Technische Universität Kaiserslautern
Fachbereich Informatik
Lehrgebiet Datenverwaltungssysteme

Verarbeitung von XML-Strömen

Katharina Bellon

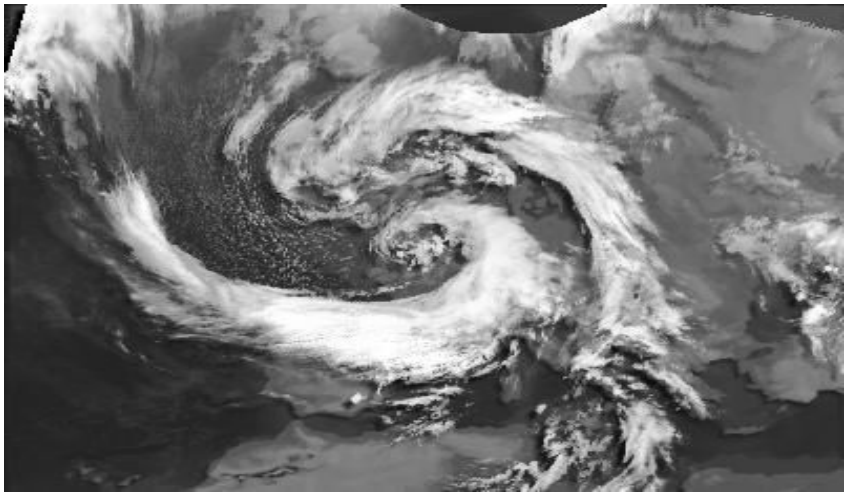
Seminar zum Thema Data Streams
Sommersemester 2005

Gliederung

- Anwendungen
- Anforderungen
- Anfragesprachen
  - XPath
  - XQuery
- Systeme zur Bearbeitung der XML-Ströme
  - SPEX
  - XMLTK (XPath-Prozessor)
  - XFilter
- Zusammenfassung

Anwendungen (1)

- Nachrichtenüberwachung
  - Börse
  - Presse
  - Meteorologie



APS LE COUP DE PROJECTEUR

Ingénieur en cognitique de l'IdC Bordeaux

UNE FORMATION UNIQUE EN FRANCE

L'institut de la cognitique (IdC) de l'université de Bordeaux 2 forme en trois ans des ingénieurs, spécialisés en sciences, techniques et technologies de la cognition, et dispense des formations de deuxième et troisième cycle. Il délivre le diplôme d'ingénieur en cognitique, reconnu par la commission des titres, et les Master, DEA et doctorats. L'originalité de cette formation réside principalement dans la réunion de domaines scientifiques tels que la psychologie et la physiologie, d'une part, et les technologies électronique et informatique, d'autre part.

Cette formation, toute nouvelle et unique en France, intéresse les domaines tels que : le management des risques et la sûreté de fonctionnement, avec la prise en compte du facteur humain, sous tous ses aspects ; le développement de produits, notamment ceux intégrant des automates, pour lesquels l'interface homme-machine (IHM) joue un rôle prépondérant ; l'ingénierie de systèmes, notamment d'imagerie à des fins de recherche médicale non invasive, pour la connaissance des mécanismes de la cognition, ou la détection de pathologies ou de troubles ; le conseil en organisation, avec la prise en compte de l'aspect social pour l'optimisation du changement organisationnel. D'autres domaines devraient être concernés : on peut penser aux méthodes d'ingénierie simultanée ; à la mise en place de plateaux virtuels, qui outre la compétence informatique réseaux, nécessitent de prendre en compte le comportement humain ; on peut également penser plus généralement à tout ce qui a trait à l'intelligence artificielle, comme par exemple, à la définition et au développement de méthodes et outils pour automatiser partiellement la conception des produits, en particulier les tâches répétitives sans grosse valeur ajoutée.

Des industriels majeurs ont déjà manifesté leur plus vif intérêt sur le sujet, notamment Airbus, Bertin Technologies, Dassault Aviation, EADS ST, EADS Sogerma, Lego, Renault, Thales...

Des élèves de deuxième et troisième cycles suivent déjà leur formation auprès de l'IdC. Les premiers élèves-ingénieurs (une trentaine) vont faire leur entrée en septembre 2004 ; le recrutement se fait auprès des élèves des classes préparatoires aux grandes écoles, auprès des titulaires - ou futurs titulaires - de diplômes du premier cycle (DEUG sciences, DUT...).

Les débouchés professionnels sont évidents pour ces diplômés, qui n'ont pas de concurrence en France.

