Extraktion von XML aus HTML-Seiten

Das WYSIWYG-Werkzeug d_2c

Diplomarbeit von Max Völkel Betreuer: Dipl.-Inform. Markus L. Noga

IPD Goos, Institut für Programmstrukturen und Datenorganisation, Uni Karlsruhe (TH)

Web-Automatisierung

- Trends
  - Webseiten als universale Schnittstelle zu Informationen, Interaktiven Auskunftsdiensten
  - Wachstum von Informationsangebot und –nachfrage
- → Automatisierung notwendig
  - Semantik der Inhalte explizit machen

Anbieter oder Anwender?

- Anbieter
  - haben kein (wirtschaftliches) Interesse an maschinen-nutzbarer Darstellung
    - Werbe-Banner würden umgangen (Einnahmen)
    - Zusätzlicher Aufwand (Kosten)
- Anwender
  - Anwenderseitige Aufbereitung notwendig
  - Erstellung von Wrappern
    - Jetzt: Mühsam
    - Mit WYSIWYG einfacher, weniger Fehler

Aufgabenstellung

Erstelle ein System, mit dem Wrapper zur **Datenextraktion aus Webseiten möglichst einfach** erzeugt werden können.

- Ergebnis als XML per Webschnittstelle
- Modulare Architektur
- Plattformunabhängig

Teil I

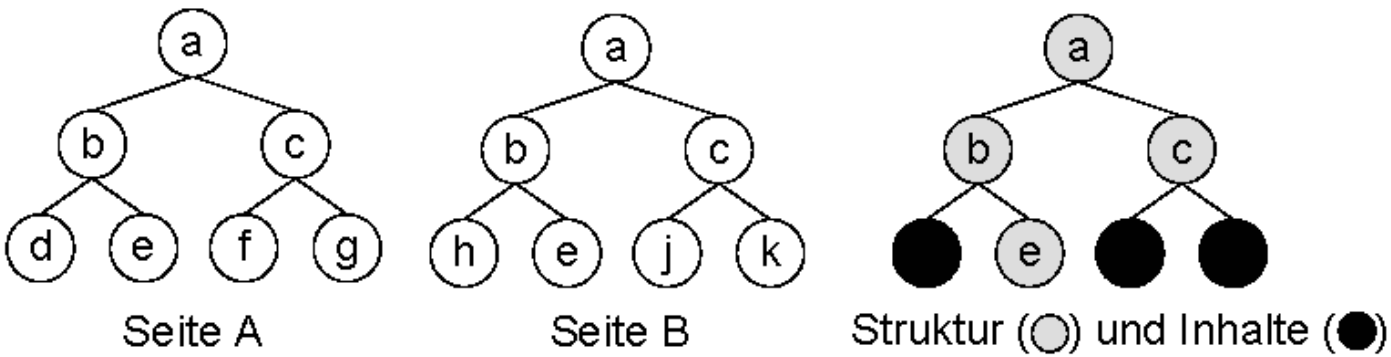
- **Grundlagen**
  - Struktur und Inhalte
  - Definition: Wrapper
  - HTTP und HTML in der Praxis
  - Definition: PH
- **Entwurf**
  - Kernidee, Entwurfsentscheidungen
  - Einschub: Die Rolle der Browser
  - Atomare und höhere Dienste
  - Wrapper in der Praxis
  - Kriterien
- **Stand der Technik**
  - Systeme zur Wrapper-Erzeugung
- **Architektur**

Teil II

- **Benchmark für PH-Zerteiler**
- **Evaluation**
  - Benutzerschnittstelle
  - Anhand der Kriterien
  - Anwendungs-möglichkeiten
- **Ausblick**

Struktur und Inhalte von Webseiten

- 80 % der Webseiten werden dynamisch generiert
  → gemeinsame *Struktur* mit verschiedenen *Inhalten* darin
- Für eine Menge M von (*gleichartigen*) Webseiten:
  - ***Struktur*** := gemeinsamer Schnittbaum der HTML-Elemente aller Seiten aus M
  - ***Inhalte*** := unterhalb des Strukturbaumes liegenden Unterbäume



- Wrapper nutzen Struktur, um Inhalte zu extrahieren

Definition: Wrapper

Wrapper sind Software-Artefakte, die von den Basistechnologien HTTP und HTML abstrahieren und als Schnittstelle Dokumente oder ausgewählte Teile davon zurückliefern.

