XML und Datenbanken

(Winter 2010/2011)

Prof. Dr. Stefan Brass
Institut für Informatik

Übungsleiter: Dr. Henning Thielemann

Geplante Themen

- XML Syntax (Kurzeinführung / Wiederholung)
- XML Schema
- XPath / XQuery Data Model (XDM), XML Infoset
- XPath
- XQuery
- XSLT (kurz)
- XML-Unterstützung in SQL (Oracle, DB2)
- Speicherstrukturen für XML (falls noch Zeit)

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 / 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 wichtigen Daten in einer relationalen DB hält, wird man in der Kommunikation mit Geschäftspartnern etc. XML benutzen. Es ist dann nützlich, wenn man gewisse Auswertungen direkt auf den XML-Dateien durchführen kann.

Motivation (2)

- 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.

Zeit und Ort

Vorlesung (2 SWS):

- Dienstags, 16^{15}–17^{45}, Raum 1.23.

Übung:

- Dienstags, 10^{15}–11^{45}, Raum 5.06, ab 19.10.2010
- Maximal 20 Teilnehmer

Klausur:

- (voraussichtlich 22.03.2011, 16–18, Raum 509)

  Achten Sie zur Sicherheit auf eventuelle weitere Ankündigungen.
  Bei zu wenig Teilnehmern mündliche Prüfungen.

Prüfungsmodalitäten (1)

Bedeutung der Klausur:

- Für Wirtschaftsinformatiker (Diplom): Prüfung zu einer Veranstaltung aus dem Datenbank-Bereich.

  Diese Veranstaltung ist neu. Erkundigen Sie sich zur Sicherheit in Ihrem Prüfungsamt.

- Für Master-Studierende: Abschlußprüfung des Moduls (Wirtschaftsinformatik nur nach Beratung).
- Für alle übrigen: Nur Leistungsschein.

  Dieser Schein wird, soweit ich weiß, nirgendwo verlangt.

Prüfungsmodalitäten (2)

Ablauf der Klausur:

- Schriftliche Unterlagen (Bücher, Skript, eigene Notizen, etc.) sind erlaubt. Rechner sind nicht erlaubt.
- Mit 60% der Punkte ist die Zensur mindestens 4.0.

  D.h. 50 FP für Wirtschaftsinformatiker. Es ist also nicht garantiert, daß die Fachpunkte genau dem erreichten Prozentsatz der Klausurpunkte entsprechen.

- Mit 95% der Punkte gibt es eine 1.0 (95 FP).
- Eventuell gibt es bis max. 10% der Klausurpunkte für Hinweise auf Fehler im Skript etc. (typisch nur 1–3%, kein Anspruch).

Prüfungsmodalitäten (3)

Zulassung zur Klausur:

- Zur Klausur wird nur zugelassen,
  - ◇ wer mindestens 50% der Hausaufgabenpunkte erreicht hat,
  - ◇ und in der Übung seine Abgaben (bzw. die seiner Gruppe) präsentieren und erklären konnte.

    Einschließlich Fragen zum Umfeld der Aufgaben beantworten. Maximale Gruppengröße: 4. Empfohlen: 1.

- Dazu muß man in der Übung hinreichend häufig anwesend sein (bei mehr als dreimaliger Abwesenheit Ermessen des Übungsleiters).

Ansprechpartner (1)

Dozent: Prof. Dr. Stefan Brass

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

  Betreff-Zeile sollte mit `[xml10]` beginnen, möglichst aussagefähig.

- Büro: Von-Seckendorff-Platz 1, Raum 313
- Telefon: 0345/55-24740
- Sprechstunde: Dienstags, 12^{15}–13^{15}
- Frühere Unis: Braunschweig, Dortmund, Hannover, Hildesheim, Pittsburgh, Gießen, Clausthal.
- Oracle8 Certified Database Administrator.
- IBM Certified Advanced DBA (DB2 UDB 8.1).

Ansprechpartner (2)

Übungsleiter: Dr. Henning Thielemann

- Büro: Von-Seckendorff-Platz 1, Raum 314.
- Telefon: 0345/55-24773.
- Email: `xml10@henning-thielemann.de`

Sekretärin: Ramona Vahrenhold

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

WWW-Seite

`http://www.informatik.uni-halle.de/~brass/xml10/`

- Aktuelle Ankündigungen
- Folien der Vorlesung (PDF oder ps 4:1)
- 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.

- Punkte-Datenbank

  Tragen Sie sich bitte bei der Punkte-Datenbank ein, sofern Sie keine Datenschutz-Bedenken haben. Wenn Sie Ihre Ergebnisse nicht in der Punkte-Datenbank wünschen, teilen Sie uns das bitte mit.

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)

- Eric van der Vlist:
  XML Schema.
  O'Reilly, 2002, ISBN 0596002521, 400 Seiten.
- Rudolf Jansen:
  XQuery, Eine praxisorierentierte Einführung.
  Software & Support Verlag, 2004, ISBN 3-935042-65-5, 167 Seiten.
- Howard Katz (Editor):
  XQuery from the Experts.
  A Guide to the W3C XML Query Language.
  Addison-Wesley, 2003, ISBN 0321180607, 512 Seiten.

Lehrbücher (4)

- 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.
- 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, 37 Euro.

Eine Bitte

- Das Gebiet ist für mich neu.
- 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.