Les sciences cognitives s'imposent dans l'activité industrielle

Les sciences cognitives constituent un domaine scientifique en pleine évolution déjà consacré sur le plan mondial, surtout en Amérique du Nord. Elles sont enseignées en France à travers certains cycles de formation (UFR). Elles débouchent aujourd'hui sur une dimension applicative susceptible de prendre une place parmi les sciences de l'ingénieur. Elles concernent notamment la question des interfaces homme/machine, centrales dans la production de biens manufacturés (aéronautique, automobile, équipements...), ainsi que l'activité recherche et développement des entreprises industrielles ; elles trouvent des applications notables dans la gestion des systèmes complexes - le système de visualisation et de pilotage dans l'aéronautique, par exemple - et dans la mise en œuvre de travail collaboratif interentreprises à partir de l'usage des nouvelles techniques d'information et de communication.

Cognitique et aéronautique

Aujourd'hui, les entreprises aéronautiques françaises sont confrontées à une concurrence mondiale très forte, en particulier d'origine américaine. L'évolution de l'offre technologique (génération graphique, optique, holographie, interfaces vocales, automatisation...), des contraintes de gestion de la complexité (gestion de programmes, gestion des situations en environnement complexe...), ainsi que des systèmes de missions civiles et militaires (conduite de vol, missions complexes intégrant des drones...) conduit désormais à faire de l'interface homme/machine un enjeu stratégique majeur.

La situation concurrentielle avec les groupes américains est telle que le marché est ouvert, y compris pour la sous-traitance des groupes aéronautiques français : par exemple, Dassault sous-traite certains de ses équipements à Honeywell, groupe concurrent de Thales.

Pour mémoire, l'activité dédiée à l'interface homme/machine peut représenter plus de 60% des effectifs d'un établissement intervenant dans ce secteur.

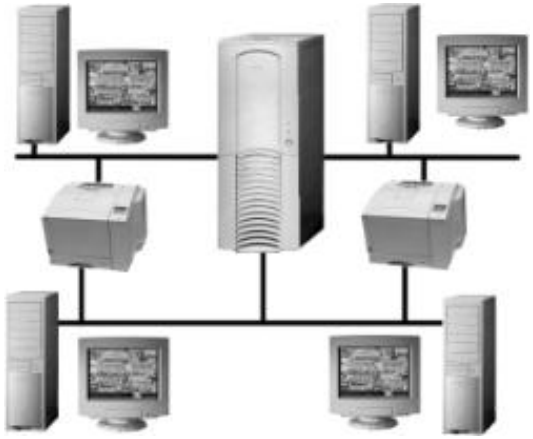
Les nouveaux équipements (casques de visée, hologrammes, systèmes avioniques de conduite de vol sur écrans, systèmes de conduite vocale, visualisations graphiques...) sont autant d'éléments centraux des grands programmes de conception et fabrication aéronautique, qui revêtent un caractère stratégique pour l'industrie française et européenne, et pour lesquels l'approche cognitique est non seulement nécessaire mais majeure.

Bernard Ladevèze

Vendredi 26 mars 2004 - Page 4

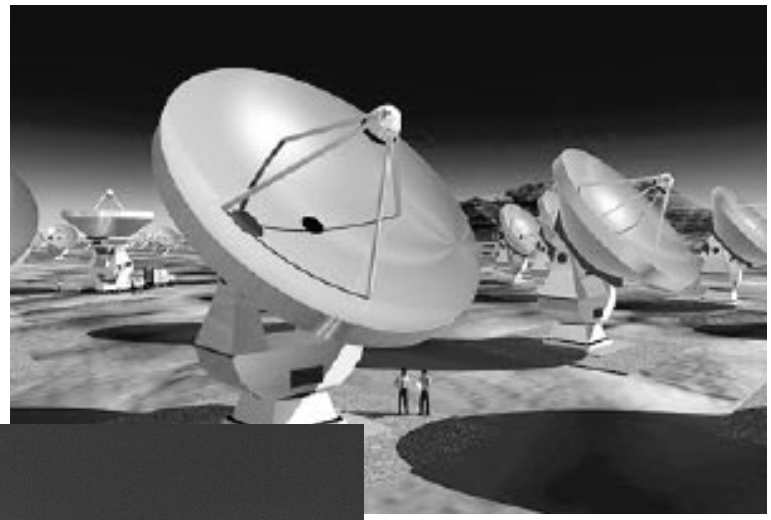
Anwendungen (2)

- Systemüberwachung und Systemsteuerung
  - Verkehrsteuerung
  - Produktionssteuerung
  - Logistik
  - Netzwerkverwaltung



Anwendungen (3)

- Analyse von wissenschaftlichen Messdaten
  - in Medizin
  - in Astronomie
  - bei der Früherkennung von Tornados



Anforderungen

- Möglichst schnelle Bearbeitung von Anfragen, zum Teil sogar in Echtzeit
- Platzsparende Verfahren
- Möglichst genaue Approximationsmechanismen
- Bearbeitung beliebig strukturierter XML-Daten
- Weiter Anforderungen:
  - Skalierbarkeit
  - Plattformunabhängigkeit
  - ...

