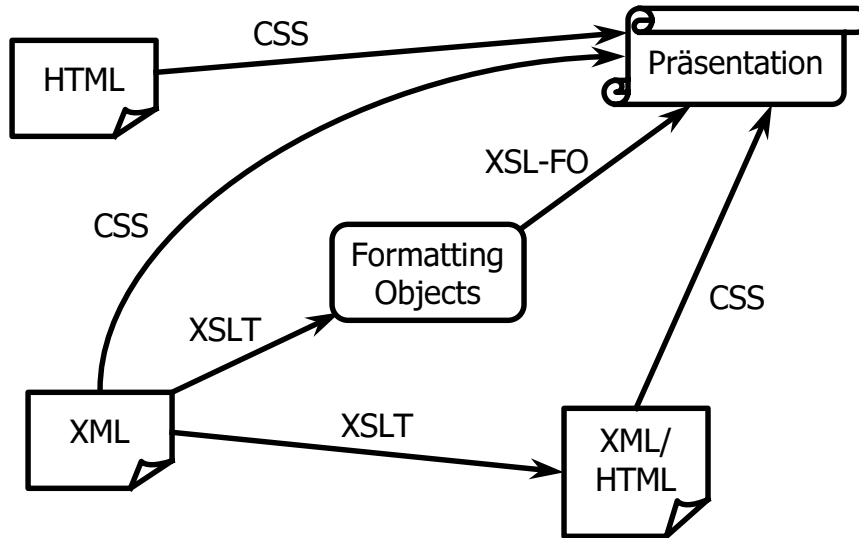
Cascading Style Sheets (CSS)

- Platzierung von Style-Informationen
 - verschiedene externe Files oder im Dokument
- Kaskadierung von Stylesheets
 - hierarchische Organisation von Stylesheets
- medienabhängige Stylesheets
 - Ergänzung zum medienunabhängigen HTML
- alternative Stylesheets
 - z.b. kompakt oder besser lesbar formatiert
- Performancefragen sind noch nicht geklärt
 - Einbussen und Gewinne sind denkbar

WWW (SS2001) - Grundkomponenten

28

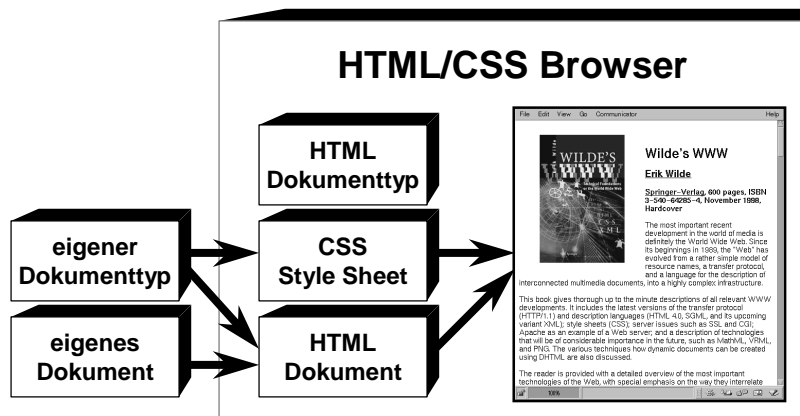
Verwendung von Style Sheets



WWW (SS2001) - Grundkomponenten

29

Publishing mit HTML/CSS



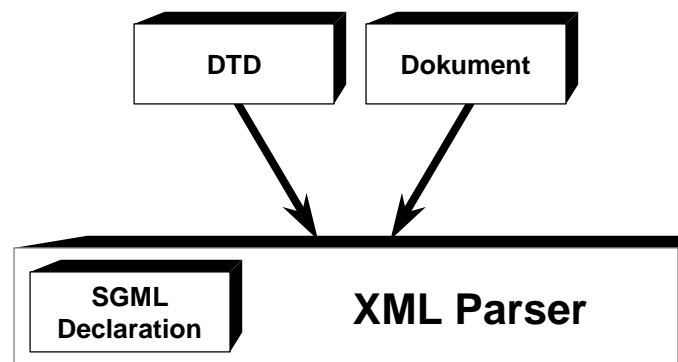
WWW (SS2001) - Grundkomponenten

30

Extensible Markup Language (XML)

