MDAW

XML – extensible markup language

Dient zur Strukturierung und Beschreibung von Daten. Trennt Inhalt von der Formatierung
Ein Dokument ist ein XML-Dokument, wenn es wohlgeformt ist. Genügte s weiteren Beschränkungen, kann es gültig sein.

Wohlgeformt:
- Logische Struktur ist korrekt
- Alle Wohlgeformtheitsbeschränkungen sind erfüllt

Gültig(valid):
- Passende DTD (Document Type Definition) bzw. Schema
- Beschränkungen der DTD sind eingehalten

Inhalt von XML-Dokumenten

- XML-Dokumente enthalten Text
- Text = Folge von Zeichen (Markup/Zeichendaten)
- Zeichen = atomare Einheit von Text (ISO/IEC)
- Markup:
- Tags (Start/Ende/Empty)
- Entity-Referenzen
- …

- Text, der kein Markup ist, bildet die Zeichendaten

Wohlgeformtheit

Basiskriterien, die sich aus der Grammatik ergeben:

- Root element trägt Namen des DTD
- Es gibt nur ein root element
- Gültige Verschachtelungen der Tags
- Leere Elemente müssen abgeschlossen sein
- Attribute müssen in („) eingeschlossen werden
- Überprüfung durch Prüfprogramme (Validierer) möglich

Grundbegriffe:

Dokument-Type-Definition
- Legt Regeln für den Aufbau von Dokumenten fest
- Beschränkungen der logischen Struktur
- Gültigkeit des XML-Dokuments

Elemente:
Befehlssatz für das XML-Dokument: wird in der DTD festgelegt

Attribute:
- Parameter mit denen die Elemente näher spezifiziert werden
- Werden zusammen mit den Elementen in der DTD festgelegt

DTD

Externe DTD:
In separater –Datei, auf die verwiesen wird

Interne DTD:
Direkter Bestandteil der DTD

Eigenschaften der DTD:
- Mehrfach-Deklarationen möglich
- Interne DTD überschreibt externe DTD
- Intern + extern werden zusammengefasst (Gesamt-DTD)

Elemente:
Für das Dokument gültige Befehle

- Elemente gestehen aus einem Start-Tag und einem End-Tag
- End-Tag muss vorhanden sein
- Schachtelungen sind möglich

Element-Arten
- Inhalt besteht nur aus Elementen
- Inhalt besteht nur aus Daten
- Mischform: Inhalt besteht aus Elementen und Daten

Wurzel- oder Dokument-Element

- Muss in jedem XML-Dokument vorhanden sein
- Muss alle anderen Elemente einschließen (Container)
- Wurzelelement = Name der Dokumenttyp-Deklaration

Leeres Element ist erlaubt, wenn in der Deklaration „empty“ hinzugefügt wird.
Schachtelungen sind möglich, wenn Element mit „any“ deklariert ist.

Attribute:

Es gibt 3 verschiedene Attribut-Typen:
- Zeichenketten-Typ
- Aufzählungs-Typ
- Token-Typen (9 verschiedene)

Zeichenketten-Typ:
- Schlüsselwort CDATA
- Beliebige Zeichenfolge als Attribut zulässig

Aufzählungs-Typ
- Hat kein Schlüsselwort
- Wird erkannt durch Wertliste
- Attribut kann nur einen Wert aus der Liste annehmen
- Default-Wert in DTD festlegen

Token-Typen:

ID:
- Identifikations-Code für das Element
- Für jedes Element im Dokument eindeutig
- Seriennummer

IDREF/IDREFS:
- Verweist auf ein (mehrere) ID-Attribute
- Dadurch: Erstellung eines Links zum Element mit dieser ID

Attributvorgaben:
Muss ein Attribut gesetzt werden, soll es einen default-Wert haben?

#REQUIERED
Angabe des Attributs ist zwingend erforderlich

#IMPLIED
Freiwillige Angabe des Attributs

„wert“
Default-Wert bei nicht gesetztem Attribut

#FIXED „wert“
Attribut hat beim Setzen immer den default-Wert

Entities
Textbausteine definierter/externe Dateien einbinden

- Allgemeine Entities ⇔ Parameter Entities
- Analysierte Entities ⇔ nicht-analysierte Entities
- Interne Entities ⇔ externe Entities

Allgemeine Entities:
- Definition von Kürzeln für das Dokument
- Werden vom Parser durch ausführlichen Text ersetzt
- Schachtelungen erlaubt, Rekursionen nicht

Parameter Entities
- Werden in DTDs anstatt im Dokument verwendet
- Gleiche Verfahrensweise wie bei allg. Entities
- Kennzeichnung durch %

Analysierte Entities
- Kürzel im Dokument/DTD
- Ersetzungstexte werden nach Markups durchsucht

Nicht-analysierte Entities
Schlüsselwort: NDATA

XML ist eine auf Verwendbarkeit reduzierte Version von SGML
Die Standard Generalized Markup Language dient der Dokumenttypbeschreibung
XML beschreibt Dokumenttypen, nicht Dokumente
HTML ist eine Instantierung eines DTDs und ist (beinahe) in XML ausdrückbar
XML erlaubt durch DTDs selbstdefinierte Tags und baumartige Strukturierung von Dokumenten

Ein XML-Schema definiert
- Die Struktur einer Instanz eines XML Dokuments
- Genau wie bei DTDs wird definiert, welche Elemente enthalten sein können, welche Elemente welche anderen Elemente enthalten können, etc
- Den Datentyp eines jeden Elements bzw. Attributs
- Etwa „eine Zahl zwischen 20 und 28“
- Ist in DTDs nicht möglich

Warum Schema statt DTD?