XPath

- XPath 1.0
  - Adressiert Knoten eines XML-Baumes
  - Beispiel: `/child::name[position()=3]`
  - Lokalisierungspfad
  - Kontextknoten
  - Achsen
  - Knotentest
  - Prädikat
  - Ergebnisobjekt vom Typ: *node-set, boolean, string, number*
- SXP (Simple XPath) wie XPath mit Ausnahme von
  - Achsen *ancestor-or-self* und *descendant-or-self*
  - Wertebasierte Vergleiche
    - der Ergebnisse von Prädikaten (Bsp.: `[child::a = desc::a]`)
    - von positionsbasierte Prädikaten (Bsp.: `[position() = 3]`)

XQuery

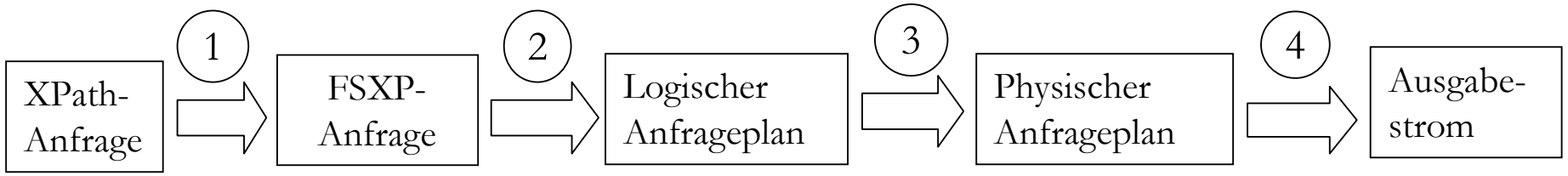
- XML Query Language basiert auf XPath 2.0
- Wurde von *Quilt*, *XML-QL* und *Lorel* abgeleitet
- Ähnliche Semantik wie SQL (*Structured Query Language*) und OQL (*Object Query Language*)
- Wird zur Transformation von XML-Dokumenten benutzt
- XQuery-Anfrage kann geschachtelte Unteranfragen enthalten
- Ausdrücke können
  - Funktionen und Operatoren enthalten
  - bedingt und quantifiziert sein
  - boolesche und Vergleichsausdrücke sein
- Möglich sind Joins und Aggregatfunktionen ähnlich wie in SQL
- Zentrale Schlüsselwörter: **FLWR**

Beispiel:

```
for $b in doc("books.xml")//book
let $c := $b//author
where count($c) > 2
return $b/title
```

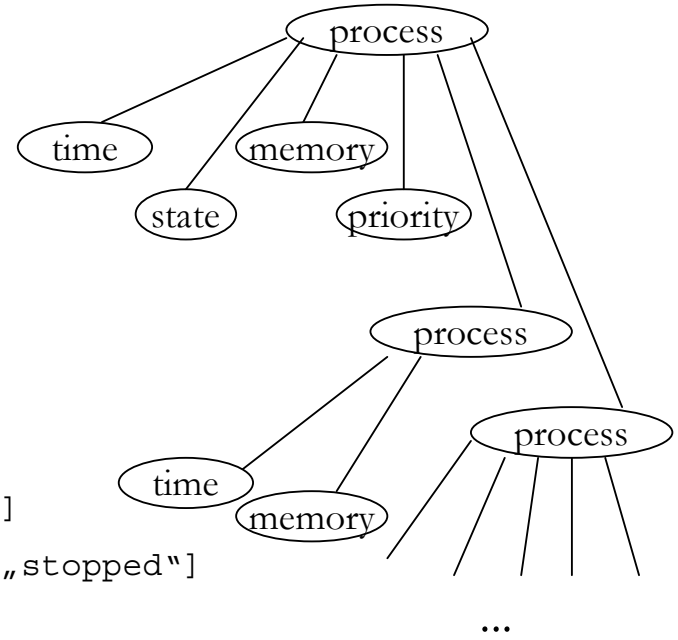
SPEX (1)

- *Streamed and Progressive Evaluator for XPath*
- Wurde entwickelt, um XPath-Anfragen gegen den XML-Datenstrom auszuwerten
- Netzwerk aus deterministischen Kellerautomaten
- Sequentieller Durchlauf
- Wertet FSXP (*Forward Simple XPath*) -Anfragen ohne Rückwärtsachsen (z. B. *parent, ancestor*) aus
- Die Verarbeitung erfolgt in vier Schritten



SPEX (2)

Schritt 1



Beispiel:

XPath-Anfrage

```
/desc::process[child::time > 24 or child::memory > 500]
/anc::process[child::priority < 10 and child::state = „stopped“]
```

