XML und Datenbanken

Kapitel 0: Organisatorisches

Prof. Dr. Stefan Brass

Martin-Luther-Universität Halle-Wittenberg

Wintersemester 2023/24

http://www.informatik.uni-halle.de/~brass/xml23/

Inhalt

Themen

- XML Syntax (Daten und DTDs)
- XML Schema
- XPath / XQuery Data Model (XDM), XML Infoset
- XML Parser Bibliotheken in Java: DOM, SAX (kurz)
- XPath
- XQuery
- XSLT (kurz)
- JSON (kurz)
- XML-Unterstützung in SQL (Oracle, DB2)

Falls noch Zeit (unwahrscheinlich): Speicherstrukturen für XML.

Ggf. mehr zum semantischen Web: RDF, SPARQL

Voraussetzungen

Dies ist eine fortgeschrittene DB-Vorlesung (für Master-Studenten).

Vorausgesetzt werden:

- Relationales Modell, ER-Diagramme
- SQL (komplexere Anfragen, `CREATE TABLE`)
- Programmierkenntnisse, Syntaxdiagramme
- (mehr oder weniger:) Englische Sprachkenntnisse

Nicht (unbedingt) vorausgesetzt: XML

- Wer hat „Grundlagen des WWW“ gehört?
- Wer kennt XML? Wer kennt HTML?

Interessieren Sie sich für das Thema?

- Es ist ein Problem, dass diese doch sehr spezielle Vorlesung offenbar zumindest gefühlt für Wirtschaftsinformatiker zu einer Art Pflichtvorlesung geworden ist.

  Das war nie abgesprochen. Während ich schon denke, dass jeder Informatiker, Bioinformatiker und Wirtschaftsinformatiker XML kennen sollte, ist das z.B. bei XQuery weniger klar. Diese Vorlesung beschäftigt sich aber zu einem großen Teil mit solchen fortgeschrittenen Technologien.

- Nicht wenige Master-Studierende der Wirtschaftsinformatik haben unzureichende Vorkenntnisse (weil BWL Bachelor).

  Das relationale Modell, SQL und Datenbank-Entwurf im ER-Modell werden wirklich vorausgesetzt! Wenn Ihnen das fehlt, müssen Sie zuerst „Einführung in Datenbanken“ hören.

- Wenn Sie diese Vorlesung nur gezwungenermaßen belegen, und ohne Vorkenntnisse, wird es für beide Seiten frustrierend.

Inhalt

Semistrukturierte Daten (1)

- Relationale Daten gelten als stark strukturiert:
  - Das Schema ist DBMS und Nutzern bekannt, sehr stabil (ändert sich nur minimal).
  - Die einzelnen Tabelleneinträge sind atomar, Auswertungen ohne manuelle Hilfe möglich.
- Texte (auch Bilder etc.) gelten als unstrukturiert:
  - DBMS-Sicht: nur Folge von Zeichen/Worten.
  - Die inhaltlich interessanten Strukturen sind dem DBMS nicht bekannt $\rightarrow$ Keine Hilfe bei Suche.

    Z.B. Warenbezeichnungen/zugehörige Preise, falls Katalog.

Semistrukturierte Daten (2)

- Mit dem Web und XML sind semistrukturierte Daten aufgekommen:
  - Zum Teil sind inhaltlich interessante Strukturen mit Tags markiert, zum Teil einfach Text.
  - Mit den Tags wird häufig sehr frei umgegangen, die Strukturen sind unregelmäßig.
  - Die Tags (Element-Typen) / das Schema sind nicht unbedingt vorab bekannt.
  - Das Schema ist ständiger Änderung unterworfen.

    Daten entstehen häufig durch Integration autonomer Quellen.

Semistrukturierte Daten (3)

- In Anfragen an relationale Datenbanken kann man sich nur auf bekannte Spalten fester Tabellen beziehen.

  Bei Bedarf kann man das Data Dictionary abfragen, um sich zuerst die Schema-Information zu beschaffen. Dies ist aber ein getrennter Schritt: Man braucht erst die Ergebnisse dieser Abfrage, um die Abfrage an die eigentliche Datenbank zu formulieren.