- Unzufriedenheit mit DTDs, weil zwei verschiedene Sprachkonzepte eingesetzt werden müssen
- Verbesserte Typgebung bei den definierten Datentypen
- Wunsch nach vordefinierten Datentypen, wie sie auch in Datenbanken vorkommen

Moderne XML-Werkzeuge unterstützen bei de Dokumenttypbeschreibungssprachen
- Anwendung von XML-Spy
- Erstellung eines DTD
- Erstellung eines XML Schemas

Metadaten

- Metadaten enthalten Informationen über Daten
- Metadaten sind maschinenlesbare Informationen im Web

Anwendung von Metadaten:

- Kataloge
- Suchmaschine, Agentensysteme
- Electronic Commerce
- Digitale Signaturen, Privacy, Urheberschutz
- Inhaltsbewertung, Ablaufdatum, etc
- Recherche-Systeme, Annotationen, Blogs

PICS

- PICS = Platform for Internet Content Selection
- PICS erlaubt es, Dokumente mit Attributen(Labels) zu versehen.
- Ursprüngliche Intention: Eltern und Lehrern ein Instrument bereitzustellen, das es ermöglicht zu kontrollieren, auf welche Information Jugendliche zugreifen können
- Die Mechanismen für die Vergabe und Verwaltung von PICS-Attributen können auch für andere Zwecke eingesetzt werden

PICS-Prinzipien

- Herstellerunabhängig und –übergreifend
- Vergabe der PICS-Attribute durch den Informationsanbieter und durch zahlreiche, unabhängige Institutionen
- Leichte Verwendbarkeit: Eltern und Lehrer sollen leicht regulieren können, welche Information ihren Kindern zugänglich ist.

PICS-Aufgabenbereiche

- Definition eines Vokabulars für die PICS-Attribute
- Definiert 4 Kategorien (language, nudity, sex, violence)
- Zuweisung von Attributen an ein Dokument
- Verteilung und Verbreitung der Attribute
- Erstellen von Filterungs-Software
- Erstellen von Filterungs-Regeln
- Anwendung der Filterung

Anwendung von Metadaten

RDF

- RDF = Ressource Description Framework
- Allgemeine Metadaten und deren Struktur kann mit RDF abgebildet werden

- RDF kann andere Definitionsvorschläge für Meta-Information im Web abbilden
- RDF kann in XML repräsentiert werden (ist aber generell unabhängig von XML)

Semantic Web

Das Semantic Web ist die Repräsentation von Daten im World Wide Web in der Form, dass die Informationen eine wohldefinierte Bedeutung erhalten um Mensch und Maschine besser kooperieren zu lassen.

Es basiert auf dem RDF, welches eine Vielzahl von Applikationen integriert unter Anwendung von XML für die sprachliche Beschreibung und URLs für die Benennung.

Semantic Web Anwendungen:
- Datenrepresentation: RDF
- Subject/Praedikat/Objekt Triple
- Identifiziert durch URLs
- Keine Namenskonflikte

Kategorien:
- Datenintegration
- Datenabhängige Agenten
- Wissensmanagement
- Semantische Indices und semantische Portals
- Verwaltung persönlicher Daten
- Metadaten für Annotationen
- Metadaten für Beschreibung, Suche und Auswahl
- Metadaten für Medien und Inhalte
- Wissensgenerierung
- Management von Katalogen und Thesauri
- Syndication

Syndication:
- Veröffentlichung von Metadaten über Ressourcen
- Erlaubt Kategorisierung und Wiederbenutzung von Material

RDF
- Beschreibung eines Kanals
- Beschreibung einzelner Artikel, Bilder
- RSS reader erlauben
- Wiederauferstehung von Push
- Weit verbreitet in Blogs

XML/SGML

- SGML – Standard Generalized Markup Language:
- Ursprünglich für den (high-end) publishing Bereich eingeführt (70er)

- Zu allgemein
- Komplexität überfordert Browser
- HTML ist eine DTD in SGML
- XML Vereinfachung von SGML
- Strukturierte Datenspeicherung
- Strukturierter Datenaustausch

XML-Entwurfsziele

- Einfache Nutzung im Internet
- Viele Anwendungen ermöglichen
- Kompatibilität zu SGML
- Einfache Erstellung von Programmen
- Unterschiedliche Darstellung auf ein Minimum reduzieren
- Gute Lesbarkeit
- Schnelles Design
- Präziser und formaler Entwurf
- Leichte Dokumentenerstellung
- Platzsparende Darstellung ist von untergeordneter Bedeutung

Das Paket von XML:

- XML
- DTD/Schema
- DOM
- XPATH
- XSL/XSLT
- CSS
- Namespace
- XLink auch XLL
- XPointer
- XQL
- Web Services

Funktionen von XML – Smart Data

- Beschreibung von Metadaten (Daten über Daten)
  - Auf Basis der Struktur
  - Auf Basis des Inhalts
  - Auf Basis der Präsentation (Rendering)
  - Auf Basis der Verknüpfungen
- Transformation von Daten unterschiedlicher DTDs
- Metasprache
- Datenaustauschformat

XML-Visionen

- **Überbrückung v on Inkompatibilität**
  - Proprietäre Datenformate
  - Unterschiedliche Plattformen
  - Kommunikation von Applikationen

- **Mehrfache Verwertung einer Publikation**
  - Druck
  - Datenformat
  - Visualisierung
  - Sprachausgabe

XML-Visionen cont.

- Internationalisierung der Datenübertragung durch Unicode
- Verbesserte Informationssuche/Suchfunktionalitäten in Suchmaschinen
- Lokale Informationsverarbeitung (Rendering, XSLT)
- Entlastung von Servern und Netzsystemen
- Verbesserte Link Funktionen: Popup, Bidirektionalität
- Inhaltsbeschreibung und Definition der Inhaltsbezeichnungen beim Katalogisieren
- Knowledge Management durch Software-Agenten
- Beschreibung von Privacy Preferences

XML – Technischer Stand

- XML über CSS und XSL in der aktuellen Browser Generation
- Zahlreiche XML Applikationen in der Entwicklung z.B. XParse, Msxmi-Parser
- Einige Standards existieren vorerst nur als Working Drafts
- Viel Marketing, wenig Marktpenetration
- Weg von Proprietärität zu Akzeptanz als int. Standard schwierig

Zukunft von XML/SGML

- Parallele Existenz von HTML, XML und SGML
- XML in Netzsystemen
  - Datenspeicherung
  - Datenbeschreibung
  - Datenaustausch
  - Spezialapplikationen
- HTML
  - Datendarstellung
  - Skripting
- SGML
  - Grosse Dokumentationsprojekte, ansonsten obsolet

Definition Markierungssprachen

Eine Markierungssprache dient zum (semantischen) Formatieren von Daten und Texten in z.B.