HTTP und HTML in der Praxis selten standardkonform

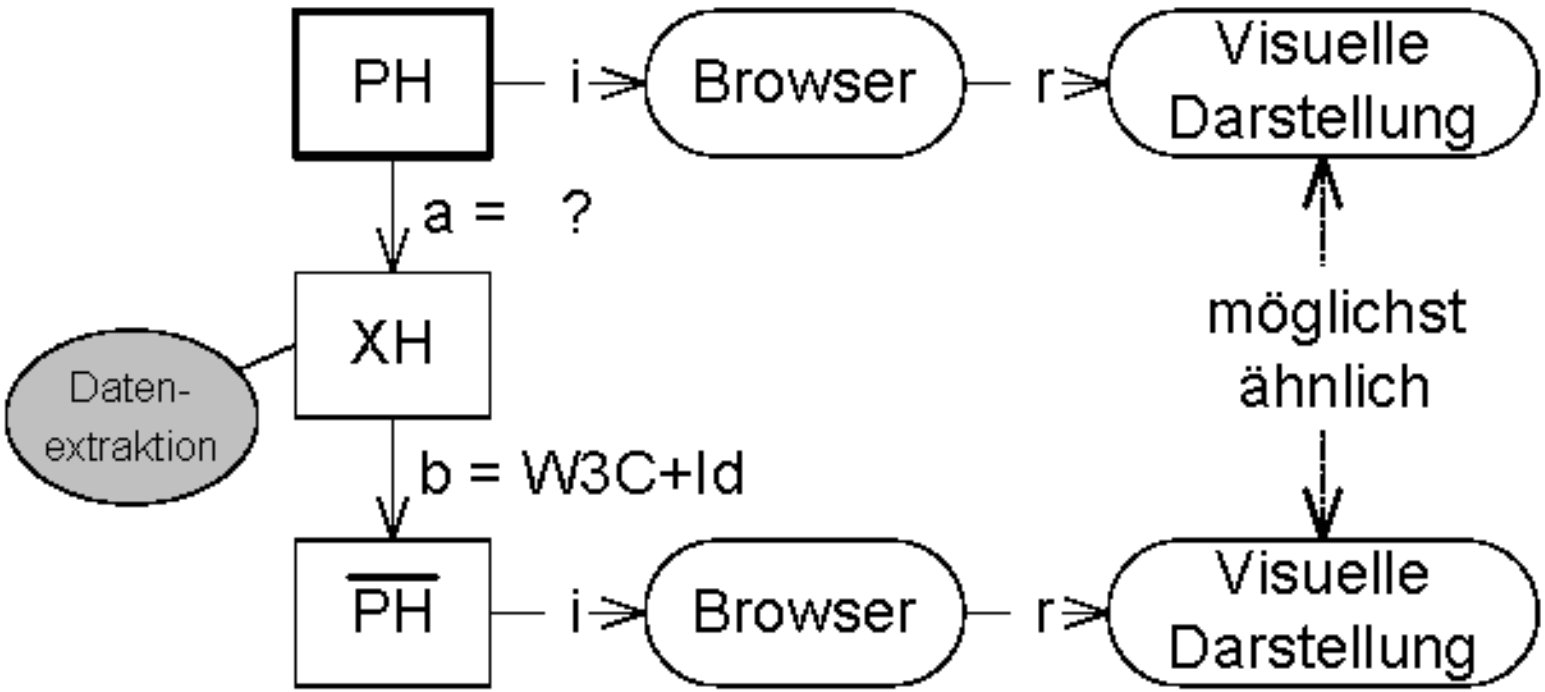
- HTTP
  - Wenige relevante Webserver
    → wenige HTTP-Dialekte
- HTML
  - Zahlreiche Autorenwerkzeuge; noch mehr Generatoren für dynamische Webseiten
  - Autoren prüfen visuell mit Browsern, diese erfüllen Standards nur teilweise
  - 0,24 % der Seiten sind syntaktisch korrekt

Definition von PH: *P*raxis-*H*TML

- Formal Zeichenkette
- Intentional HTML
- Browser zerteilen PH in einen Syntaxbaum
  - Zerteilung abhängig vom verwendeten Browser
  - marktbeherrschende Browser stellen den *de facto* Standard für die Interpretation von PH dar
  - → PH muss wie von einem solchen Browser interpretiert werden

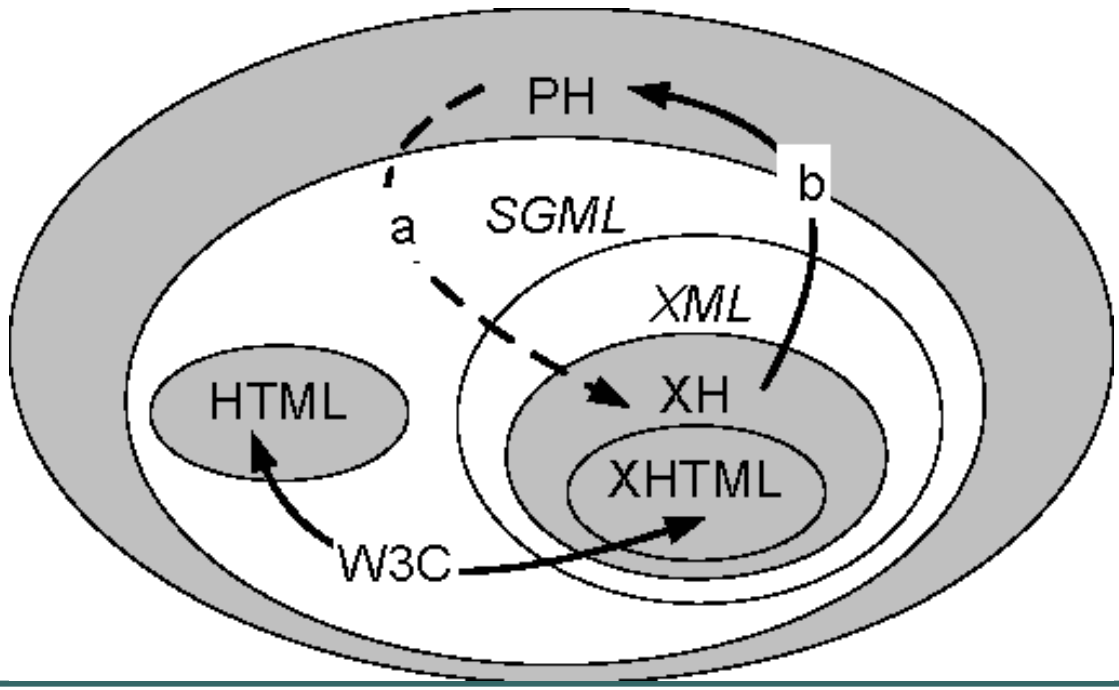
Kernidee des Entwurfs

Extraktion aus XML, welches analog einem Browser aus PH erzeugt wurde



Kernidee des Entwurfs

Extraktion aus XML, welches analog einem Browser aus PH erzeugt wurde



Woher kommt das XSLT-Stylesheet?

- PH wird als XH interpretiert
- Daten werden mit XSLT extrahiert
  - Optimierungen nutzbar
- Erstellen des XSLT-Stylesheets
  - Formulierung als Lernproblem
  - Selektiertes Element:
    - `<xslt:value-of select=„DOM-Pfad“/>`
  - → Eine Seite genügt
  - → Selektionsproblem

Einschub: Die Rolle der Browser

- Verbreitet zur Nutzung des WWW
- Werkzeug zur *visuellen* Datenintegration
  - Text, Bilder, CSS, JavaScript, Java, Plugins, ...
- Minimales Bedienungskonzept:
  - „Anklicken bei Interesse“
  - Anklicken von Verweisen und Knöpfen zur Navigation
- → Ziele für d_2c
  - als HTML darstellbar = als XML extrahierbar
  - Erstellung von Wrappern ähnlich zur Web-Navigation → WYSIWYG