- Bei XML-Daten macht es dagegen Sinn, dass man sich den Inhalt von beliebigen Tags, die auf „name“ enden, an beliebiger Stelle im Dokument anzeigen lassen möchte.

  Mit einem intelligenteren Texteditor könnte man auch danach suchen.

Motivation (1)

- Es geht viel schneller, Daten im XML-Format zu erfassen, als eine relationale Datenbank anzulegen.

  Solange die Datensammlung klein ist, reicht ein Texteditor. Trotzdem können die Daten so strukturiert sein, dass man mit XQuery alles an Anfragen/Auswertungen berechnen könnte, was auch mit einem relationalen DBMS möglich wäre. Das Risiko ist allerdings, dass die Daten im Laufe der Zeit immer schlechter strukturiert werden (wenn man nicht bewusst eine DTD/ein Schema entworfen hat und die Einhaltung erzwingt).

- XML ist ein wichtiges Daten-Austauschformat.

  Selbst wenn man seine Unternehmens-Daten in einer relationalen DB hält, wird man in der Kommunikation mit Geschäftspartnern etc. XML benutzen. Es ist dann nützlich, wenn man manche Auswertungen auch direkt auf den XML-Dateien durchführen kann.

Motivation (2)

- XML eignet sich auch zur langfristigen Aufbewahrung (Archivierung) von Daten.

  XML ist ein firmenunabhängiger Standard, der allgemein anerkannt, verbreitet, und sehr stabil ist. Man kann davon ausgehen, dass es auch in 10–20 Jahren noch funktionierende Software geben wird, die heute erstellte XML-Dateien lesen kann. Das wird auch dadurch unterstützt, dass viele XML-Software „Open Source“ ist, man sie also ggf. auf einem dann aktuellen Betriebssystem neu compilieren könnte, bzw. erforderlichenfalls anpassen. Die Struktur von XML-Daten ist auch relativ einfach. Wichtig ist natürlich, dass die konkrete Anwendung von XML gut dokumentiert ist (Bedeutung der verwendeten Tags etc.).
  Im Gegensatz dazu sind die Dateien von Datenbanksystemen immer systemspezifisch, und können sich von Version zu Version ändern. Datenbank-Software ist auch sehr komplex.

Motivation (3)

- XML unterstützt komplex strukturierte Objekte.

  Am relationalen Modell wird kritisiert, dass man die Objekte zur Speicherung in einfache Tupel zerlegen muss.

- Manche Leute meinen, XML sei „die Zukunft“ auch im DB-Bereich: XQuery wird als das „SQL des 21. Jahrhunderts“ bezeichnet.

  Ich halte das für stark übertrieben. Obwohl man feststellen muss, dass vieles von der schönen Einfachheit, die das relationale Modell gebracht hat, heute kaputt gemacht wird: Alles wird immer komplexer (meiner persönlichen Meinung nach häufig ohne rechte Not).

- Es gibt für XML sehr viele freie Werkzeuge.

- Interessante neue Forschungsprobleme.

Motivation (4)

- Beispiele für Anwendungen von XML:
  - XHTML
  - Google Sitemaps
  - SEPA-Lastschriftmandate (z.B. Verein muss eine XML-Datei an seine Bank schicken, um Mitgliedsbeiträge einzuziehen).
  - Offenlegung/eBilanz für Firmen

    Ich habe eine UG (kleine GmbH) für Feuerwerke und muss das machen.
  - Daten für Firewall in Linux

    Zumindest „Rocky Linux“, Definition von Zonen und Services.
- JSON ist ziemlich ähnlich zu XML (aber einfacher).
  - XQuery 3.1 hat auch Unterstützung für JSON.

Inhalt

Zeit und Ort

Vorlesung/Seminar (2 SWS):

- Dienstags, 12^{15}–13^{45}, Raum 3.07.

