

XML und Datenbanken

Kapitel 0: Organisatorisches

Prof. Dr. Stefan Brass

Martin-Luther-Universität Halle-Wittenberg

Wintersemester 2022/23

<http://www.informatik.uni-halle.de/~brass/xml22/>

Inhalt

- 1 Inhalte
- 2 Motivation
- 3 Zeit und Ort
- 4 Kontaktdaten
- 5 Prüfung
- 6 Literatur

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?

Beschränkung der Teilnehmerzahl

- 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).

Studierende, die sich wirklich für XML Technologien interessieren, und die Voraussetzungen mitbringen, wären ja willkommen. So ist es aber ein Problem, das gelöst werden muss. Solange aber alles zu funktionieren scheint . . .

- Die geplanten Seminarvorträge gehen bei 35 Teilnehmern schon nicht mehr. Meine Zeit ist begrenzt.

Inhalt

- 1 Inhalte
- 2 Motivation**
- 3 Zeit und Ort
- 4 Kontaktdaten
- 5 Prüfung
- 6 Literatur

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 → 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.

Inhalt

- 1 Inhalte
- 2 Motivation
- 3 Zeit und Ort**
- 4 Kontaktdaten
- 5 Prüfung
- 6 Literatur

Zeit und Ort

Vorlesung/Seminar (2 SWS):

- Dienstags, 12¹⁵–13⁴⁵, 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:

- Montags, 16¹⁵–17⁴⁵, online.
- StudIP-Eintrag der Übung, Reiter „Meetings“.

Melden Sie sich bitte mit Mikrofon an (stumm geschaltet).

- 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

- 1 Inhalte
- 2 Motivation
- 3 Zeit und Ort
- 4 Kontaktdaten**
- 5 Prüfung
- 6 Literatur

Ansprechpartner (1)

Dozent: Prof. Dr. Stefan Brass

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

Betreff-Zeile sollte mit [xm122] beginnen, möglichst aussagefähig.

- Büro: Von-Seckendorff-Platz 1, Raum 313
- Telefon: 0345/55-24740
- Sprechstunde: Montags, 12⁰⁰–13⁰⁰
- Frühere Unis: Braunschweig, Dortmund, Hannover, Hildesheim, Pittsburgh, Gießen, Clausthal.
- Oracle8 Certified Database Administrator.
- IBM Certified Advanced DBA (DB2 UDB 8.1).

Ansprechpartner (2)

Weitere Mitarbeiter der Gruppe:

- PD Dr. Alexander Hinneburg

Raum 314, Telefon: 0345/55-24732, Email: hinneburg@...

- M. Sc. Mario Wenzel

Raum 315, Tel. 0345/55-24776, Email: mario.wenzel@...

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/xml22/>

- 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

- 1 Inhalte
- 2 Motivation
- 3 Zeit und Ort
- 4 Kontaktdaten
- 5 Prüfung**
- 6 Literatur

Studienleistung (1)

- Studienleistung: 67% der Hausaufgabenpunkte
 - Die Hausaufgaben sind einzeln zu bearbeiten.
 - Die Hausaufgaben werden nur stichprobenartig korrigiert, aber in der Übung besprochen.

Wenn Ihre Hausaufgabe nicht korrigiert wurde, bekommen Sie vorläufig die volle Punktzahl. Es ist möglich, dass Hausaufgaben noch nachträglich korrigiert werden, und die Punktzahl dann gesenkt wird. Damit müssen Sie insbesondere dann rechnen, wenn Sie wegen eines Plagiats auffallen, oder wegen der Abgabe einer Lösung, die offensichtlich kein sinnvoller Lösungsversuch ist.
 - Wenn Sie sich in der Übung freiwillig melden, vorzurechnen, werde ich nur in extremen Fällen Punkte abziehen.

Sollte ich Ihre Hausaufgabe zufällig später für die Stichprobe auswählen und dann doch Punkte abziehen, können Sie mich daran erinnern, dass Sie freiwillig vorgerechnet haben.

Studienleistung (2)

- Falls Plagiate bei Hausaufgaben oder sehr still bei Übungen: Änderung (Verschärfung) der Regeln möglich.

Modulbeschreibung: „Aktive Mitarbeit in den Übungen inkl. Kurzvorträge über Hausaufgaben und Beantwortung von Fragen zum Umfeld der Aufgaben“.

- Helfen Sie Ihren Mitstudierenden, aber geben Sie nicht Ihre Lösung zum Abschreiben.

Minimal 0 Punkte für alle Beteiligten bei „zu ähnlichen“ Lösungen.
Im Wiederholungsfall keine Studienleistung in diesem Jahr!

- 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%).

Studienleistung (3)

- Falls Sie häufig nicht an der Übung teilnehmen können:
 - Reden Sie mit 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).
- 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.

Prüfung

- Klausur: 21.02.2023, 10⁰⁰–12⁰⁰ (Vorschlag).

Der Termin kann sich eventuell noch ändern. Bitte achten Sie auf Ankündigungen auf der Webseite. Informieren Sie den Dozenten möglichst bald über eventuelle Terminkonflikte.

- Zweiter Termin: mündl. Prüfung, ca. 28.03.2023.
- 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.

Inhalt

- 1 Inhalte
- 2 Motivation
- 3 Zeit und Ort
- 4 Kontaktdaten
- 5 Prüfung
- 6 Literatur**

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)

- 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 praxisorientierte 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 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!

- Nehmen Sie an den Übungen in der Vorlesung teil.

Bleiben Sie nicht einfach nur passiver Zuhörer!