- Textverarbeitungen wie MS Word
- Layoutpro9grammen wie z.B. QuarkExpress
- Internetseiten
- Multimedianwendungen z.B. CD-Rom Präsentationen mit Macromedia Director

Entstehung von SGML

1967 – William Tunnicliffe hielt Vortrag über die Trennung von Inhalten und Formatierungen
in Dokumenten

60er – Idee von Stanley Rice, Sprachenkatalog mit strukturierten Tags für Dokumente zu
entwickeln. Das CGA registrierte GenCode als Trademark, Gründung des GenCode
Comittee um eine solche Formatierungssprache zu entwickeln.

1969 – Charles Goldfarb, Edward Mooker und Raymond Lorrie entwickeln die Generalized
Markup Language (IBM) Ziel: Verarbeitung des gleichen Basisdokument für verschiedene Systeme

1978 – Goldfarb trit5t dem ANSI (American National Standards Institute) bei. Ziel: Textver-
arbeitungssprache auf GML basierend

80/85 – Entwicklung einer auf GML basierenden Textverarbeitungssprache. Prüfung durch
ISO (International Organisation for Standardization)

1986 – 15. Oktober Veröffentlichung von ISO 8879. Name: SGML

90er – Tim Berners Lee, Angestellter bei CERN entwickelt HTML

Kritiken

- Textverarbeitungssysteme wurden zunehmend besser
- Markup verschwand in den Hintergrund => Anwender dachten Systeme kämen ohne Markup aus
- EDV-Entwickler und Manager misstrauten dem ISO-Standard

DTP (Desktop Publishing)

Entwicklung von leistungsfähigen DTP Systemen (z.B. QuarkExpress)

Gesellschaftsbildung

- Bildung eins kleinen, geschlossenen Entwicklungskreis von SGML-Experten
- Trafen (und treffen) sich auf der IMC (International Markup Conference) in Amsterdam
- Außenstehende hatten zu diesem Kreis keinen Zugang
- Führte zu Ablehnung von SGML

Anerkennung von SGML

- Schnelles Bekannt werden und Akzeptanz von HTML machte SGML wieder populär
  - ⇨ steigendes Interesse an SMGL-Projekten
  - ⇨ Interesse an SGML Veranstaltungen
  - ⇨ man erkannte Notwendigkeit von Markierungssprachen
- Forderung nach weboptimierten SGML
- Entwicklung von XML (Mitte 90er)
  - SMGL – unnötiger Komplexität ü Netzwerkeffizienz

Wichtige SGML-Pionierprojekte

DAPHNE (Document Application Processing) Project (1984-86)
Austausch von technischen bzw. Wissenschaftlichen Texten über deutsche Forschungsnetze. Entwicklung eines dafür optimierten SGML. Erster SGML-Parser. Zusammensetzung von Dokumentteilen aus mehreren Quellen.

DOCDEL II Project (84-86)
Elektronisches Publizieren von Dokumenten. Erstellung von einem einheitlichen Format für die Ausgaben auf unterschiedlichen Textverarbeitungssystemen unter Verwendung von SGML

strukText (1986)
Strukturierung von Texten. Allgemeine SGML-Strukturen für den breiten Gebrauch standardisieren. Sah den Mangel an standardisierten Auszeichnungsschema (wie später HTML)

Entwicklung von zweisprachigem Wörterbuch (Brockhaus 87-88)
Entwicklung einer speziellen SGML-Sprache zur Strukturierung eines zweisprachigen Wörterbuches seit Anfang der 90er. Alle Wörterbücher und Enzyklopädien bei Brockhaus werden mit dem selbst entwickelten SGML basierenden Redaktionssystem erstellt und gepflegt.

Arten von Markierungssprachen

Es gibt 2 verschiedene Arten von Markierungssprachen

Besondere Formatierungssprachen (z.B. RTF und HTML)
Menge an Tags in Formatierungsmöglichkeiten beschränkt. Dokument nicht portabel d.h. kann nur von einem Programm z.B. Word interpretiert werden.

Verallgemeinerte Markierungssprachen (z.B. SGML und XML)
Markierung beschreibt Struktur eines Dokuments. Definition von wieder verwendbaren Objekten möglich.

Die Formatierungssprache RTF

RTF gehört zu den besonderen Formatierungssprachen. Sie besteht aus einer fest definierten Menge von Tags z.B.

- Zeilenumbrüche (/par) einfügen
- Text fett (/b), kursiv (/i) oder unterstrichen (/u) darstellen

RTF wird für die Programme WordPad und Word verwendet

RTF in MS Word

RTF formatierte Dokumente sind relativ leicht lesbar.
In MS Word sollte Lesbarkeit vermieden werden

⇨ Dokumente werden dort binär abgespeichert

- ⇨ Von außen keinen Zugang
- ⇨ Geschlossenes Markierungsformat

Struktur von SGML

Markierung beschreibt Struktur des Dokumentes, nicht dessen Formatierung oder Stilcharakteristika
Strenge Markierungssyntax

- Code kann von Software und Mensch gelesen werden

Möglichkeiten von SGML

Auszeichnung der Dokumentstruktur
Identifikation der Zeichen, die in einem Dokument verwendet werden sollen = Zeichenspezifikation

Identifikation von Objekten, die im gesamten Text verwendet werden sollen.
=>bequem verwendbar
=> Veränderung einer solchen Deklaration ändert jedes Vorkommen dieses Objektes im Dokument
Einbindung von externen Daten z.B. Bilder

Die Markierungssprache XML

- Offene, textbasierte Markierungssprache
- Liefert für Daten strukturelle und semantische Informationen
- Für das Web optimierte Untermenge von SGML
- XML ist eine Metasprache

