

# Web Services – Einsatzmöglichkeiten für das Information Retrieval im WWW

Fabio Tosques & Philipp Mayr

Frankfurt am Main, den 24. Mai 2005

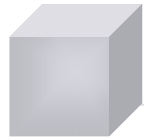
27. Online-Tagung der DGI 2005



# Überblick

---

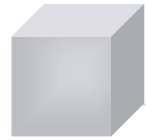
- Datenanalyse mittels „screen scraping“
- Datenanalyse mit Hilfe von Web Services
- Web Services - Entwicklung und Funktionsweise
- Web Services - Beispielanwendungen
- Web Services - Vorteile, Nachteile, Probleme
- Ausblick



# Problemstellung

---

- im WWW finden sich viele nützliche und interessante Daten:  
„The data is out there, the data is free, and the data is extremely interesting.“ (Erik Benson)
- Informationsgewinnung aus Daten im WWW
- Viele dieser Daten sind für das menschliche Auge aufbereitet, Programme können sie nur schwer interpretieren
- Wie können die Daten für die Informationsgewinnung verwertet werden?
- Wie sehen Daten aus, die „normalerweise“ geliefert werden?



## Fürs menschliche Auge ...

---

### [Deutsche Gesellschaft für Informationswissenschaft und ...](#)

Die wissenschaftliche und berufsständische Fachgesellschaft der deutschen Informationsspezialisten. Informationen über Mitgliedschaft, Arbeitskreise, ...

[www.dgd.de/](#) - 1k - [Im Cache](#) - [Ähnliche Seiten](#)

### [WG: Einige Highlights der \*\*DGI-Tagung\*\*](#)

... die gemeinsame **Tagung** um 10:30 Uhr im Raum Europa von **DGI**-Praesident Dr. Horst ... Heimreise geben und den Abschied von der **DGI-Tagung** 2001 erleichtern. ...

[www.ub.uni-dortmund.de/listen/inetbib/msg18189.html](#) - 6k - [Im Cache](#) - [Ähnliche Seiten](#)

### [DGI-Tagung 8/99 in Wien](#)

... Wir bemühen uns, das Umfeld für eine interessante **Tagung** vorzubereiten.

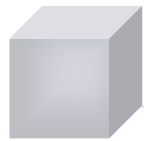
Der Erfolg der Veranstaltung hängt jedoch zum größten Teil von Ihrem Mitwirken ...

[www.medaustria.at/DGI99/](#) - 5k - [Im Cache](#) - [Ähnliche Seiten](#)

### [\[wiss-org\] - TAGUNG: DGI-Tagung](#)

... Frankfurt am Main, Messengelände Halle 4 Erstmals wurden die **DGI-Online-Tagung** und die **DGI**-Jahrestagung zu einer uebergreifenden **Tagung** zusammengelegt. ...

[index.bonn.iz-soz.de/~sigel/ ISKO/wiss-org-archive/msg00459.html](#) - 6k - [Im Cache](#) - [Ähnliche Seiten](#)