FSXP-Anfrage

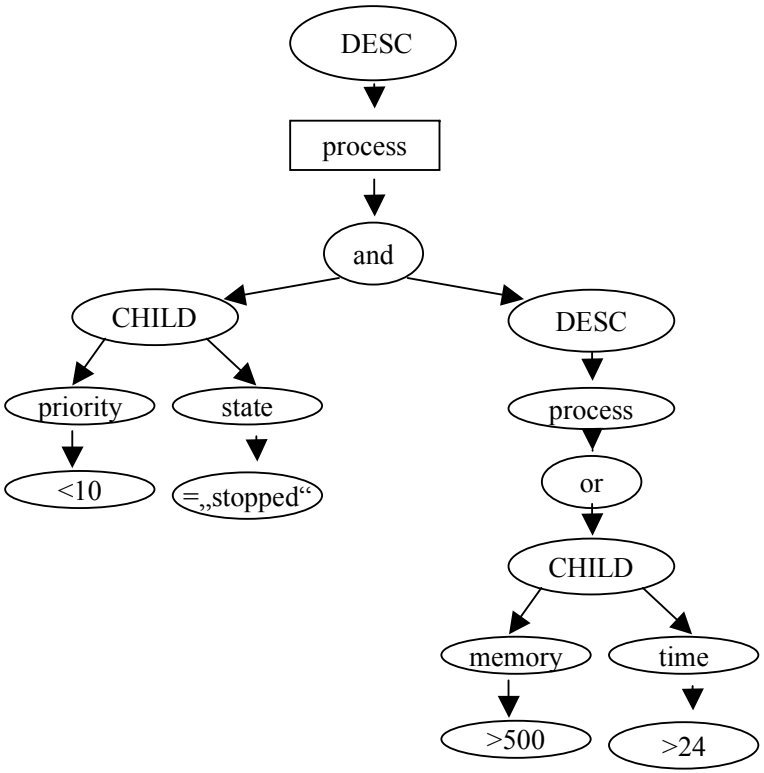
```
/desc::process[child::prority < 10 and child::state = „stopped“ and
 desc::process[child::time > 24 or child::memory > 500]]
```

SPEX (3)

Schritt 2

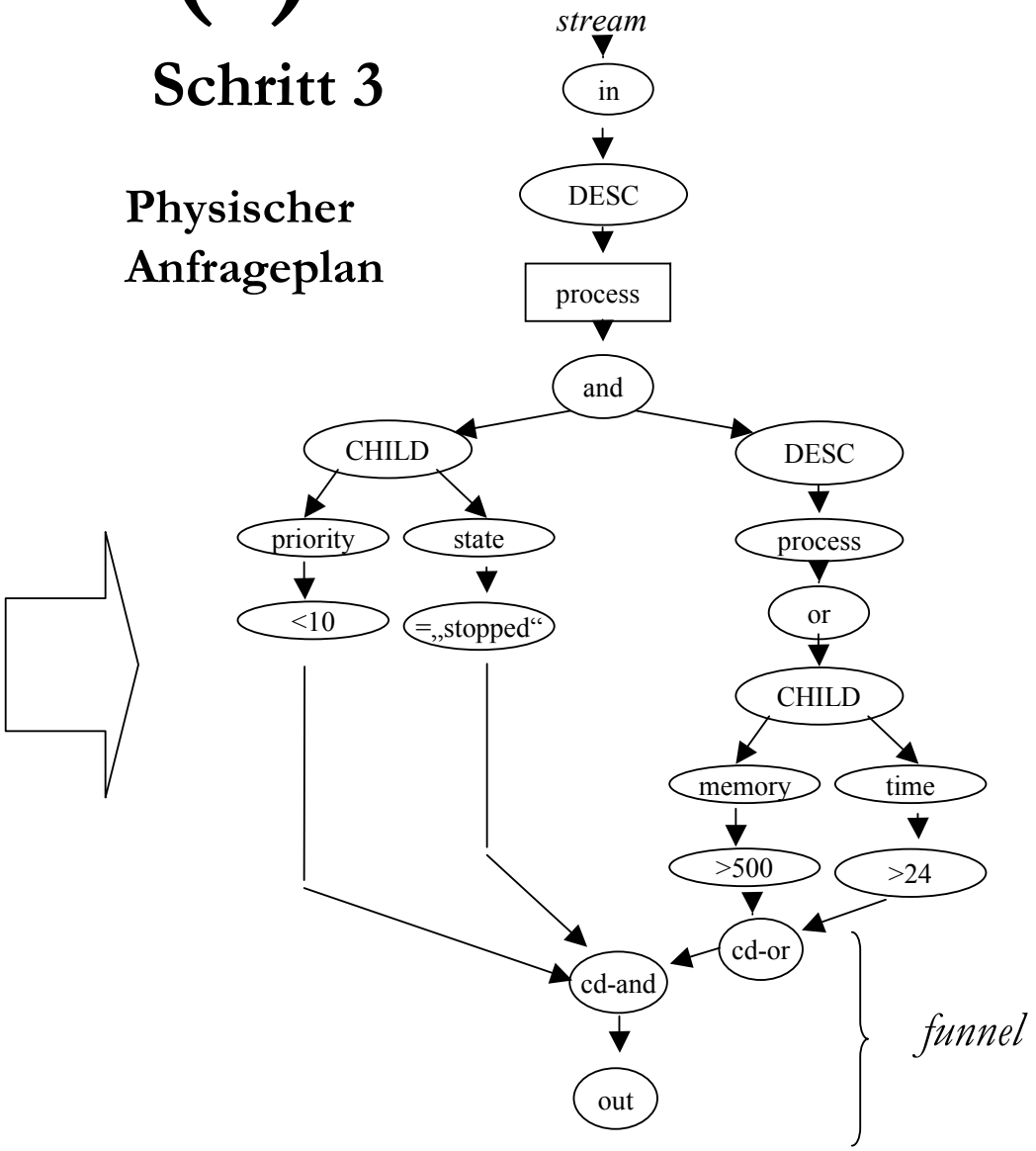
```
/desc::process[child::prority < 10
    and child::state = „stopped“
    and /desc::process[child::time > 24
          or child::memory > 500]]
```