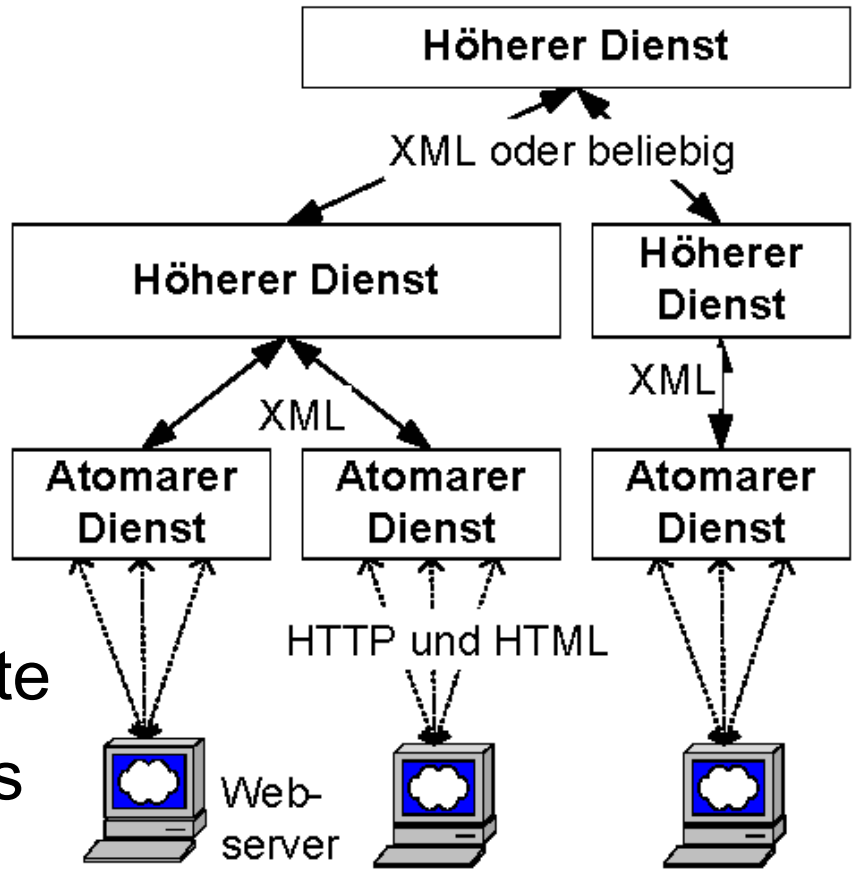
Selektionsproblem

- WYSIWYG
  - Einfach, intuitiv
  - Weniger Fehlermöglichkeiten
- Im Browser
  - Für Benutzer gewohnt, keine Kontextwechsel
  - Instrumentalisierte Webseiten im Browser
  - Erzeugt mit „**Filterndem Proxy**“
  - Bedienung über „**Bookmarklets**“

Beispiel für Selektion im Browser

Atomare und höhere Dienste

- Atomarer Dienst
  - Abstrahiert von HTTP und HTML
  - extrahiert
- Höherer Dienst
  - Nutzt atomare und andere höhere Dienste
  - Aggregiert, wertet aus



Schritte eines atomaren Dienstes

- 1) Herunterladen der HTML-Seite über HTTP
- 2) Extraktion
  - A) Fehlertoleranter Zerteiler bereitet PH in Baumstruktur auf
  - B) Anschließend werden aus der Baumstruktur Daten extrahiert
- 3) Speichern der extrahierten Daten zur Weiterverarbeitung

Wrapper in der Praxis

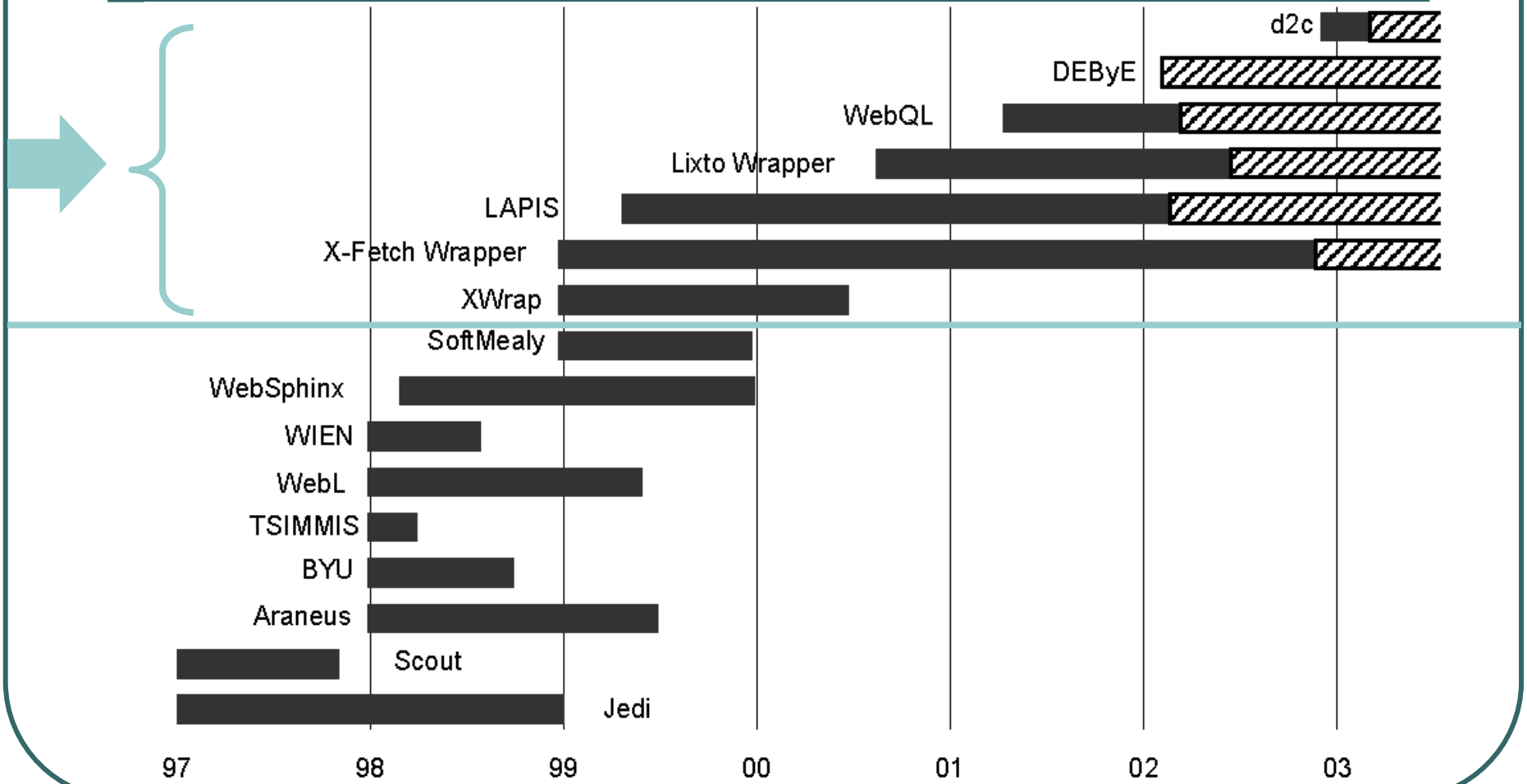
Webseiten ändern sich...