  Die Vorlesung wird aufgezeichnet. Die Aufzeichnung steht nach einigen Stunden in StudIP, Reiter „OpenCast“ (muss dann noch geschnitten werden). Sie sind auf der Aufzeichnung praktisch nicht zu hören, wenn Sie eine Frage stellen.

Übung:

- Gruppe 1: Mittwochs, 10^{15}–11^{45}, Raum 3.02.
- Gruppe 2: Mittwochs, 12^{15}–13^{45}, Raum 3.02.

  Raum 3.02 ist der Multimedia-Pool. Es soll auch praktisch am Rechner geübt werden. Natürlich werden auch die Hausaufgaben besprochen.

- Beginn in der ersten Semesterwoche.

Zeitliche Belastung

- Diese Vorlesung hat 5 Leistungspunkte.

  Auch „credit points“ genannt.

- Entspricht 150 Stunden studentischer Arbeitszeit:

Lernform	SWS	Stunden
Vorlesung	2	30
Selbststudium	0	45
Übung	2	30
Hausaufgaben	0	30
Spezielle Prüfungsvorbereitung	0	15

Das Selbststudium ist wirklich wichtig. Lesen Sie ein Buch oder passende Internet-Quellen.

Inhalt

Ansprechpartner (1)

Dozent: Prof. Dr. Stefan Brass

- Email: brass@informatik.uni-halle.de

  Betreff-Zeile sollte mit [xml23] beginnen, möglichst aussagefähig.

- Büro: Von-Seckendorff-Platz 1, Raum 313
- Telefon: 0345/55-24740
- Sprechstunde: Montags, 12^{00}–13^{00}
- Frühere Unis: Braunschweig, Dortmund, Hannover, Hildesheim, Pittsburgh, Gießen, Clausthal.
- Oracle8 Certified Database Administrator (aktuell: 19C/21C).
- IBM Certified Advanced DBA (DB2 UDB 8.1).

Ansprechpartner (2)

Übungsleiter: MSc. Mario Wenzel

- Büro: Von-Seckendorff-Platz 1, Raum 315
- Telefon: 0345/55-24776
- Email: mario.wenzel@informatik.uni-halle.de

Sekretärin: Ramona Vahrenhold

- Büro: Von-Seckendorff-Platz 1, Raum 324
- Telefon: 0345/55-24750, Fax: 0345/55-27333
- Email: vahrenho@informatik.uni-halle.de

Webseite

http://www.informatik.uni-halle.de/~brass/xml23/

- Aktuelle Ankündigungen
- Folien der Vorlesung (PDF, es gibt auch Druckversion)
- Übungsblätter
- Alte Klausuren
- Verweise auf Literatur im WWW

  Zu WWW-Themen gibt es sehr viel nützliche Literatur im WWW selbst (z.B. Standards, Tutorials). Falls Sie empfehlenswerte Quellen finden, schicken Sie mir bitte eine EMail mit der URL.

- Verweise auf Software im WWW

Inhalt

Studienleistung (1)

- Die Studienleistung ist Voraussetzung für den erfolgreichen Abschluss des Moduls (d.h. die LP).

  Die Studienleistung ist nicht Voraussetzung für die Teilnahme an der Prüfung.
  Es macht aber keinen Sinn, die Klausur ohne Studienleistung zu schreiben.
  Hausaufgaben bereiten auf die Klausur vor und sind sowieso noch zu erledigen.

- 50% der Hausaufgabenpunkte

  Wenn es möglich ist, die Lösung zu testen (z.B. mit einem Validator),
  gibt es 0 Punkte für offensichtlich nicht getestete Lösungen (z.B. Syntaxfehler).

- Die Hausaufgaben sind einzeln zu bearbeiten.

  Helfen Sie Ihren Mitstudierenden, aber geben Sie nicht Ihre Lösung zum Abschreiben. Minimal 0 Punkte für alle Beteiligten bei „zu ähnlichen“ Lösungen.

- Außerdem wird aktive Mitarbeit in den Übungen verlangt, dazu gehört auch das Vorstellen von Hausaufgaben.