Logischer Anfrageplan



Schritt 3

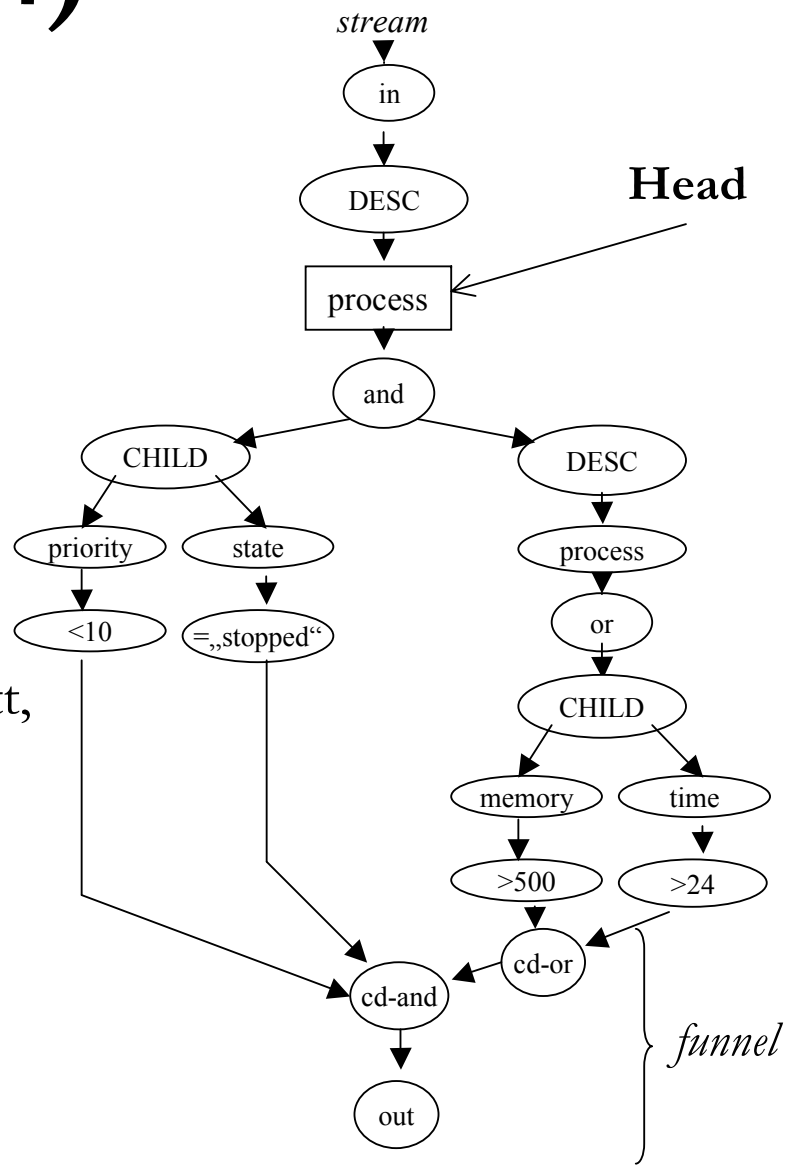
Physischer Anfrageplan



SPEX (4)

Schritt 4

- Strom besteht aus XML-Token
- *Transduktor*
  - benutzt Stack, um die Tiefe des Knoten zu merken.
  - leitet den Token unverändert oder annotiert weiter.
- *Head*: entspricht einem Lokalisierungsschritt, der zur gewünschten Ergebnismenge führt.
- *Funnel:* sammelt potentielle Antworten und fügt sie zusammen