- Inhaltlich
  - → Wrapper liefert neue Daten
- Strukturell
  - → Problem?
  - Eher selten (2 x im Jahr)
  - Robuste Wrapper → Fehleranfällig
  - Wrapper-Verifikation

Kriterien

- 1) Eigenschaften erzeugter Wrapper
  - A) Atomare Dienste
  - B) Interpretation von PH wie ein Browser
- 2) Erstellung von Wrappern
  - Möglichst einfach → WYSIWYG im Browser
- 3) Kompatibilität und Wiederverwendung
  - A) Ergebnis als XML über Webschnittstelle
  - B) Modular, Standards

Systeme zur Wrapper-Erzeugung



Systeme zur Wrapper-Erzeugung

	PH-Modell	WYSIWYG im Browser?
DEBye	Text	■ ■ □ □
LAPIS	Text	■ □ □ □
Lixto	Baum	■ ■ □ □
WebQL	?	■ □ □ □
XWrap	Baum	■ ■ □ □
X-Fetch	Text	■ □ □ □

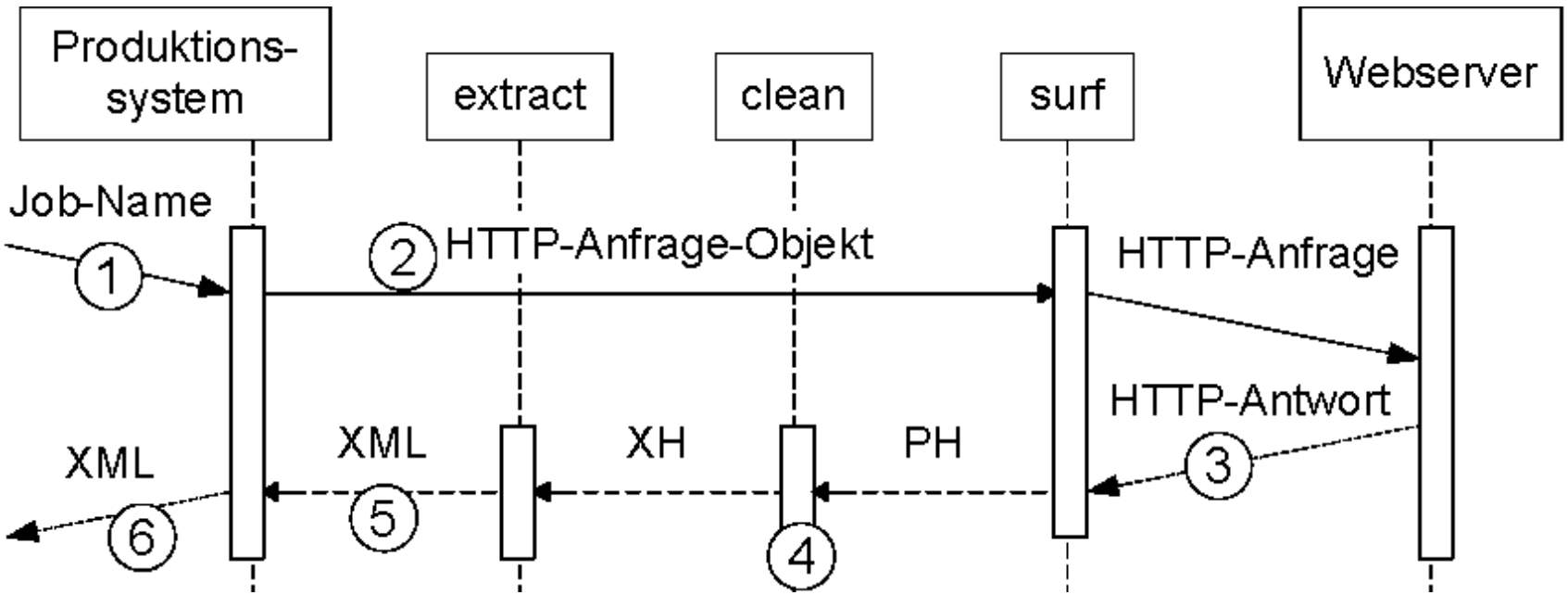
- Alle Systeme
  - Proprietäre Extraktionssprache
  - Kein PH-Zerteiler (RegExp. auf Text)