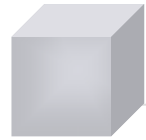


## ... und was die Maschine sieht

```

<table width=100% border=0 cellpadding=0 cellspacing=0><tr><td bgcolor=#3366cc><img
width=1 height=1 alt=""></td></tr></table><table width=100% border=0 cellpadding=0
cellspacing=0 bgcolor=#e5ecf9><tr><td bgcolor=#e5ecf9 nowrap><font
size=+1>&nbsp;<b>Web</b></font>&nbsp;</td><td bgcolor=#e5ecf9 align=right nowrap><font
size=-1 color=>Ergebnisse <b>1</b> - <b>10</b> von ungefähr <b>9.710</b> für <b>dgi
tagung</b>. (<b>0,18</b> Sekunden)&nbsp;</font></td></tr></table><!--a--><div><p
class=g><!--m--><a href=http://www.dgd.de/>Deutsche Gesellschaft für
Informationswissenschaft und <b>...</b></a><br><font size=-1>Die wissenschaftliche un
berufsständige Fachgesellschaft der deutschen<br>
Informationsspezialisten. Informationen über Mitgliedschaft, Arbeitskreise,
<b>...</b><br><font color=#008000>www.dgd.de/ - 1k - </font><nobr> <a class=fl
href="http://216.239.59.104/search?q=cache:8tXAax1H2q4J:www.dgd.de/+dgi+tagung&hl=de"
- <a class=fl
href="/search?hl=de&lr=&q=related:www.dgd.de/">Ähnliche&nbsp;&nbsp;Seiten</a></nobr></font>
<p class=g><!--m--><a href=http://www.ub.uni-dortmund.de/listen/inetbib/msg18189.html
Einige Highlights der <b>DGI</b>-<b>Tagung</b></a><br><font size=-1><b>...</b> die
gemeinsame <b>Tagung</b> um 10:30 Uhr im Raum Europa von <b>DGI</b>-Praesident Dr.
Horst<br>
<b>...</b> Heimreise geben und den Abschied von der <b>DGI</b>-<b>Tagung</b> 2001
erleichtern. <b>...</b><br><font
color=#008000>www.ub.uni-dortmund.de/listen/inetbib/msg18189.html - 6k - </font><nob
<a class=fl
href="http://216.239.59.104/search?q=cache:0LpX7KtqydsJ:www.ub.uni-dortmund.de/listen

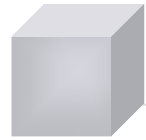
```



# Oder von Yahoo!...

---

1. [DGI / Tagungen und Fachkonferenzen](#)   
Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI) ... Mai. 27.  
**DGI-Online-Tagung.** Leitbild Informationskompetenz - Positionen, Praxis, Perspektiven im europäischen Wissens ...  
(zugleich 57. Jahrestagung der DGI) 23. Mai bis 25 ...  
[www.dgd.de/tagungen](http://www.dgd.de/tagungen) - 54k - [Im Cache](#) - [Weitere Seiten dieser Web-Site](#)
2. [Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. \(DGI\)](#)   
Vereinigung für Informationswissenschaft und Informationspraxis.  
Kategorie: [Firmen > Information > Berufs- und Fachverbände](#)  
[www.dgd.de](http://www.dgd.de) - 13k - [Im Cache](#) - [Weitere Seiten dieser Web-Site](#)
3. [fachmesse für wissensmanagement](#)   
... neue Internet-Präsenz für die COMiNFO 2004 und die 26. **DGI-Online-Tagung** "Information Professional 2011"  
besuchen ... COMiNFO 2003. 25. **DGI-Online-Tagung** Competence in Content ...  
[www.cominfo2004.de](http://www.cominfo2004.de) - 12k - [Im Cache](#) - [Weitere Seiten dieser Web-Site](#)



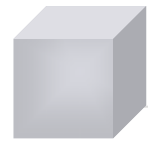
## ...und was die Maschine sieht

```

<div id=yschweb><div class=yschhd><h2>WEB-ERGEBNISSE</h2></div><ol start=1>
<li><div><a class=yschttl
href="http://de.wrs.yahoo.com/S=2114718003/K=dgi+tagung/v=2/SID=e/l=WS1/R=1/IPC=de/SH
/ Tagungen und Fachkonferenzen</a>
<a
href="http://de.wrs.yahoo.com/S=2114718003/K=dgi+tagung/v=2/SID=e/l=WS1/R=1/IPC=de/SH
target=_blank"></a>
</div>
<div class=yschabstr>Deutsche Gesellschaft für Informationswissenschaft und
Informationspraxis e.V. (<b>DGI</b>) ... Mai. 27. <b>DGI</b>-Online-<b>Tagung</b>.
Leitbild Informationskompetenz - Positionen, Praxis, Perspektiven im europäischen Wis
... (zugleich 57. Jahrestagung der <b>DGI</b>) 23. Mai bis 25 ...</div><em
class=yschurl>www.dgd.de/<b>tagung</b>en</em>
- <em>54k</em>

- <a
href="http://de.wrs.yahoo.com/S=2114718003/K=dgi+tagung/v=2/SID=e/l=WS5/R=1/;_ylt=AoT
Cache</a>
- <a
href="http://de.wrs.yahoo.com/S=2114718003/K=dgi+tagung/v=2/SID=e/l=WS3/R=1/;_ylt=Aip
Seiten dieser Web-Site</a>
<li><div><a class=yschttl
href="http://de.wrs.yahoo.com/S=2114718003/K=dgi+tagung/v=2/SID=e/l=WS1/R=2/SS=202506
Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (<b>DGI</b>)</a>

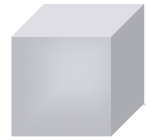
```