=> kann zur Erzeugung von Untersprachen zur Lösung spezifischer Sprachen verwendet werden

- XSL: Formatierungssprache
- SMIL, CML, MathML, MusicML, TM,…: Anwendungen

Vergleich XML – SGML

- XML ist Teilmenge von SGML
- XML ist für das Web optimiert, kann mit HTML zusammenarbeiten
- Schlanker als SGML
- =>SGML Spezifikation umfasst hunderte Seiten
- XML unterstützt Stylesheets: erlaubt Erstellung von Stilen

Logische Struktur von XML-Dokumenten

Zeigt, wie ein Dokument aufgebaut ist welche Elemente in welcher Reihenfolge

Zwei Teile:

- Prolog =>Spezifikationen über den Dokumenttyp
- Element => eigentlicher Inhalt des Dokuments

Prolog

Besteht optional aus
- XML-Deklaration => identifiziert verwendete XML-Version
- Dokumenttyp-Deklaration
    - Grammatische Regeln für Dokumenttyp
    - Intern/extern

Element

- Das Dokument-Element enthält alle Daten
- Beliebige Anzahl verschachtelter Unterelemente
- Begrenzt durch Start- und End-Tags, dazwischen Inhalt
- Inhalt kann sein:
    - Elemente
    - Zeichendaten
    - Kommentare
- Kein Inhalt =>leeres Element

Physische Struktur

- Besteht aus dem gesamten Inhalt des Dokuments
- Unterteilt diesen in Entities (Speichereinheiten)
- Dokument-Entity ist Anlaufstelle für Parser
- Analysierte/nicht analysierte Entities
- Interne/externe Entities

Werden in der DTD deklariert
Auch durch Entity-Referenzen eingebunden

XML-Schema

Ein XML-Schema stellt den Dokumenttypen in reiner XML-Entity-Syntax dar. Dabei werden einfache und komplexe Typen definiert.

Einfache Typen (simple types)
- Elemente
- Attribute
- Restriktionen

Komplexe Typen (complex types)

Ein komplexer Typ beinhaltet andere Typen und/oder Attribute
Es gibt vier Arten von komplexen Typen
- Leere Elemente
- Elemente, die lediglich andere Elemente enthalten (Container)
- Elemente, die nur Text enthalten
- Elemente die Text und weitere Elemente enthalten

Indikatoren

Ordnungsindikatoren:
All, choice, sequence

Quantifizierende Indikatoren
Minoccurs, maxoccurs

Erweiterungen

Any Element
Erlaubt die Erweiterung des XML Dokuments um Elemente, die nicht im Schema definiert sind

Any Attribut
Wie oben, erlaubt das Erweitern von XML Dateien um Attribute, die ursprünglich nicht im Schema definiert sind

XSL – Stylesheet Language of XML

Grundlagen DOM und XPATH

Um bestimmte Elemente ansprechen zu können, müssen wir im Dokumentbaum navigieren können: XPATH
XPATH Elemente beschreiben absolute oder relative Beziehungen

XPATH Ausdrücke

- XPATH Ausdrücke sind absolute oder relative Identifikationen
- Oder Bedingung
- Oder Auswertungen
- Oder Flusssteuerungsanweisungen

XSL

Die extensible Stylesheet Language oder besser die Stylesheet Language für XML

XSL ist eine Beschreibung dessen, wie ein Prozessor ein XML-Dokument von einer Struktur in eine andere zu transformieren hat. Eine mögliche Transformation ist es, das XML-Dokument von einer semantischen Struktur in eine Anzeigestruktur zu überführen.

Durch Trennung von Daten und Formatierung ist es notwendig ein Werkzeug zur Verfügung zu haben, mit dessen Hilfe man in der Lage ist, die Daten umzustrukturieren. Ein Spezialfall der Umstrukturierung wäre die Darstellung der Daten in Form der Anzeigestruktur. Dabei dienen die XSL-Vorgaben als Vorlage, Muster für die Struktur des Ausgabedokuments.

XSL bietet die Möglichkeit die Struktur von XML-Dokumenten zu verändern. Dazu enthält die XSL-Vorlage sowohl XSL-Anweisungen wie auch Bestandteile der Zielstruktur. Die häufigste Anwendung ist die Überführung der XML-Daten in eine Ansichtsstruktur z.B. für einen Browser.

Für die Ansicht gibt es 2 Möglichkeiten:

- Für einen HTML-Browser erstellt der Host mittels XSL-Vorlage die Ansichtsstruktur
- Für einen XML-Browser stellt der Host nur das XML-Dokument und die XSL-Vorlage bereit und der Browser erzeugt die Ansichtsstruktur.

Mit XSL lässt sich eine Transformation auf Basis eines Teilbaumes definieren. Die damit festgelegten Regeln umfassen:

- Das Erkennen (matchen) eines bestimmten Elementes
- Das Festlegen der Struktur des darunterliegenden Subbaumes
- Die Art und Weise, wie der Subbaum bearbeitet werden soll

Stylesheets werden in Templates beschrieben

- XSL Stylesheets sind XML Dokumente
- Stylesheets geben eine Struktur wieder
- Templates können Inhalte aus dem Quelldokument einfließen lassen

Wesentlich interessanter ist der Einsatz von XSL templates zur Dokumenttransformation
Templates „matchen“ bestimmten Knoten (nodes, roots)

Um bestimmte Elemente ansprechen zu können, müssen wir im Dokumentbaum navigieren können: XPATH

XPATH Elemente beschreiben absolute oder relative Beziehungen zum aktuellen Node.

- XPATH ist eine Sprache zur Definition von Teilen eines XML Dokuments
- XPATH verwendet Pfade um XML Elemente zu definieren
- XPATH definiert eine Bibliothek von Standard Funktionen
- XPATH ist ein wesentliches Element für XSLT
- XPATH ist nicht in XML beschrieben
- XPATH ist ein W3C Standard

XSL besteht aus 3 Teilen
- XSLT
- XPATH
- XSL Formatting Objects

XSLT

CSS – Die Style Sheets für HTML
Weil HTML vordefinierte Tags verwendet, werden diese wohl verstanden und Browser wissen, wie die markierten Texte darzustellen sind. Erweiterungen der Darstellung durch Style Sheets in CSS sind einfach zu erstellen und einfach verständlich.