  Es ist kein Problem, wenn Sie ca. drei Mal fehlen. Sonst siehe nächste Folie.

Studienleistung (2)

- Es ist möglich, dass es in den Übungen Präsenzaufgaben gibt, die auch Hausaufgabenpunkte liefern.

  Diese Aufgaben können nur bei Übungsteilnahme abgegeben werden.
  Sie machen aber nur einen kleinen Teil der Punkte aus (maximal 15%).

- Falls Sie häufig nicht an der Übung teilnehmen können:
  - Reden Sie mit dem Übungsleiter oder dem Dozenten.
  - Die Übung kann nicht aufgezeichnet werden, da ja die Studierenden sprechen sollen.

    Eventuell gibt es auch in den Übungen einige Folien, aber damit wird keineswegs die ganze Übung abgedeckt. Z.B. nicht die Diskussion der vorgestellten Hausaufgaben und die Antworten auf Fragen.

  - Sie müssen davon ausgehen, dass Sie zusätzliche Aufgaben bekommen (z.B. Seminarvortrag).

Prüfung (1)

- Klausur: 20.03.2024, 14^{00}–16^{00}.

  Der Termin kann sich noch ändern. Bitte achten Sie auf Ankündigungen auf der Webseite. Informieren Sie den Dozenten möglichst bald über eventuelle Terminkonflikte. Geplant ist eine elektronische Prüfung im Prüfungscenter der Universität (Mansfelder Str. 15, 06108 Halle, „Lührmann Gebäude“, Straßenbahn-Haltestelle „Saline“).

- Zweiter Termin: mündl. Prüfung, ca. 09.–12.07.2024.

- Bei der Klausur sind drei DIN A4-Blätter mit Notizen erlaubt (Vorder- und Rückseite bedruckt/beschrieben), keine Bücher, Aktenordner, Rechner.

  Praktische Anwendung, Verstehen, wenig Auswendiglernen.
  Z.B. XML DTD und Datendatei für gegebene relationale DB,
  XML Schema (ggf. nur Teilstück), Anfragen in XPath und XQuery.
  Eventuell XDM Datenstruktur zeichnen. Kurzer Aufsatz zu Frage.

Prüfung (2)

- Ich wundere mich über Studierende, die in der Klausur Konstrukte nutzen, die sie vorher in den Hausaufgaben offenbar nie ausprobiert haben.

  Sonst wüssten sie, dass es nicht geht. Z.B. das Schlüsselwort `AND` in einem XPath-Ausdruck groß schreiben. XML ist case-sensitiv, die Anfragesprachen sind es auch, und alle Schlüsselworte werden klein geschrieben.

- Klausuraufgaben sind ähnlich zu Hausaufgaben vorher.

  Nutzen Sie die Gelegenheit und beschäftigen Sie sich ernsthaft mit den Hausaufgaben. Da können Sie ja auch herumprobieren.

- Es ist damit zu rechnen, dass Sie bei der Klausur Ihre Anfragen u.s.w. nicht ausprobieren können.

  Eventuell entwickeln wir noch eine eigene Webseite zum Ausprobieren. Ansonsten müssen Sie sozusagen „auf Papier programmieren“, obwohl es eine elektronische Klausur ist. Wenn Sie bei den Hausaufgaben viel Zeit zum Debuggen brauchen, sind Sie noch nicht reif für die Klausur.

Inhalt

Lehrbücher (1)

- Erhard Rahm, Gottfried Vossen (Hrsg.):
  Web & Datenbanken.

  Konzepte, Architekturen, Anwendungen.

  dpunkt.verlag, 2003, ISBN 3-89864-189-9, 488 Seiten.

- Meike Klettke, Holger Meyer:
  XML & Datenbanken.

  Konzepte, Sprachen, Systeme.

  dpunkt.Verlag, 2003, ISBN 3-89864-148-1, 428 Seiten.

- Georg Lausen:
  Datenbanken. Grundlagen und XML-Technologien.

