

Zu meiner Person

- Background eigentlich in Soziologie/Philosophie (Uni Frankfurt)
- Erste Kontakte mit SGML/DSSSL in 1996
- 1999-2005: Mitarbeiter an der Technischen Universität Darmstadt, Schwerpunkt: wiss. Informationssysteme, digitale Bibliotheken, Knowledge Management
- 2005/06: Heidelberger Akademie der Wissenschaften
- Seit 2004: Unabhängiger Entwickler; diverse Projekte rund um eXist

Die Anfänge

- Sommer 2000/Frühjahr 2001: Erste Versuche zur Implementierung von Indexierungsverfahren für XML-Dokumente

Inspiration:

D. Shin, H.Jang, H.Jin: "Bus: An Effective Index and Retrieval Scheme in Structured Documents". In Proceedings of the 3rd ACM International Conference on Digital Libraries, 1998, Pittsburgh, PA.

Erste Schritte

- Januar 2001: erste Versionen basieren noch auf relationaler DB (Oracle/MySQL), die Zerlegung der Dokumente ist aber wesentlich zu langsam
- Mai 2001: Natives XML Backend mit Berkeley DB (schwierig zu installieren)
- 2002: Erste reine Java-Version
- 2003: Beginn der XQuery-Implementierung
- 2006: Neues Indexierungsschema (DLN), Redesign des Datenbankkerns

Motivation

- Entwicklung wurde vorrangig durch konkrete Anwendungen bestimmt
- Features wie XQuery, XUpdate oder Crash Recovery entstanden aus der Zusammenarbeit mit Anwendungsprojekten

eXist war von Anfang an für Anwendungsfälle gedacht, die sich nicht oder nur umständlich mit einem RDBMS umsetzen lassen, sollte aber nicht eine relationale DB ersetzen!

Wer benutzt eXist?

Die Gruppe der eXist-Nutzer ist sehr heterogen!

Idealtypisch gibt es zwei große Fraktionen mit z.T. gegensätzlichen Interessen:

- Klassische Datenbank-Anwender: kommen von der relationalen DB zu eXist
- SGML/XML Anwender: kennen XML/XSLT/XQuery, haben aber keine Erfahrung in Datenbank-Programmierung

Wer benutzt eXist?

- eXist kam eher aus der SGML/XML Ecke:
- Ausrichtung auf dokumenten-zentriertes XML, komplexes Markup mit sehr unregelmäßiger Struktur
- Viele eXist-Nutzer sind in den Geisteswissenschaften beheimatet (TEI-Standard)
- Daten-zentrierte Applikationen wurden im Laufe der Zeit aber immer wichtiger

Vorteil von XML DBs liegt gerade in der Kombination und Verlinkung von dokumenten-zentrierten mit daten-zentrierten Informationen!

Wofür wird eXist benutzt?

- Elektronische Editionen, textkritische Analyse, Linguistik:
<http://www.zeit.de/akademie>, "Der Junge Goethe in seiner Zeit"
- Archivierung, Museumssysteme:
<http://museicapitolini.net/>, <http://www.perseus.org/>
- Technische Dokumentation:
FIAT Wartungsdokumentation, ASML,
<http://erules.veristar.com/dy>

Wofür wird eXist benutzt?

- Portale, Content Management Systeme:
z.B.: <http://erules.veristar.com/dy>
- Autorensysteme, Publishing
- Produktion, Wartung, Überwachung:
ASML, Trapeze Networks
- Online-Shops:
<http://rosavtotorg.ru/>
- Biologie/Umwelt/Landwirtschaft
Biodiversität im Amazonas, FAO Terminology
<http://www.fao.org/faoterm/search>

Derzeitige Projekte

- Elektronische Edition der Werke des Jungen Goethe (Ludwigs-Maximilians-Universität München)
- Network Monitoring Application (Trapeze Networks, USA)
- Buddhistische Steininschriften in China (Heidelberger Akademie der Wissenschaften)
- Text Analysis Portal: TAPoR (University of Victoria, Canada)
- Wartung/Steuerung von Litographie-Maschinen zur Chipproduktion (ASML, Niederlande)

Das eXist-Team



- Kernteam besteht im Moment aus 4 bis 5 Personen, quer über Europa verteilt
- Größerer Kreis umfasst vielleicht 20 bis 30 Entwickler, die unregelmäßig Beiträge leisten

Das eXist-Team

- Vorteil: das Projekt bleibt überschaubar, die Kommunikation ist einfach und direkt
- Nachteil: das Tagesgeschäft (Bug-Fixes, Support, Qualitätsmanagement) bleibt an wenigen Personen hängen

Rückblick 2006

- Neues Indexierungsschema im Kern von eXist
- Versionen 1.0 und 1.1 wurden zur gleichen Zeit veröffentlicht
- Version 1.1 beseitigt die bisherige **Größenbeschränkung für Dokumente**
- Bessere Performance für Updates einzelner Knoten

Rückblick 2006

Standardkonformität

eXist besteht derzeit 92.8% der insgesamt 14.637 Tests in der offiziellen XQuery Test Suite des W3C

- Im Jan. 2006 fiel eXist noch bei 60% der Tests durch!
- Durch die Test Suite wurden viele Unsicherheiten beseitigt
- XQTS bedeutet mehr Sicherheit für die Benutzer

Schwerpunkte 2007: Performance

- Verbesserter Query-Optimizer
- intelligentes Umschreiben der Query bei Abfragen auf sehr große Knotenmengen
- Reduktion des Hauptspeicherbedarfs
- Optimierungsentscheidung soll früher getroffen werden
- Optimierung sollte für den User besser nachvollziehbar sein

Viele der heutigen Performance-Probleme sind nur durch eine intelligentere Query-Optimierung zu lösen

Schwerpunkte 2007: Redesign der Index-Architektur

- Modularisierung: Auslagerung der sekundären Indexstrukturen aus dem Kern
- Verwendung von Index-Informationen für die Optimierung
- Pipeline-Architektur zum Einfügen neuer User-definierter Indexstrukturen
- neue Indexe: N-Gram, Spatial ...
- bessere Konfigurierbarkeit bestehender Indexe

Weitere Pläne für 2007

- 1 Einheitliches Modell für temporäre (in-memory) und persistente Dokumente!
- 2 XQJ XQuery API for Java
- 3 Zu hoher Speicherbedarf für die Verwaltungsinformationen von Collections
- 4 Redesign der internen Locking Protokolle
- 5 XQuery Debugger
- 6 Anpassung an aktuelle W3C Drafts: XQuery Update Extensions, Fulltext Extensions