XSL – Style Sheets für XML
Weil XML keine vordefinierten Tags verwendet, Weiß weder der Mensch noch die Maschine wie ein XML Dokument darzustellen ist. Dazu gibt es XSL Style Sheets, die eine Transformationen festlegen. XSL Transformationen bestimmen die Regeln der Überführung von einem XML Dokument in ein anderes.

XSL Transformationen

- XSLT kann neue Elemente hinzufügen oder Elemente entfernen
- Es kann das Dokument neu arrangieren oder sortieren
- XSLT kann Elemente prüfen und anhand der Bedingungen feststellen, ob diese im Ergebnis wieder angezeigt werden sollen.
- Mittels XPATH werden Teile des Original XML Dokuments identifiziert (match)
- Ein identifiziertes Objekt wird gemäß Templates transformiert

XML in Nachrichtenagenturen

Nachrichtenagenturen sind Organisationen zur Sammlung von Nachrichten mit einzigem Zweck:
Übertragung und Verbreitung von Fakten an Nachrichtenunternehmen oder an Privatpersonen

Gremien der Nachrichtenagenturen

- NAA (News Association of America)
- IPTC (International Press Telecommunication Council)
- RTNDA (Radio- Television News Directory Association)

IPTC

1965 gegründet, vertritt die telekommunikativen Interessen von Nachrichtenagenturen der gesamten Welt. Seit den späten 70ern hauptsächlich Verabschiedung von Nachrichtenstandards.

Standards

ANPA 1312 / IPTC 7905
- Standards für Papier entworfen
- Heute noch in Gebrauch
- Eingesetzt in Fernschreiben
- Nicht verwendbar für Multimedia

- Nicht flexibel
- Keine Trennung von Inhalt und Metadaten
- Schlecht für Software verarbeitbar
- Durchsuchbar
- Viele menschliche Ressourcen nötig

IIM – Information Interchange Model

- Sollte ANPA/IPTC ersetzen
- 1992 fertig gestellt
- 1993 erneuert
- Datenformat zur Übermittlung von Informationen, Bildern
- Heute nicht mehr vom IPTC weiterentwickelt

NITF (SGML) – News Industry Text Format

- Erste Version (SGML) bereits 1989 fertig gestellt
- Erreichte weite Ersetzung von ANPA/IPTC
- Eigenschaften
    - Trennung von Struktur & Inhalt
    - Ermöglicht Referenzen zu externen Dateien
    - Starke Ähnlichkeit zu HTML
    - Langlebige Nachrichten
    - Flexibel

NITF(XML)

- 1998 erscheint erste XML-DTD des IPTC
- Integration der Vorteile von XML
- Hohe Flexibilität
- Multinationale Zeichenkodierung, Unicode
- XSL einsetzbar & XLL inkludiert

NewsML

- XML-Codierung von Nachrichten
- Entwickelt für deren Erstellung und Übertragung

Ziele:

- Medienunabhängigkeit
- Mehrfache Darstellung einer Nachricht
- Alle Formate und Medientypen werden gleichermaßen erkannt
- Erleichterung der Entwicklung/Erstellung von Nachrichtenelementen
- Simples Hinzufügen oder Ausblenden von Nachrichtenelementen
- Collections of news items
- Verknüpfung zwischen Nachrichtenelementen
- Datenanhang

Die Zukunft:

- Noch relativ unvollständig
- Textlicher Inhalt mit NITF geregelt

- Schnelle Entwicklung
- Wird NewsML alle alten Standards ablösen?

Akzeptanz bei etablierten Agenturen:

- Wird nach und nach umgesetzt
- Viele Kunden fordern Kontinuität
- Unruhiger Standardstatus noch abzuwarten

Omnipaper

Ein Projekt aus dem Information Society Technologies (IST) Programm unter dem fünften Rahmenprogramm RTD der europäischen Union

Agenda:

Ausgangslage
    Motivation, Problematik, Einschränkungen
Projektziele
    Objectives, Programm
Projekt Partner
    Partner aus Portugal, Spanien, Belgien und Österreich

Ausgangslage/Problematiken

- Unfassbare Menge an Information im Internet
- Ständig wachsende Anzahl an Nutzern und Hosts
- Notwendigkeit, Informationen zu verknüpfen
    - Mannigfaltige Zugriffsmöglichkeiten
    - Vergleichbarkeit der Informationen von verschiedenen Quellen
    - Assoziatives Verbinden von Informationen
- Vielfache Problematiken
    - Zugriff auf Information wird erschwert durch verschiedene Formate, Methoden, Protokolle, Strukturen, Plattformen, Software, Betriebssysteme, Sprachen, Gesetze, ….

Objektdefinition/Einschränkungen

- Konzentration auf Online Nachrichten/Zeitungen
- Inhalte liegen bereits digitalisiert vor
- Formate größtenteils Text und Bild, gut Verarbeitungsmöglichkeiten vorhanden
- Verteilt, Mehrsprachig, International

Projektziele/Objectives

- Ausgangslage
    - Wachsender Bedarf an Multi-Newspaper Access
    - Information ist zu umfangreich und unüberschaubar
    - Vernetzung bleibt schwierig
- Vorschläge, an diese Probleme heranzutreten. Information soll
    - Rascher erreichbar und auffindbar

    - Inhaltlich genauer
    - Und wertvoller durch selbstlernende Systeme werden

Projektziele/Objectives cont.