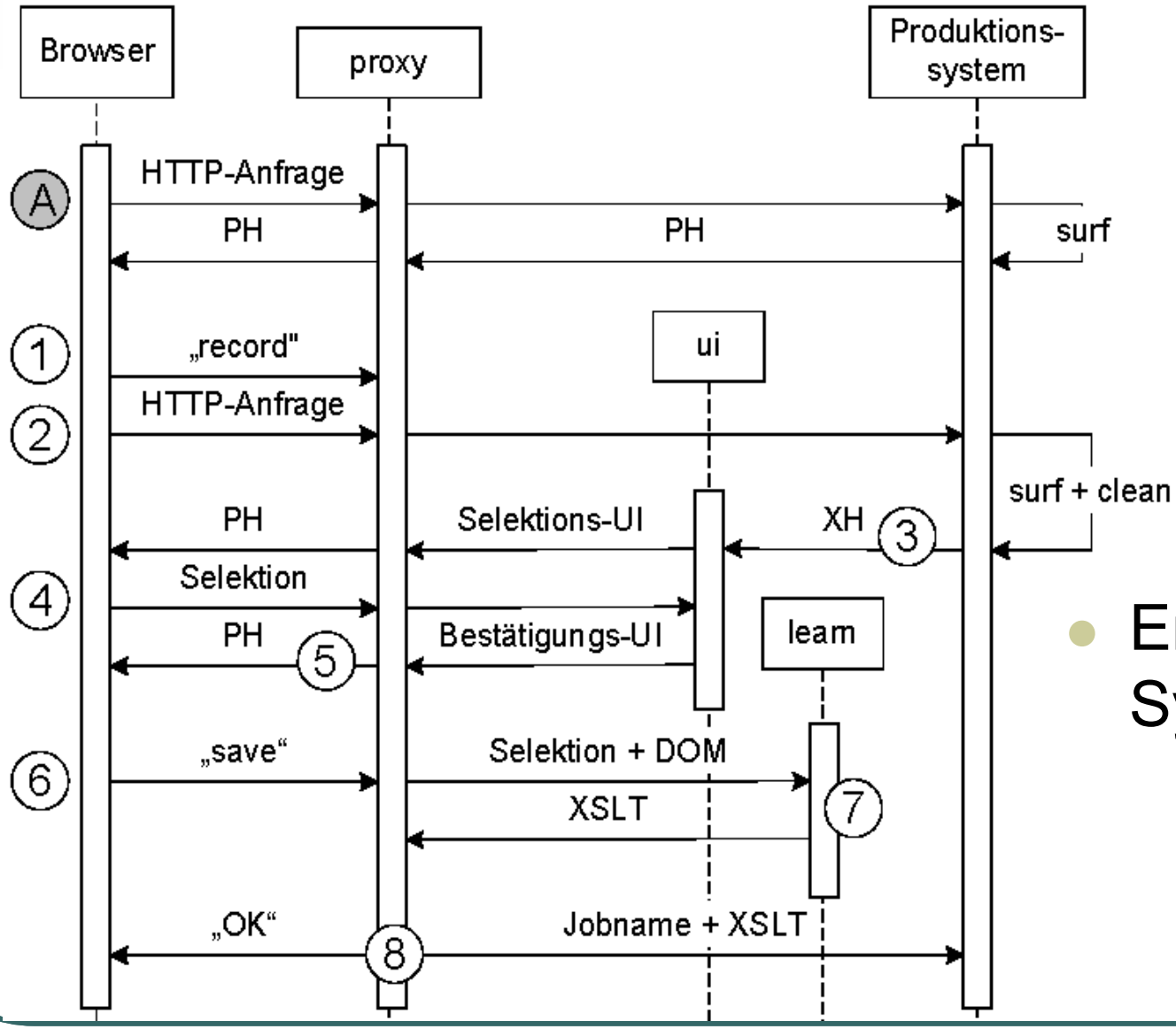
  oder
  - PH-Zerteiler geringer Qualität (Swing)