# Datengewinnung – mögliche Verfahren

---

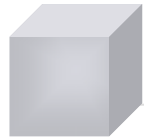
- Auswertung von Hand
- Auswertung mit speziellen Tools
- Auswertung mit Hilfe von Programm-Modulen
- „screen scraping“



# Screen scraping - Funktionsweise

---

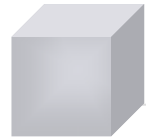
- html-Code von Hand analysieren
- html-Code parsen
- mit Hilfe von regulären Ausdrücken die gewünschten Daten extrahieren
- Daten weiterverarbeiten



# Screen scraping?

---

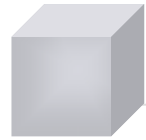
- „Screen scraping is the act of capturing data from a system or program by snooping the contents of some display that is not actually intended for data transport or inspection by programs. [...] Nowadays it often refers to parsing the HTML in generated web pages with programs designed to mine out particular patterns of content.“ (Wikipedia)



# Screen scraping - Probleme

---

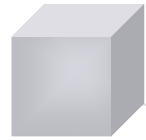
- die Struktur und der Aufbau von Webseiten ändert sich
  - Programme müssen aufwendig angepasst werden
  - jeder arbeitet mit eigenen „Hacks“
- ➔ „mediaeval torture“ (Lincoln Stein, Nature 2002)



## Web Services - Idee

---

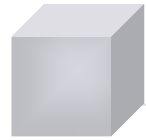
- wohldefinierte Schnittstellen sollen den Zugriff auf Daten erleichtern
- Idee solcher Schnittstellen ist nicht neu: Sun-RPC, COM+, CORBA
- neu ist: standardisierter Austausch mit Hilfe XML-kodierter Daten über bekannte Datenübertragungsprotokolle (http, https, smtp, ...)



## Web Services – worum geht es?

---

- „Ein Web Service ist eigentlich nicht viel mehr als eine dynamische Webseite. Der Unterschied besteht darin, dass normale Webseiten die Sprache HTML verwenden und sie normalerweise ein Browser dem Anwender präsentiert. Im Unterschied dazu liefert ein Web Service XML-Daten, die nicht für die unmittelbare Anzeige gedacht sind, sondern von einem speziellen Client-Programm weiterverarbeitet werden ...“ (c‘t 10/2005)