- Verbindung von Informationsquellen, die weit verstreut vorliegen, so-dass das Resultat mehr als bloß die Summe der Einzelinformationen darstellt.
- Hauptziel (key objective) ist die Erzeugung einer mehrsprachige Navi-gations- und Assoziationsplattform über verteilten Informationsquel-len in einer selbst-lernenden IT-Umgebung

Partner

- KULRD (Katolike Universiteit Leuven, Belgien)
- Uminhu (Universidade do Minho, Portugal)
- My News S.L. (Spanien)
- CURE (Center of Usability Research and Engineering, Österreich)
- UPM (Universidad Polytechnico de Madrid, Spanien)
- DAEDALUS S.A. (Spanien)
- Mediargus (Belgien)

Objectives/geplante Ergebnisse

- Navigation
    - Wie erreicht man einfache und intuitive Navigation
    - Einsatz von multilingualen Thesauri
    - Suchmöglichkeiten unterstützt von AI
    - Link Management
- Integration
    - Zusammenführung verteilter Information
    - Verschiedene Plattformen, Formate, Zugriffsmethoden

Erreichte Lösungen

- Architekturmodell eines europaweiten Nachrichtenarchives
    - Mehrsprachige Thesauri, Linking
- Allgemeingültiges Blueprint
    - Beschreibung der Projektergebnisse für allgemeinere Anwendun-gen
- Cross evaluation der Applikationsbereiche

Testbare Prototypen

- Prototypen-Entwicklung
    - Proof of concept
    - Technisch und inhaltlich detailliert umsetzen
    - Graduelle Vorgehensweise
    - Bedacht auf Benutzer-Integration und Mehrsprachlichkeit
    - Cross evaluation der Applikationsbereiche

Dokumenttypen definieren

- XML DTD
  - Elemente und Attribute definieren
- XML Schema
  - Elemente, Attribute, Datentypen und semantische Einschränkungen definieren

Dokumente transformieren

Quelldokumente zu Zieldokumenten

- XML Style Sheets
  - XSL Definition
  - Templates
  - Beispiele
  - Regeln
  - XPATH

DOM

DOM steht für document object model, das Dokumentobjektmodell
Spezifikation DOM (Level 1) von W3C als Empfehlung anerkannt im Oktober 1998
DOM Level 2 ist Empfehlung seit Nov 2000

Die DOM Programmierschnittstelle (API – Application programming interface) erlaubt:

- Definition der logischen Struktur eines XML oder HTML Dokumentes und Definition des Zugriffs und der Veränderung eines Dokuments
- DOM sieht Dokumente als logische Struktur, ähnlich einem Baum oder Wald, strenge Gliederung der Objekte

Dokumente mit DOM

- Anwendung des „Baum-Modells auch auf XML Dokumente
- Ein Dokument folgt einer vorgegebenen Datenstruktur
- DOM betrachtet Dokumente als aus individuellen Objekten zusammengesetzt
- Bearbeitung bzw. Identifikation z.B. mit XPATH

XML-Parser

Zwei verschiedene Ansätze für XML Parser:

- Baumbasierte Parser
- Ereignisbasierte Parser

Simple API for XML

SAX, die Simple API for XML ist eine Standardschnittstelle für ereignisgesteuerte XML Parser, die in Zusammenarbeit mit den Mitgliedern der XML-DEV

mailing list entwickelt wurde. SAX 2.0 wurde am Freitag, den 5. Mai 2000 veröffentlicht und ist für die kommerzielle, wie auch die nicht kommerzielle Verwendung frei.

Parserdefinition
Ein Parser nennt man ein Programm, das zur syntaktischen Analyse ein Quellprogramm als Eingabe bekommt. Dies erhält dieser aus der lexikalischen Analyse vom so genannten Scanner und in Form einer Folge von Token. Er hat die Aufgabe, daraus einen Ableitungsbaum zu erstellen und hinsichtlich syntaktischer Korrektheit zu überprüfen.

XHTML

XHTML ist
Eine W3C Recommendation seit 26 Januar 2000

Drei Ausprägungen sind definiert
- XHTML Strict
- XHTML Transitional
- XHTML Frameset

XHTML 2.0 ist ein W3C Working Draft seit Mai 2003

XML für Grafiken und Multimedia

Problem:
Das Web benutzt Grafik-Formate für 2D und 3D: Diese sind generell Pixel basiert.

Zoomen führt zu unschönen Ergebnissen. Lösung: Verwendung von Vector basierten Formaten. Diese sind durch mathematische Funktionen beschrieben.

Scalable Vector Graphics

SVG ist eine XML Sprache zur Beschreibung von zweidimensionalen Grafiken in XML.

SVG biete drei Typen von Grafik-Objekten
- Vector Grafiken
- Bilder
- Text

SVG 1.0 ist eine W3C Recommendation seit Sept 2001
SVG 1.1 ist eine W3C Recommendation seit Jan 2003
SVG 1.2 wird zurzeit entwickelt

SVG wird nicht direkt im IE oder Netscape unterstützt. Trotzdem gibt es Unterstützung in vielen Software-Tools:
- SVG2PDF
- SVG Toolkit
- Jakaroo

Extensible 3D (X3D)

Component based approach (profiles)

- Scalable, thin clients, DOM support
- Core model
- Minimum set of viewer functions
- 3D geometries, animations, interactions, rendering
- X3D-VRML profile
- Compatibility to VRML97
- User specific profiles

Major X3D contributors

- Sun
- Shout interactive
- Blaxxun
- Draw
- Sony
- NIST
- Lucid Acutal
- NPS

X3D process

- Actual: Implementation & Evaluation
- Next: standardization
- Web3D is a member of W3C

SMIL

Synchronize Multimedia on the web using an XML based language

- To enable simple authoring of TV-like multimedia presentations such as training courses on the web
- HTML-like, easy to learn
- A smile presentation can contain components of streaming audio, streaming video, Images, text or any other media type

SMIL Authoring Tools
Allaire Homesite
CWI SMIL Validator
Oratrix Grins
….

SMIL Players
Quicktime 4.1
Compaq HPAS
Real player 7+
….

VoiceML/VoiceXML

Designed for
- Synthesized speech
- Digitized audio
- Regognition of spoken and DTMF key input
- Recording of spoken input

MusicML – DTD for sheetmusic
- No standards
- Connection factory, dutch enterprise

MML – DTD for Music
- Organization elements
- Time elements
- Frequency elements
- Lyric elements
- Notation elements

MathML
Ist eine Auszeichnungssprache die speziell für mathematische Ausdrücke gedacht ist.