  Spektrum Akademischer Verlag, 2005, ISBN 3827414881, 281 Seiten.

Lehrbücher (2)

- Harald Schöning:
  XML und Datenbanken. Konzepte und Systeme.

  Hanser Fachbuchverlag, 2002, ISBN 3446220089, 300 Seiten.

- Wassilios Kazakos, Andreas Schmidt, Peter Tomczyk:
  Datenbanken und XML.

  Konzepte, Anwendungen, Systeme.

  Springer, 2002, ISBN 354041956X, 352 Seiten.

- Akmal B. Chaudhri, Awais Rashid, Roberto Zicari:
  XML Data Management.

  Native XML and XML-Enabled Database Systems.

  Addison-Wesley, 2003, ISBN 0201844524, 688 Seiten.

Lehrbücher (3)

- Margit Becher:
  XML: DTD, XML-Schema, XPath, XQuery, XSL-FO, SAX, DOM.

  Springer Vieweg, 2. Aufl., 2022, 448 Seiten ISBN-10: 3658354348.

- Priscilla Walmsley:
  Definitive XML Schema, 2nd Ed.

  Prentice Hall, 2012, ISBN 0-13-288672-3, 768 Seiten.

- Eric van der Vlist:
  XML Schema.

  O'Reilly, 2002, ISBN 0596002521, 400 Seiten.

Lehrbücher (4)

- Wolfgang Lehner, Harald Schöning:
  XQuery: Grundlagen und fortgeschrittene Methoden.

  dpunkt.verlag, 2004, ISBN 3898642666, 304 Seiten.

- Howard Katz (Editor):
  XQuery from the Experts.

  A Guide to the W3C XML Query Language.

  Addison-Wesley, 2003, ISBN 0321180607, 512 Seiten.

- Priscilla Walmsley:
  XQuery: Search Across a Variety of XML Data.

  O'Reilly Media, 2nd Ed., 2016, 733 Seiten, ISBN 978-1491915103.

- Rudolf Jansen:
  XQuery. Eine praxisorierentierte Einführung.

  Software & Support Verlag, 2004, ISBN 3-935042-65-5, 167 Seiten.

Lehrbücher (5)

- Jim Melton, Stephen Buxton:
  Querying XML.

  Morgan Kaufmann, 2006, ISBN 1-55860-711-0, 848 Seiten.

- Michael Brundage:
  XQuery: The XML Query Language.

  Addison-Wesley Professional, 2004, 536 Seiten, ISBN-10 :0321165810.

- Michael Seemann:
  Native XML Datenbanken im Praxiseinsatz.

  Software & Support Verlag, 2003, ISBN 3-935042-35-3, 316 Seiten, mit CD.

- Bastian Gorke:
  XML-Datenbanken in der Praxis.

  bomots verlag, 2006, ISBN 3-939316-19-9, 130 Seiten.

Lehrbücher (6)

- Michael Kay:
  XSLT 2.0 and XPath 2.0 Programmer's Reference.

  Wrox, 4th Ed., 2008, ISBN 0470192747, 1368 Seiten.

- Bob DuCharme:
  XML: The Annotated Specification.

  Prentice-Hall, 1998, ISBN 0-13-082676-6, 339 Seiten.

- Elliotte Rusty Harold, W. Scott Means:
  XML in a Nutshell, A Desktop Quick Ref., 3rd Ed.

  O'Reilly, Okt. 2004, ISBN 0-596-00764-7, 689 Seiten.

Eine Bitte

- Das Gebiet ist auch für mich noch ziemlich neu.

  Es entwickelt sich ja auch recht schnell.

- Wahrscheinlich weiß mancher von Ihnen zumindest über manches Detail mehr als ich.

  Das ist mir nicht peinlich, ich lerne gerne.

- Teilen Sie Ihr Wissen mit uns allen!

  Korrigieren Sie Fehler, falls Sie sie bemerken.

- Stellen Sie Fragen!

  Bleiben Sie nicht einfach nur passiver Zuhörer!

- Verbesserungsvorschläge sind willkommen.