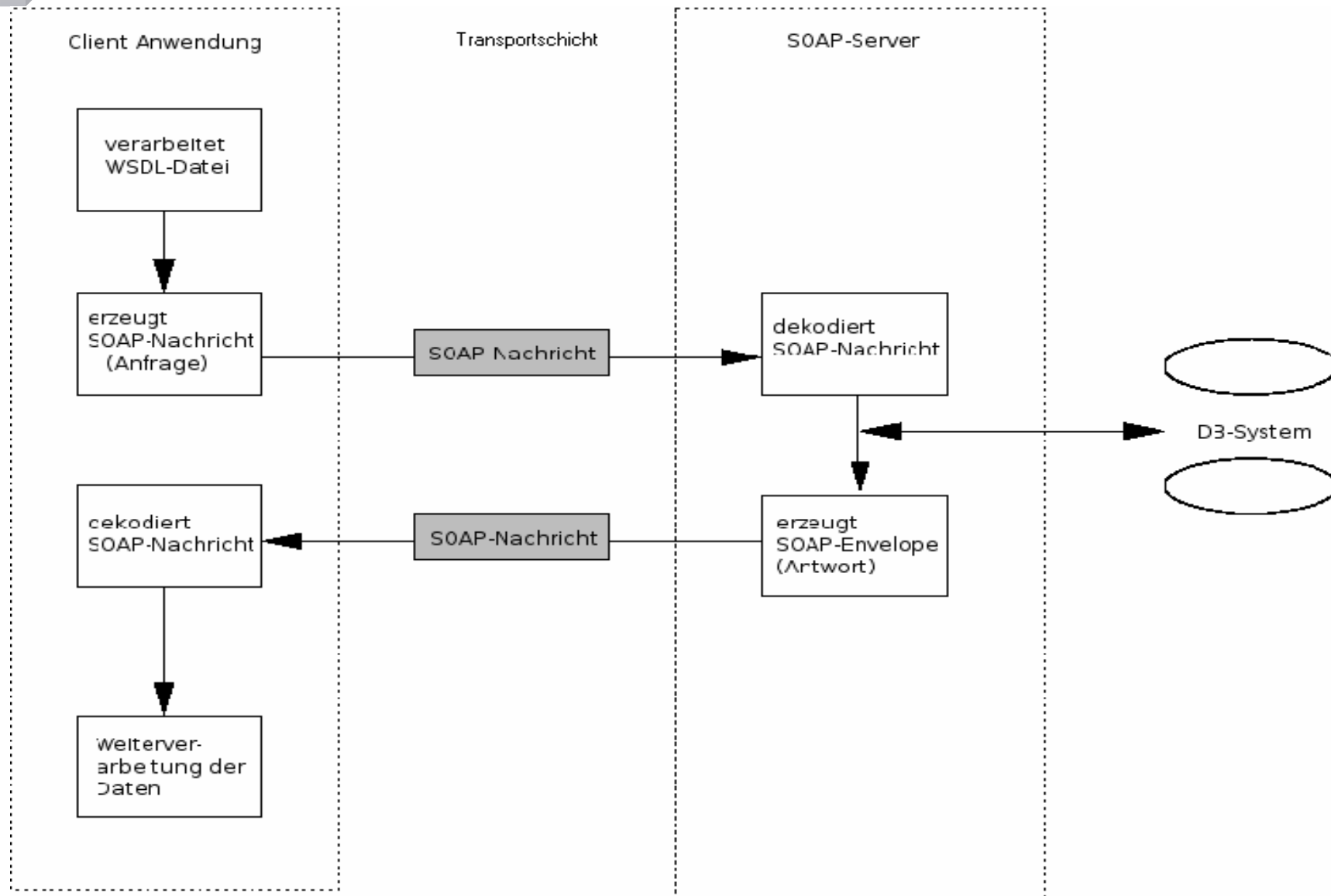
# Web Services - Entwicklung

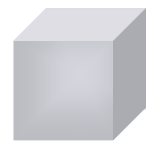
---

- 1998: Dave Winer veröffentlicht XML-RPC
- 2000: W3C-Konsortium (Microsoft, IBM, Sun ...) verabschieden SOAP-Spezifikation (Simple Object Access Protocol)
- 2000: REST (Representational State Transfer) wird von Roy Fielding in seiner Dissertation beschrieben
- 2001: WSDL (Web Services Description Language)
- 2002: Google und Amazon veröffentlichen Web-Service-Schnittstellen
- 2005: Yahoo! veröffentlicht Web-Service-Schnittstelle



# Web Services – Arbeitsweise mit SOAP/WSDL



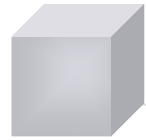


# Web Services – Datenstruktur (response)

```

- <Result>
  <Title>DGI / Tagungen und Fachkonferenzen</Title>
  - <Summary>
    Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI) ...
  </Summary>
  <Url>http://www.dgd.de/tagungen</Url>
  - <ClickUrl>
    http://rds.yahoo.com/S=96857362/K=dgi+tagung/v=2/XP=yws/SID=w/I=WS1/R=1/H=
  </ClickUrl>
  <ModificationDate>1116226800</ModificationDate>
  <MimeType>text/html</MimeType>
  - <Cache>
    - <Url>
      http://rds.yahoo.com/S=96857362/K=dgi+tagung/v=2/XP=yws/SID=w/I=WS5/R=1/I=
    </Url>
    <Size>53710</Size>
  </Cache>
</Result>