WAP – Wireless Application Protocol
Übertragungsprotokoll für mobile Endgeräte

Markupsprache, die auf XML basiert. Sprache orientiert sich an den beschränkten Möglichkeiten mobiler Endgeräte, vor allem
- Kleine, oft monochrome Displays
- Beschränkte Eingabemöglichkeiten
- Netzanbindung mit geringer Bandbreite
- Begrenzter Speicher und Rechenkapazität

Anwendungsmöglichkeiten
- Handys
- PDAs

Aufgaben von WML – Wireless markup language
- Textdarstellung und Layout
- Strukturierung der Information
- Navigation zwischen Dokumenten
- Interaktion mit Benutzer

Struktur eines Dokuments
- Ein WML Dokument ist immer ein gültiges XML Dokument
- Deck: Gesamtes WML Dokument
- Card: Informationseinheit, die dargestellt wird

Deck-Ebene

Zugriffskontrolle
- Access Element: Zugriff nur von bestimmter Domain und Pfad möglich

Template:
- Events, die auf Deck-Ebene definiert werden, sind für alle cards gültig. Überschreiben der Events in cards möglich
- Do-Element

Card Ebene
Eine card wird komplett auf dem Display dargestellt
Beinhaltet
- Text
- Eingabeelemente
- Events, die nur auf card-Ebene gültig sind
- Links

Events:

Arten von Events:
- Onenterforward: wird beim Aufruf der card ausgelöst
- Onenterbackward: wird durch einen prev-task ausgelöst
- Ontimer: wird von Timer ausgelöst
- Onpick: wird bei einer Auswahlliste durch Auswahl oder Abwahl ausgelöst

Mögliche Aktionen:
Go, prev, noop, refresh

Do-Element

Weist einer Funktionstaste eine der folgenden Funktionen zu:
- Go: Sprung zu einer URL
- Prev: Springt eine Seite zurück
- Refresh: Bewirkt ein Update der Browser-context
- Noop: No Operation

Statusmodell

Der Status des User Agent beinhaltet:
Variablen
History
Implementierungsabhängige Zusatzinformationen

Wird in card-Element zurückgesetzt, falls zugehöriges newcontext-Attribut den Wert true besitzt

History

Wird vom User Agent (Browser) in Form eines Stacks organisiert. Speichern von URL und einigen anderen Daten, aber kein Cache. Prev-task ruft letzte Card wieder auf.

WAP-Entwicklungsumgebungen

- Nokia WAP Toolkit 2.1
- Ericsson WAP DIE SDK 2.0
- Phone.com UP.SDK 4.1

Notwendigkeit

- Ständig wachsende Anzahl von Informationen
- Austauschformat für Metadaten nötig
- Neutral – für jede Anwendung nutzbar
- Maschinenverständlich
- Maschinenverarbeitbar
- Muss über Computernetze übertragbar sein
- Soll erweiterbar wein
- Verständlich für den Menschen

RDF-Datenmodell

- Gerichteter und beschrifteter Graph
- Definiert eindeutig die Zuordnung zwischen Subject, Predicate und Object
- Subject (Resource): zu beschreibende Quelle
- Predicate (property): in einer Ressource definierte Eigenschaft
- Object (Literal): Wert der Eigenschaft (Wert oder wieder Ressource)

RDF-Grobüberblick

- Standardisiertes Datenmodell für Metadaten
- PICS – Schema zur Beurteilung von Internet-Inhalten
- DC (Dublin Core) – Schema zur Beschreibung von Inhalten

Damit Maschinen die Bedeutung von Metadaten erfassen können, sind Standard-Schemas nötig
Basis-Elemente der Semantic-Web Initiative des W3C

Zusammenfassend:

Es gibt einige Standards für DTDs/Schemata für bestimmte Themenbereiche und Medientypen.
Für Interoperabilität sollten solche Standards auch genutzt werden

CM als Anwendungsgebiet des XML

- Content Management richtig verstehen
- Content Management zielgenau einsetzen
- Von modernem Content Management profitieren

…richtig verstehen

- Was ist Content Management?
- Welche Komponenten sind daran beteiligt?
- Wozu dienen diese Komponenten?

…zielgenau einsetzen

- E-content wird durch Dienstleiter einfacher nutzbar
- Moderne Redaktionssysteme bieten Administration für jedermann
- Content Syndication offeriert Zugriff auf vielfältiges Inhaltsangebot
- Kontrollierter Einsatz dieser Hilfsmittel trifft genau die Bedürfnisse der Kunden

…erfolgreich nutzen und davon profitieren

- Qualitativ hochwertiger Inhalt bindet Kunden an den Dienstag Professionelle Content-Provider erleichtern Content-Integration
- Einfache Wartung der eigenen Homepage durch modernste Technik

Content Syndication

Wen fragen, wenn es um Inhalte geht?

Informationsbeschaffung
- Integration vielfältigster Quellen

Informationsverwaltung
- Content Management in aktuellsten Formaten

Informationsauslieferung
- Versand und Zugriff nach modernster Technologie

Content Administration

Wie verwalte ich meine eigene Site?

Administrationsansätze
- Seite für Seite
- Site-orientiert
- Redaktionssysteme

Wichtigste Punkte
- Einfache Benutzbarkeit
- Schnell und stabile Installation und Wartung
- Gute Erweiterbarkeit, Zukunftssicherheit

Content Integrationsvarianten

Wie übernehme ich externen Content?

Content Teaser
- Integration von Verweisen auf den content

Content als Applikation
- Integration von Content Visualisierungssoftware (ASP)

Content als Produkt
- Physikalische Integration von Content

Web Services

Die W3C koordiniert in seiner Web Services Activity mehrere Gruppen, die sich um den Informationsaustausch zwischen Applikationen kümmern.