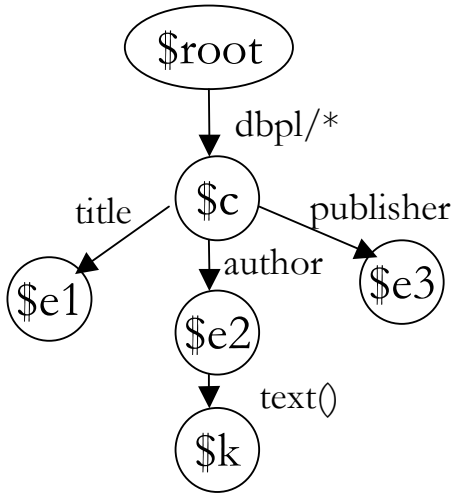


XMLTK (1)

- **XML Toolkit** ist ein System bestehend aus mehreren Kommandozeilen-Werkzeugen, die XML-Daten verarbeiten:
  - **`xsort:`** sortiert den XML-Strom
  - **`xdelete:`** löscht Element oder Attribut
  - **`xnest:`** gruppiert Elemente
- Beispiel:

```
xsort -c /dblp/* -e title -e author -k text() -e publisher
```

```
$c in $root/dblp/*
$e1 in $c/title
$e2 in $c/author
$k in $e2/text()
$e3 in $c/publisher
```



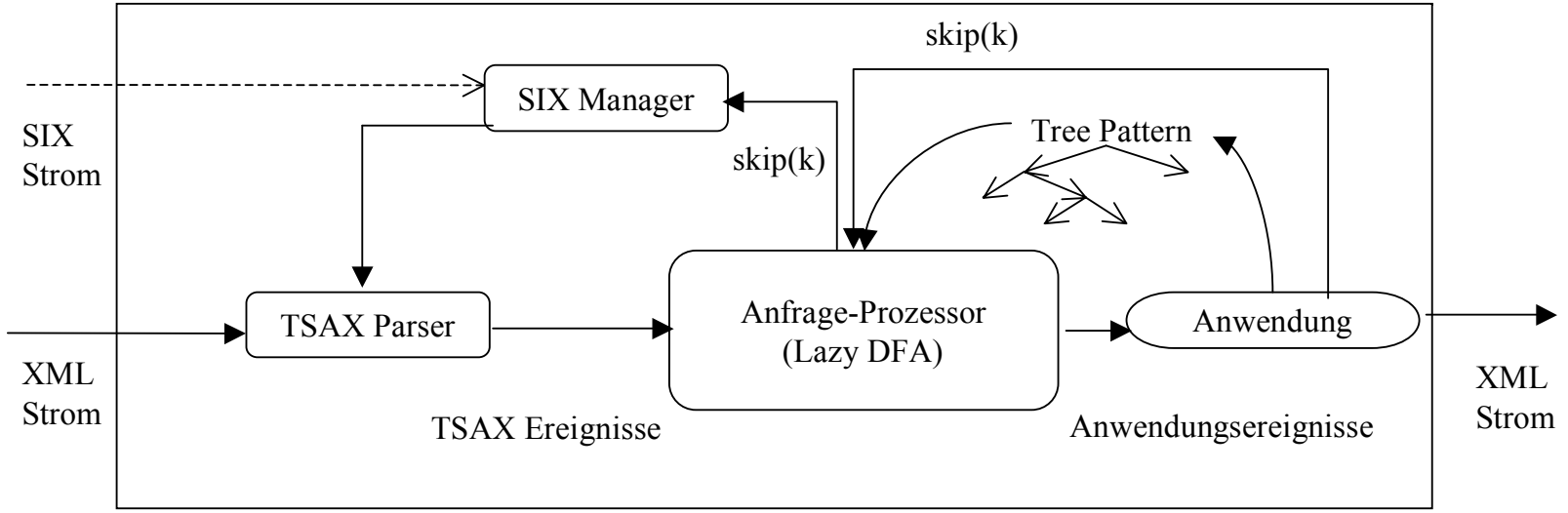
Tokenized SAX:

```
startVariable($root)
startDocument()
    startVariable($c)
    startElement(,book‘)
        startVariable($e2)
        startElement(,author‘)
            startVariable($k)
            characters(,Elliot‘)
            endVariable($k)
          ...
```

XMLTK (2)

XPath-Prozessor

- Bekommt als Eingabe den Anfragebaum und Strom von TSAX-Ereignissen
- Konvertiert den Anfragebaum in NEA, dann in DEA
- Benutzt Stack
- Vergleicht Elemente des XML-Baumes und der Anfrage
- Liefert Ereignisse an die Anwendung



XFilter (1)

- Filtermechanismus entwickelt für SDI-Systeme
- SDI (*Selective Dissemination of Information*):
  - speichert und vergleicht Benutzerprofile mit eingehenden XML-Dokumenten
  - filtert und verteilt relevante Information an Empfänger