- benutzerdefinierbare Dokumenttypen
- überwindet Einschränkungen von HTML
 - beliebige Dokumenttypen
 - neues Problem: Semantik von Elementen
 - begleitende Mechanismen werden notwendig
- überwindet Komplexität von SGML
 - fest definierte konkrete Syntax (SGML Declaration)
 - keine Markup Minimization (immer volles Markup)
 - reduzierte Zahl an erlaubten Attributtypen
- Ziele sind Einfachheit und Flexibilität

XML Parser



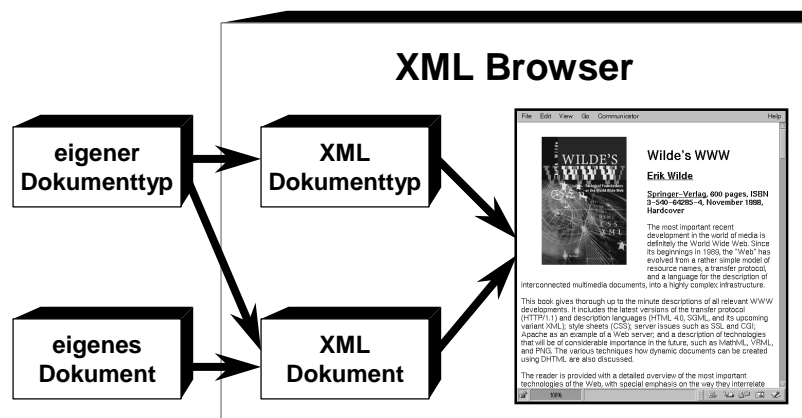
Vergleich SGML/XML/HTML

	SGML	XML	HTML
SGML Declaration	frei	fix	fix
DTD	frei	frei	fix
Dokument	frei	frei	frei