- Das World Wide Web wird mehr und mehr Kommunikation zwischen Applikationenbenutzt. Das programmtechnische Interface dazu wird mit Web Services bezeichnet.
- Arbeitsgruppen zu den Themen
    - Web Service Architecture
    - XML Protocol/Web Service Protocol (XML/SOAP)
    - Web Service Description Working Group (WSDL)
    - Choreografie

Web basierte XML Services für Kommunikation und Datenaustausch zwischen Applikationen
- Service Aufrufe über http und XML (Standards)
- Web Service Protokolle und Dienstleistungen
- Einsatz von SOAP im EU-Projekt OmniPaper

Verteilter Zugriff auf Ressourcen hat bereits längere Tradition
- Sun RPC
- OMGs CORBA
- MS' DCOM, COM+

Lösungen für LAN und WAN ganz gut
- Web Services zielen auf Internet ab
- Keine Protokolleinschränkungen durch Firewalls
- Standardisierte Formate der Datenübertragung

Web Service Begriffe

- SOAP (Simple object access protocol)
- XML als Präsentationsschicht
- HTTP als Transportprotokoll
- UDDI (Universal Description, Discovery & Integration) als “DNS” für Web Services
- WSDL (Web Service Description Language)

Das SOAP Protokoll

- SOAP erlaubt Funktionsaufrufe auf entfernt gelegenen SOAP Servern
- Es bedient sich http als Transportprotokoll und verwendet XML als Notation für die Argumente und Funktionsergebnisse
- Funktionsaufrufe und Ergebnisse sind in so genannten SOAP Envelopes eingepackt
- Ein Aufruf besteht aus Anfrage und Antwort
- Die genaue Syntax wird vom W3C standardisiert

UDDI: Suchen und Finden

- Zur Nutzung eines Web Services muss er bekannt sein bzw. gefunden werden können
- Mit DNS vergleichbar
- Erste Implementierungen verfügbar (UDDI4J bei IBM)
- Beschreibung der Funktionen in WSDL

WSDL:

- Beschreibung von Web Services
- Hier wird beschrieben, wie auf einen Web-Service zugegriffen werden kann
- Beschreibungen in XML abgefasst
- Aus WSDL können automatisch Programmstubs erzeugt werden

Implementierungen

- MS: .net
- IBM: Geschenk an Apache, WSTK
- Sun: ONE, Brazil, Forte for Java
- Borland: Delphi direkte Unterstützung für SOAP Programmierung
- Open-Source: Apache-Projekt, SOAPLite für PERL

Topic Maps

Eine Topic Map ist eine Sammlung von Topics und (inhaltliche) Beziehungen zwischen diesen Topics. Topic Maps verlinken die Ressourcen hinter URLs. XTM dient als XML-basiertes Austauschformat (Repräsentationsformat) für Topic Maps.

Topic Maps sind aufgesetzte semantische Linknetzwerke. Die Verbindung zwischen Topics und Ressourcen sind URLs. Topic Maps betreffen Objekte der realen Welt als auch abstrakte Begriffe und Konzepte, wie zB „SVG“. Sie werden in Topic Maps nicht absolut definiert, sondern relativ zueinander.

Topic Maps können mit nicht vollständigem Wissen umgehen. (I know that Prince Charles was married but I do not know the name of his wife)
Topic Maps können zusammengeführt werden (Maybe someone else knows that someone called Diana was married to a british prince)

Topic Maps sollten möglicherweise mit tausenden Topics arbeiten. Wurden gebaut um Informationen zu beschreiben, nicht Wissen. Wurden nicht für eine spezifische Applikation gebaut, sondern sollen in verschiedensten Kontexten einsetzbar sein.

Topic Maps bestehen aus
- Topics
- Assoziationen
- Occurences

Topics (Objekte) können sein
- Web Ressourcen (Stock, Quotes, documents)
- Reale Objekte (people, countries)

Topics haben Charakteristiken
- Names: welche Bezeichnung hat das Topic
- Occurences: welche Objekte werden repräsentiert
- Types: is a relationship mit anderen Topics

Topic Names

Jedes Topic hat eine eindeutige ID innerhalb einer Topic Map. Diese ID ist für den internen Gebrauch. Jedes Topic kann (einen oder mehreren) Namen haben.

Topic Occurences

Referenzieren externe Ressourcen
- Dokumente: via URLS
- Definiert durch URNs
- Nicht definiert, aber global eindeutig:

Ein Topic kann beliebig viele Occurences haben

Topic Types

Ein Topic kann beliebig viele Typen haben. Jeder Typ ist wiederum selbst ein Topic
Entweder innerhalb derselben map oder definiert über ein anderes Dokument

Topic Typen bestimmen eine Typenhierarchie. Jede Topic Map hat ihre eigene Typenhierarchie. Es existiert kein globales Typensystem (keine Ontologie).

Associations

Topics können Teil von Beziehungen (Assoziationen) sein

Topics spielen darin
- Eine Rolle
- Und sind ein member

Typische Assoziationen
- Is loacated in, lived in, written by
- Is facility provided by, requires to have

Probleme
- Es ist unklar, welcher Art die Assoziation ist (is located in)
- Es ist unklar, was wo liegt (=>Rollen einführen)

Für alle neuen Topics kann definiert werden (is located in, building, location)
Alle Topics können mit weiteren Assoziationen vernetzt sein.
Assoziationen können eine beliebige Anzahl von members haben.

In der Praxis ist es oft schwierig entsprechende Assoziationen zu finden. Topics können auch dieselbe Rolle in einer Assoziation spielen (Die Regierung besteht aus einem Kanzler und mehreren Minister)

Ressourcen/Occurences

Jedes Topic kann
- Entweder auf externe Ressourcen verweisen (resourceRef)
- Oder Inhalte direkt enthalten (internal resource)

Scopes

Nicht alle Topic Charakteristiken sind in allen Kontexten gültig. Scopes schränken die Charakteristiken auf einen bestimmten Bereich ein. Scopes sind wiederum selbst Topics.

Wenn kein Scope definiert wurde, so sind die Charakteristiken in allen scopes gültig
- ⇨ unconstrained scope