Unterschied zu DBS: Speicherung und Verwaltung von Anfragen anstelle von Daten

Motivation:

Das XML-Dokument genügt den Ausdrücken q1 und q2 aber nicht q3

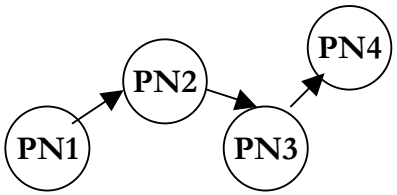
```
q1: /catalog/product//msrp
q2: //product/price[@currency =
    "USD"]/msrp
q3: //product[price/msrp<300]/name
```

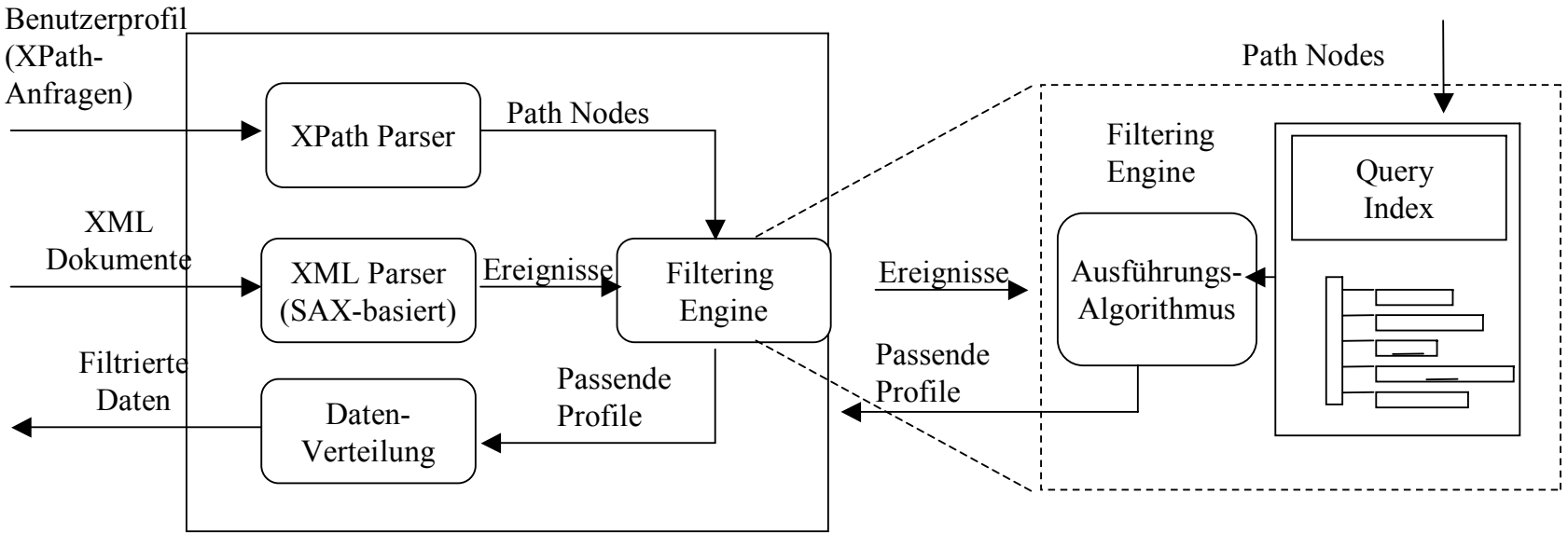
```
<?xml version="1.0"?>
<catalog>
 <product id="Kd-245">
  <name> Color Monitor </name>
  <price currency="USD">
   <msrp> 310.40 </msrp>
  </price>
 </product>
</catalog>
```

XFilter (2)

Beispiel: `/katalog//drucker/*/details[preis/euro<150]/name`

- *`Path Nodes`*: Elementknoten sind Zustände des endlichen Automanten
- *`Query Index`* enthält die Path Nodes
- *`Filtering Engine`*: Ereignisse steuern den Filterprozess





XFilter (3)