```



# Google vs. Yahoo!

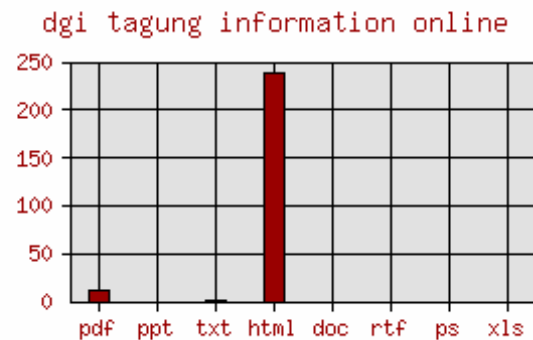
	<b>Yahoo!</b>	<b>Google</b>
Protokoll	REST	SOAP / WSDL
Registrierung	pro Anwendung	pro Schlüssel
max. Trefferzahl	50.000 pro IP	10.000 pro Schlüssel
Zugriffskontrolle	IP-basiert	Schlüssel-basiert
Suchmöglichkeiten	Image Search, Local Search, News Search, Video Search, Web Search	Web Search

# Beispiel 1: Dateitypanalyse mit Yahoo!

## Yahoo! Web Search: Filetype Analysis

Enter Your Query:

How many results?  ▼



Erstes PDF-Dokument an Stelle: 26

Erstes Text-Dokument an Stelle: 168

**Anzahl der bearbeiteten Dokumente: 250**

Dateityp	Anzahl	Prozent
html	238	95.20 %
pdf	11	4.40 %
txt	1	0.40 %

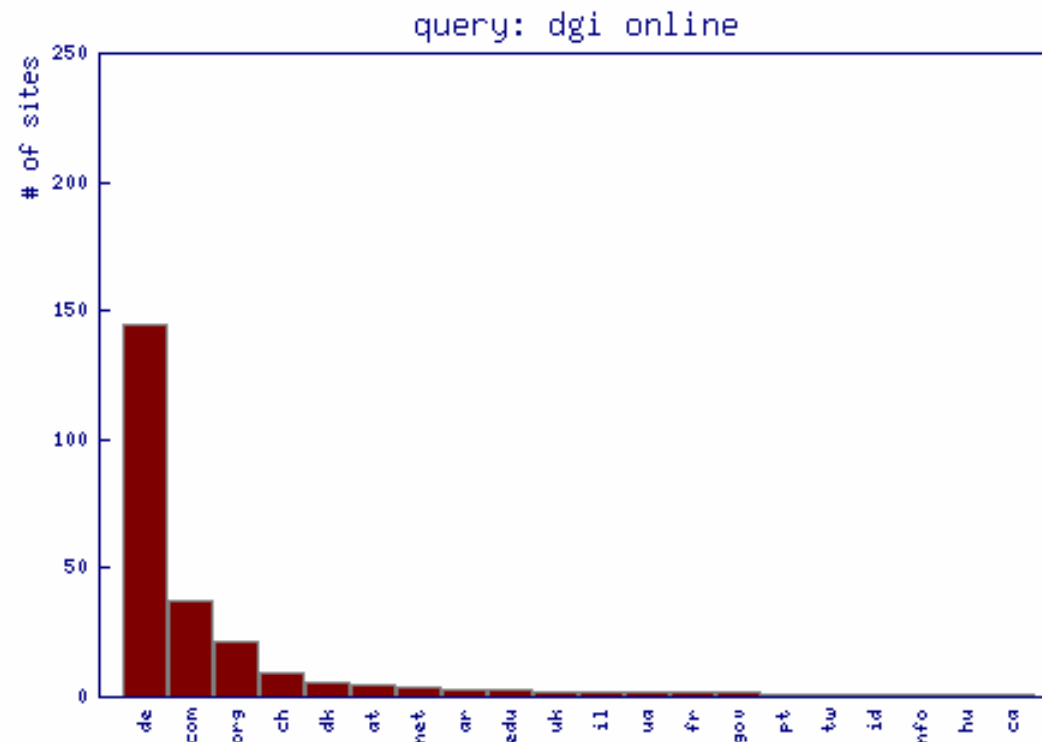
# Beispiel 2: Domainanalyse

## Analysing Top Level Domains

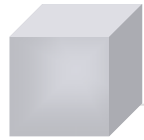
Enter Your Query here:

Please note that queries with more than 100 results can take one minute and more!

How many results?