Nicht interaktiv | interaktiv | WYSIWYG | WYSIWYG im Browser

Architektur

- Produktionssystem



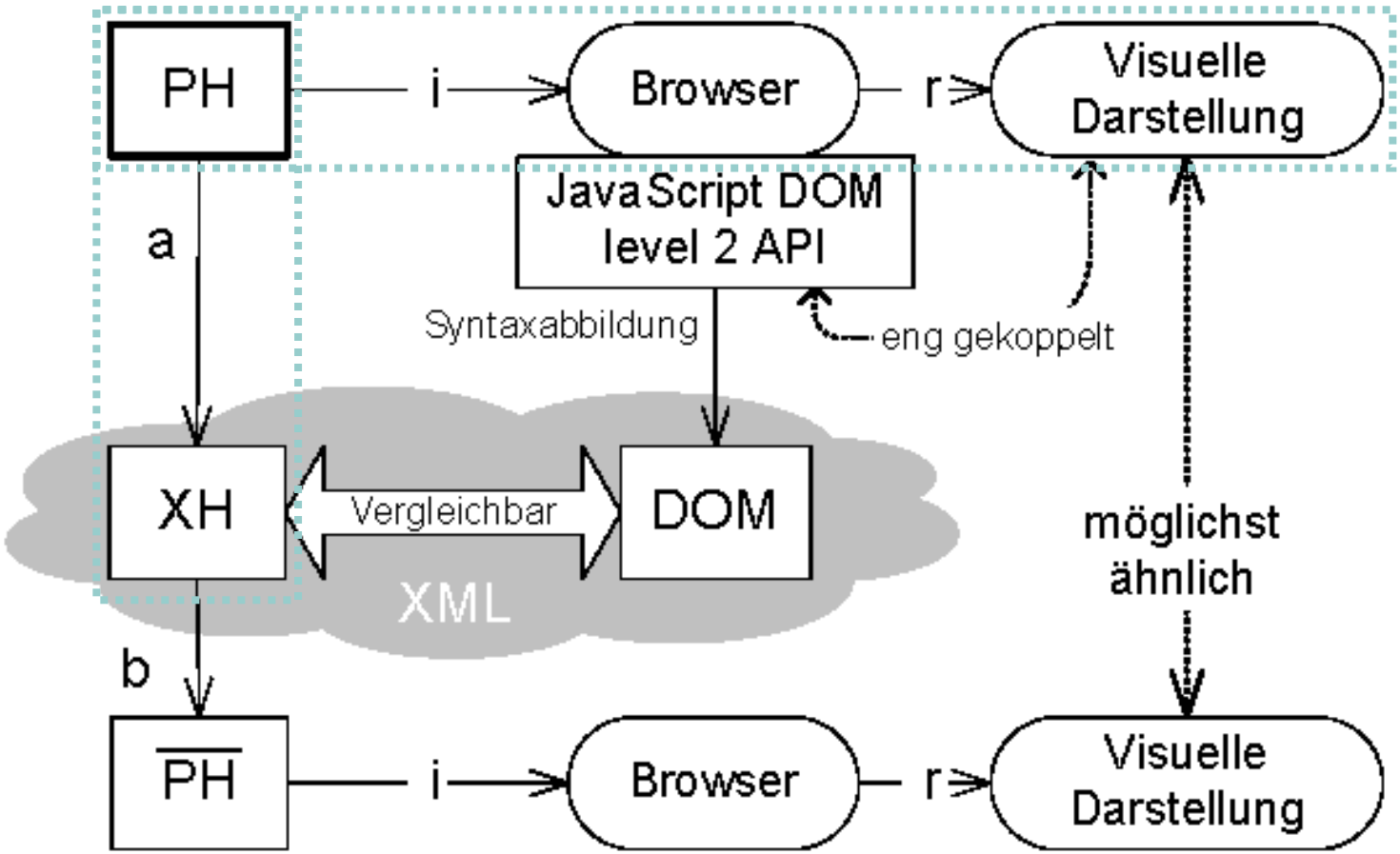


- Entwicklungs-System

Teil II

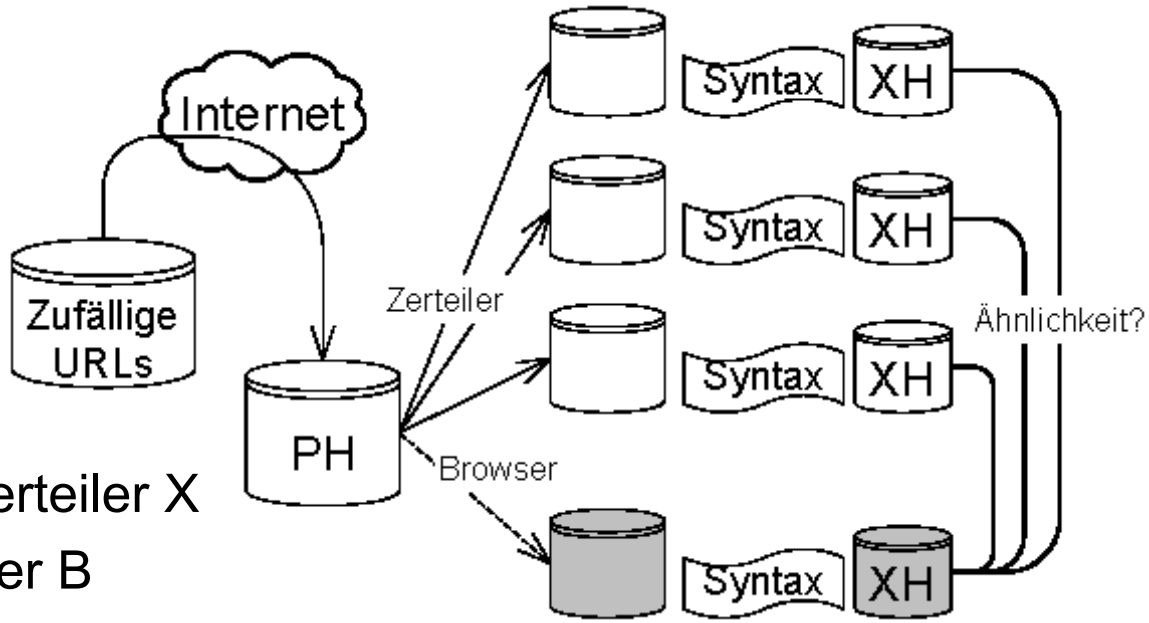
- Benchmark für PH-Zerteiler
- Evaluation
  - Benutzerschnittstelle
  - Anhand der Kriterien
  - Anwendungsmöglichkeiten
- Ausblick

Benchmark für PH-Zerteiler: Die Idee



Benchmark für PH-Zerteiler: Vorgehen

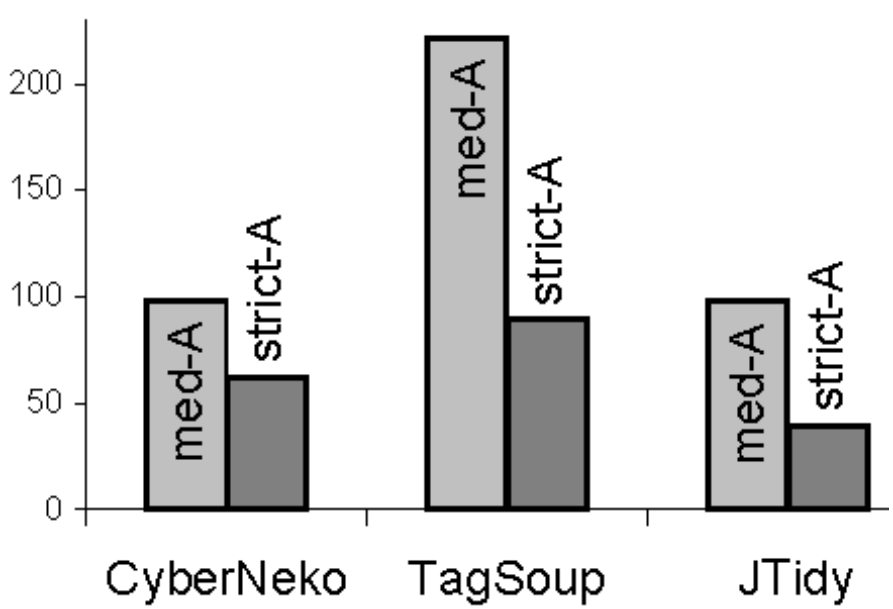
- Erzeugen von PH-Testdaten
- Zerteilen von PH als XH
  - Zu prüfender Zerteiler X
  - Browser-Zerteiler B
- Analyse
  - Syntaktisch korrektes XML?
  - Möglichst semantischer Vergleich der Ergebnisse von X mit denen von B