Query Index

Query ID: eindeutige Bezeichnung der Anfrage

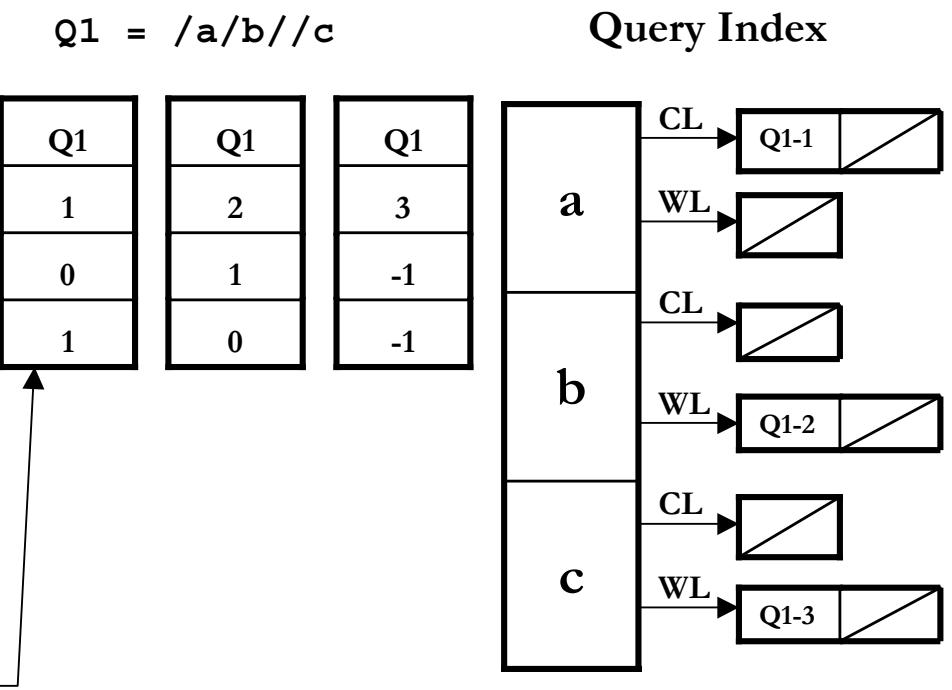
Position: Position eines Knoten in der Anfrage

RelativePos: beschreibt den Abstand zwischen dem betrachtetem Knoten und seinem Vorgänger

0	falls Knoten auf 1.Position steht
-1	falls vor dem Knoten //-Operator steht
2	falls vor dem Knoten *-Operator steht
1	sonst

Level: bezeichnet die Tiefe des Knoten

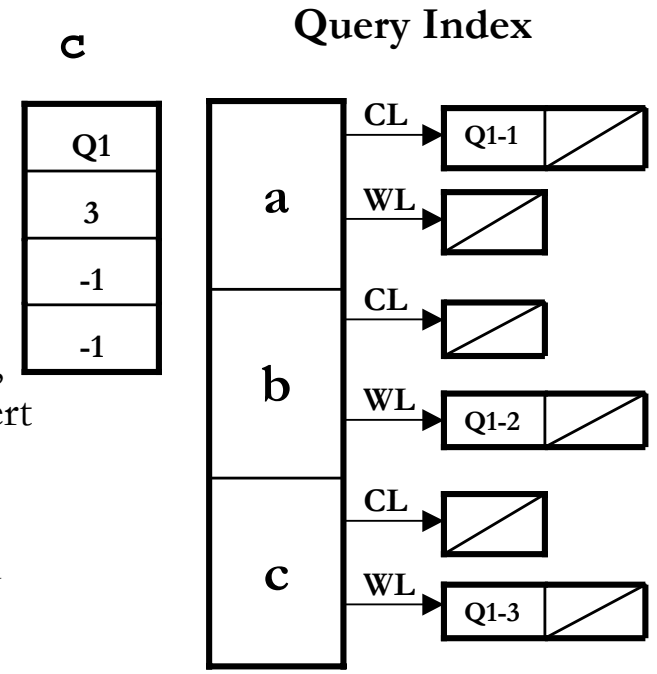
1	falls 1.Knoten und absolute Distanz zum Wurzel spezifiziert
-1	wenn RelativePos = -1
0	sonst



XFilter (4)

Ausführungsalgorithmus

- **Start Element Handler:**
  - Level Check: falls *Path Node* nicht negativen Levelwert enthält, müssen beide Levels gleich sein
  - Attribut Filter Check
  - Nach dem Bestehen beider Tests wird das nächste Element, falls vorhanden, von der *Wait List* in die *Candidate List* kopiert ~Zustandsübergang
- **End Element Handler:** beim Auftreten des Ende-Tag-Ereignisses wird entsprechender *Path Node* von der *Candidate List* gelöscht
- **Element Character Handler:**
  - Arbeitet ähnlich wie der Start Element Handler
  - Wird aufgerufen wenn der Dateninhalt gefiltert werden soll
  - Ist auch in der Lage den Zustandsübergang zu bewirken



Zusammenfassung

- XPath:
  - Adressierung der Knoten im XML-Baum
  - Basis für viele Anwendungen
- XQuery:
  - Anfrage- und Transformationssprache
  - De-facto Standard
- SPEX:
  - Sequentielle Bearbeitung
  - Netzwerk aus Transduktoren
- XMLTK:
  - System aus Werkzeugen
  - XPath-Prozessor

  Kein Speichermanagement ⇨ Beschränkte Anzahl von Elementen
- XFilter:
  - SDI-System
  - Filtering Engine

Danke für die Aufmerksamkeit!