# of tlds	rank	tld	with # of sites
1	1	de	145
2	2	com	37
3	3	org	21
4	4	ch	9
5	5	dk	6
6	6	at	5
7	7	net	4
8	8	ar	3
9	8	edu	3



## Beispiel 3: Permutation von Suchbegriffen

Enter Your Query here:

Enter: max. 4 Query terms or phrase search in apostrophe ""

Result Counts by Permutation	
Query	Count
tagung online dgi	2310
online tagung dgi	2310
online dgi tagung	2310
tagung dgi online	2320
dgi tagung online	2320
dgi online tagung	2330

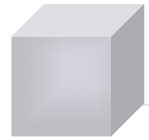
Top Results across Permutations	
Score	Result
57	<p>Die COMiNFO 2004 war schÄ¶n und wir danken unseren Ausstellern ...  <a href="http://www.cominfo2004.de/">http://www.cominfo2004.de/</a>            ... <b>DGI-Online-Tagung</b> "Information Professional 2011" besuchen. ... <b>DGI-Online-Tagung</b></p>

# Beispiel 4: Kombination von Yahoo! / Google

## Yahoo! Search WS / Google Web APIs

Enter Your query:

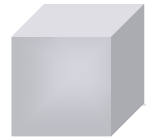
Yahoo! Search Web Services	Google Web APIs
Total: 4700	Total: 2330
1	1
<b>DGI / Tagungen und Fachkonferenzen</b> <a href="http://www.dgd.de/tagungen">http://www.dgd.de/tagungen</a>	<b>InterRed auf der DGI Online Tagung: Praxisworkshop für den Mittelstand</b> <a href="http://www.contentmanager.de/magazin/news_h11620_interr">http://www.contentmanager.de/magazin/news_h11620_interr</a>
2	2
<b>COMiNFO Forum CeBIT 2005</b> <a href="http://www.cominfo2005.de/">http://www.cominfo2005.de/</a>	<b>COMiNFO Forum CeBIT 2005</b> <a href="http://www.cominfo2005.de/">http://www.cominfo2005.de/</a>
3	3
<b>fachmesse für wissensmanagement</b> <a href="http://www.cominfo2004.de/">http://www.cominfo2004.de/</a>	<b>DGI-Online-Tagung 2005</b> <a href="http://www.online-tagung.de/">http://www.online-tagung.de/</a>



## weitere Beispiele

---

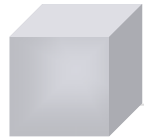
- Amazon Web Services
- Melvilsearch (Web-Service-Angebot DDC - Die Deutsche Bibliothek)
- Bioinformatik (Genomdaten)
- Berliner Flughafen
- Börse
- xmethods.org
- ...



# Zusammenfassung

---

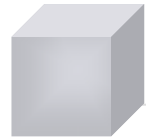
- bietet interessante Möglichkeiten der Datenanalyse
- XML-basierte Anwendung
- wird von allen bedeutenden Softwarefirmen unterstützt
- wird von Diensten im WWW vermehrt angeboten



# Vorteile von Web Services

---

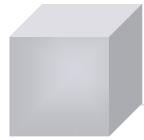
- Kontrolle der Daten und des Layouts
- erweiterte Suchmöglichkeiten können implementiert werden (z.B. Permutation)
- Weiterverarbeitung der Daten
- Kombination von verschiedenen Services
- Analyse der XML-Daten einfacher als bei binären Datenübertragungsprotokollen
- relative niedrige Einstiegshürde
- XML-Datenübertragung vereinfacht die Fehlersuche, da textbasiert
- Einsatz bekannter Protokolle (http, https, smtp)
- Plattformunabhängig



# Nachteile von Web Services

---

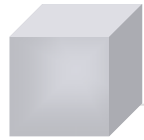
- Performance
- Hype
- Entwicklungsstadium
- Grundkenntnisse von Internetprotokollen erforderlich
- SOAP vs. REST
- Programmierkenntnisse
- offene Baustellen



# Ausblick

---

- interessante Entwicklung für Untersuchungen von Suchmaschinen (z.B. Webometrie)
- Recherchen und Auswertungen können automatisiert werden
- Kundengewinnung durch neue Services
- Microsoft .NET und MSN



## Kontakt

---

Vielen Dank für die Aufmerksamkeit!

Fabio Tosques, [abutux@web.de](mailto:abutux@web.de)

Philipp Mayr, [philippmayr@web.de](mailto:philippmayr@web.de)

Beispiele unter <http://bsd119.ib.hu-berlin.de/~ft/>