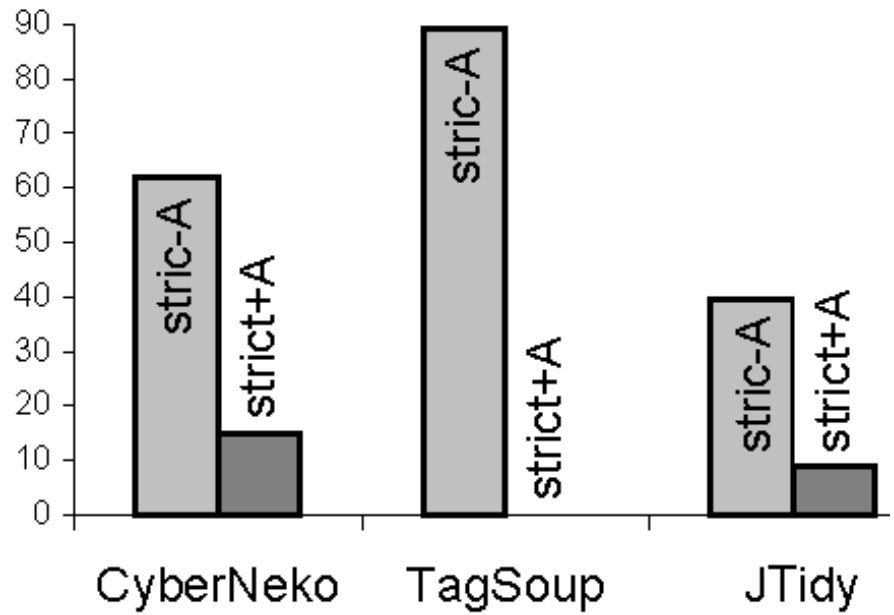


Ähnlichkeit von XML-Bäumen? Möglichst semantischer Vergleich

- Syntaktischer Vergleich uninteressant
  - <B><I>text</I></B> == <I><B>text</B></I>
- Minimale Editierdistanz von Bäumen
  - NP-vollständig
  - Viel Literatur, keine frei verfügbare Implementierung vorhanden
- Größter Gemeinsamer Schnittbaum (ggS) ab der Wurzel
  - Attribute berücksichtigen? Vergleich von Kinder-Knoten-Listen streng oder mit min. Editierdistanz?
  - Ähnlichkeit von A und B aus M := $$\frac{|\,ggS(A,B)\,|}{Max(\,|A|,\,|B|\,)}$$

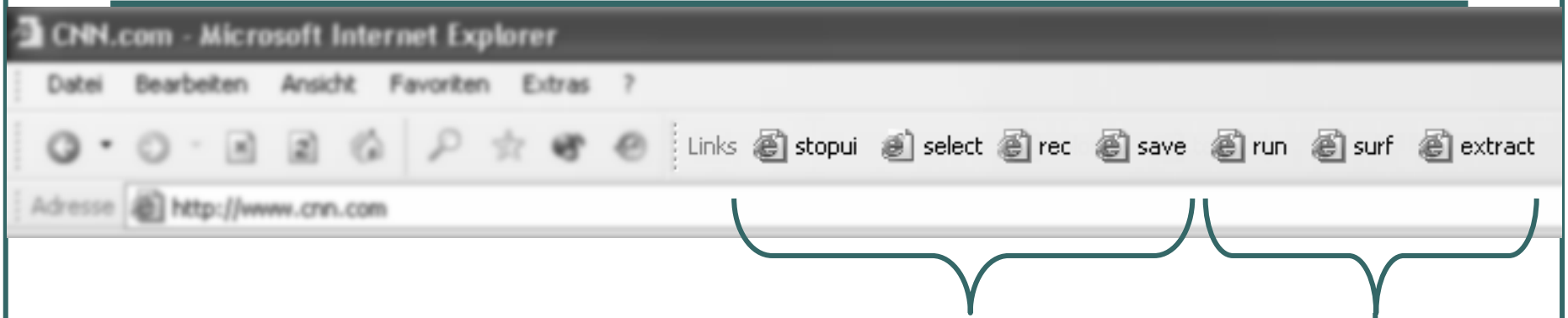
Benchmark für PH-Zerteiler: Ergebnisse





- Links: strenge Variante (strict) ~ min. Editierdistanz (med)
- Rechts: +A liefert extremere Werte (Attribute i.d. Wurzel)
- → scrict-A liefert brauchbare Ergebnisse (und schnell)

Benutzer-Schnittstelle



- Erstellen eines Wrappers
- Testen eines Wrappers
- Integration
  - Abruf der XML-Ergebnisse über URL mit Jobname als Parameter, weitere Parameter angebbar

CNN.com

FREE COUPONS: Click on a category below.
grocery coupons Go! | baby coupons Go! | online coupons Go! | free stuff Go! | store coupons Go!

SPECIAL COVERAGE: WAR IN IRAQ

irkuk Wolfowitz lays out plan for Iraq transition to democracy

SEARCH The Web CNN.com Search ENHANCED BY Google

Home Page
World
U.S.
Weather
Business at CNNMONEY
Sports at SI.com
Politics
Law
Technology
Science & Space
Health
Entertainment
Travel
Education
Special Reports

Get 4 FREE trial issues

Updated: 05:10 p.m. EDT (2110 GMT) April 10, 2003

Visit International edition

Kurds celebrate Thursday as they drive through the northern Iraqi town of Khanaqin.

Coalition pushes on northern front

- Iraqi army's 5th Corps and governor-general of Mosul expected to surrender, U.S. military sources tell CNN
- Iraqi forces "rapidly collapsing" as Kurds, U.S. forces push into Kirkuk
- Four Marines reported wounded after suicide bombing in Baghdad; Marine killed, 22 hurt in fight at mosque
- Prominent Shiite Muslim leader shot, stabbed to death in attack that began in Imam Ali Mosque in Najaf

DEVELOPING STORY

? **Video:** Celebrations in north | Baghdad palace search
? **Interactive:** U.S. Special Operations Forces | **Specs:** MOAB
? **Maps:** City of Baghdad | Troop movement
? **TIME.com:** Why Turks, Kurds both want Kirkuk

CNN.com

FREE COUPONS: Click on a category below.
grocery coupons Go! baby coupons Go! online coupons Go! free stuff Go! store coupons Go!

SPECIAL COVERAGE: WAR IN IRAQ

ays out plan for Iraq transition to democracy Suicide bomber :

SEARCH The Web CNN.com Search ENHANCED BY Google

- Home Page
- World
- U.S.
- Weather
- Business at CNNMONEY
- Sports at SI.com
- Politics
- Law
- Technology
- Science & Space
- Health
- Entertainment
- Travel
- Education
- Special Reports

Get 4 FREE trial issues of

Updated: 05:10 p.m. EDT (2110 GMT) April 10, 2003

Visit International edition

Kurds celebrate Thursday as they drive through the northern Iraqi town of Khanaqin.