WWW (SS2001) - Grundkomponenten

33

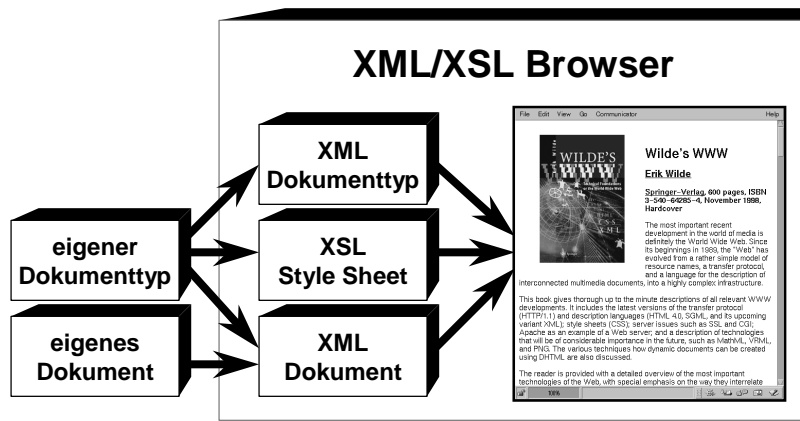
Publishing mit XML



WWW (SS2001) - Grundkomponenten

34

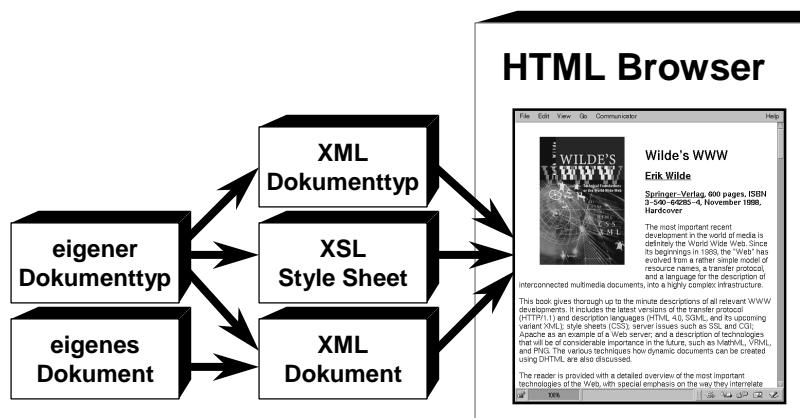
Publishing mit XML/XSL (Client)



WWW (SS2001) - Grundkomponenten

35

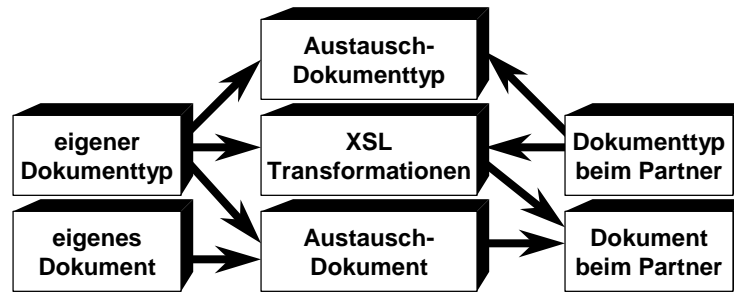
Publishing mit XML/XSL (Server)



WWW (SS2001) - Grundkomponenten

36

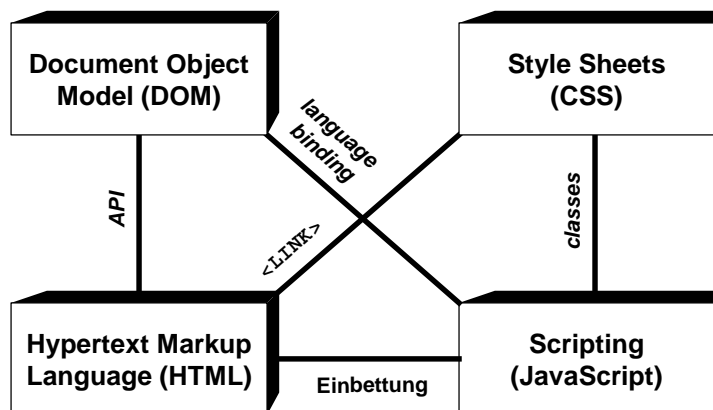
XML/XSL (Server) als B2B-Lösung



WWW (SS2001) - Grundkomponenten

37

Dynamic HTML (DHTML)



WWW (SS2001) - Grundkomponenten

38

Document Object Model (DOM)

- entstanden für portables JavaScript
- Programmierinterface für Dokumente
- unabhängig vom Dokumentenmodell
 - core definiert einen strukturellen Kern
 - HTML definiert HTML-spezifische Aspekte
 - XML erlaubt den Zugriff auf XML-Dokumente
- unabhängig von der Programmiersprache
 - Definition in Interface Definition Language (IDL)
 - language bindings für diverse Sprachen
 - bisher ECMAScript und Java

Zusammenfassung

- WWW als Sammelplatz für Technologien
 - erfunden als Hypermediasystem
 - benutzt als verteilte Benutzerschnittstelle
- interessante Konvergenzen
 - z.B. NeWS von Sun als "PostScript X Windows" heute neu erfunden als HTML und JavaScript
- viele Technologien werden im WWW recyclet
 - Java ist wohl das erfolgreichste Beispiel...
 - aber Recycling ist nicht immer erfolgreich...
- ein offenes System braucht offene Standards
 - viel Nachholbedarf bei vielen Firmen
 - Problembewusstsein oft nicht oder kaum vorhanden