Coalition pushes on northern front

- Iraqi army's 5th Corps and governor-general of Mosul expected to surrender, U.S. military sources tell CNN
- Iraqi forces "rapidly collapsing" as Kurds, U.S. forces push into Kirkuk
- Four Marines reported wounded after suicide bombing in Baghdad; Marine killed, 22 hurt in fight at mosque
- Prominent Shiite Muslim leader shot, stabbed to death in attack that began in Imam Ali Mosque in Najaf

DEVELOPING STORY

? **Video:** Celebrations in north | Baghdad palace search
? **Interactive:** U.S. Special Operations Forces | **Specs:** MOAB
? **Maps:** City of Baghdad | Troop movement
? **TIME.com:** Why Turks, Kurds both want Kirkuk


CNN.com
FREE COUPONS: Click on a category below.
grocery coupons Go!
baby coupons Go!
online coupons Go!
free stuff Go!
store coupons Go!
SPECIAL COVERAGE:
WAR IN IRAQ
q, watch CNN tonight 8 p.m. EDT
SEARCH
The Web
CNN.com
Search
ENHANCED BY Google
Home Page
World
U.S.
Weather
Business at CNNMONEY
Sports at SI.com
Politics
Law
Technology
Science & Space
Health
Entertainment
Travel
Education
Special Reports
Updated: 05:10 p.m. EDT (2110 GMT) April 10, 2003
Visit International edition
Kurds celebrate Thursday as they drive through the northern Iraqi town of Khanaqin.
(AP PHOTO)
Coalition pushes on northern front
• Iraqi army's 5th Corps and governor-general of Mosul expected to surrender, U.S. military sources tell CNN
• Iraqi forces "rapidly collapsing" as Kurds, U.S. forces push into Kirkuk
• Four Marines reported wounded after suicide bombing in Baghdad; Marine killed, 22 hurt in fight at mosque
• Prominent Shiite Muslim leader shot, stabbed to death in attack that began in Imam Ali Mosque in Najaf
DEVELOPING STORY
? Video: Celebrations in north | Baghdad palace search
? Interactive: U.S. Special Operations Forces | Specs: MOAB
? Maps: City of Baghdad | Troop movement
? TIME.com: Why Turks, Kurds both want Kirkuk
TIME
WILL HIS PLAN MAKE US SAFER?
Get 4 FREE trial issues of TIME

CNN.com

FREE COUPONS: Click on a category below.
grocery coupons Go! | baby coupons Go! | online coupons Go! | free stuff Go! | store coupons Go!

SPECIAL COVERAGE: WAR IN IRAQ

ar in Arab world Restoring order in Iraq, watch CNN tonight 8

SEARCH The Web CNN.com Search ENHANCED BY Google

- Home Page
- World
- U.S.
- Weather
- Business at CNNMONEY
- Sports at SI.com
- Politics
- Law
- Technology
- Science & Space
- Health
- Entertainment
- Travel
- Education
- Special Reports

Get 4 FREE trial issues of TIME

Updated: 05:10 p.m. EDT (2110 GMT) April 10, 2003

Kurds celebrate Thursday as they drive through the northern Iraqi town of Khanaqin.

Visit International edition

Coalition pushes on northern front

- Iraqi army's 5th Corps and governor-general of Mosul expected to surrender, U.S. military sources tell CNN
- Iraqi forces "rapidly collapsing" as Kurds, U.S. forces push into Kirkuk
- Four Marines reported wounded after suicide bombing in Baghdad; Marine killed, 22 hurt in fight at mosque
- Prominent Shiite Muslim leader shot, stabbed to death in attack that began in Imam Ali Mosque in Najaf

DEVELOPING STORY

? **Video:** Celebrations in north | Baghdad palace search
? **Interactive:** U.S. Special Operations Forces | **Specs:** MOAB
? **Maps:** City of Baghdad | Troop movement
? **TIME.com:** Why Turks, Kurds both want Kirkuk

CNN.com

SPECIAL COVERAGE: WAR IN IRAQ

U.S. reinforcements arrive in Kirkuk Wolfowitz lay

SEARCH The Web CNN.com ENHANCED BY Google

- World
- U.S.
- Weather
- Business at CNNMONEY
- Sports at SI.com
- Politics
- Law
- Technology
- Science & Space
- Health
- Entertainment
- Travel
- Education
- Special Reports

Updated: 06:10 p.m. EDT (2110 GMT) April 10, 2003

Visit International edition

Kurds celebrate Thursday as they drive through the northern Iraqi town of Khanaqin.

Coalition pushes on northern front

- Iraqi army's 5th Corps and governor-general of Mosul expected to surrender, U.S. military sources tell CNN
- Iraqi forces "rapidly collapsing" as Kurds, U.S. forces push into Kirkuk
- Four Marines reported wounded after suicide bombing in Baghdad; Marine killed, 22 hurt in fight at mosque
- Prominent Shiite Muslim leader shot, stabbed to death in attack that began in Imam Ali Mosque in Najaf

DEVELOPING STORY

? **Video:** Celebrations in north | Baghdad palace search
? **Interactive:** U.S. Special Operations Forces | **Specs:** MOAB
? **Maps:** City of Baghdad | Troop movement
? **TIME.com:** Why Turks, Kurds both want Kirkuk

Beurteilung mit Kriterien

- Erzeugte Wrapper
  - Beherrschen atomare Dienste
  - Interpretieren von PH mit TagSoup, ähnlich einem Browser
- Erstellung von Wrappern
  - durch WYSIWYG im Browser sehr einfach
- Kompatibilität und Wiederverwendung
  - Ergebnis als XML über Webschnittstelle abruft
  - Module leicht austauschbar (HttpClient, PH-Zerteiler, Servlet-Engine)
  - Nutzung von Standards: XML, XSLT, Servlets

Anwendungsmöglichkeiten

- Extraktion von Daten zur maschinellen Weiterverarbeitung
- Informations-Integrierende Dienste mit Kernproblem: Erstellung und Verwaltung von hunderten von Wrappern
  - Meta-Suchmaschinen
  - Preisvergleichsdienste
- Erstellung von Wrappern durch Laien
  - Z.B. als Service eines Portals

Ausblick

- Weiterer Zuwachs von Diensten, die von der Web-Oberfläche abstrahieren
  - Meta-Suchmaschinen
  - Suchmaschinen-Eintrags-Helfer
  - Sniper-Programme
  - Preisvergleichsdienste
- Weitere Entwicklung
  - Komplette Nutzung der Webschnittstelle über generierte Programmierschnittstelle
  - Browser-Markt stagniert
  - PH-Zerteiler werden besser

Vielen Dank für ihre Aufmerksamkeit.

Benchmark für PH-Zerteiler: Browser-DOM