

Internet Briefing @ Bern
Wie Google alle Seiten indexiert...



Bern, 12. März 2008
Dr. Bernd Langkau, Partner

Menuvorschlag

	Thema	Stichworte
1	Ein Tag im Leben einer Suchmaschine...	<ul style="list-style-type: none"> • Abgrenzung → Der Beitrag behandelt nur einen Teil davon • Grundlegende Begriffe
2	Definieren und Feststellen der Vollständigkeit	<ul style="list-style-type: none"> • Wie ist meine Ausgangslage heute? • Sind alle Seiten das Ziel?
3	Technische Anpassungen	<ul style="list-style-type: none"> • Typische Probleme und Stolpersteine für Suchmaschinen • Bewertung resp. bessere Produktwahl
4	Betrieb	<ul style="list-style-type: none"> • Überwachen und Erhalten des erreichten Status

Ein Tag im Leben einer Suchmaschine...



Vier Hauptaufgaben einer Suchfunktion

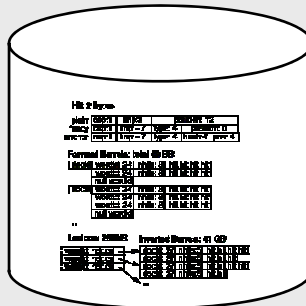
1. Daten

Akquisition und Speicherung aller Dokumente der Kollektion



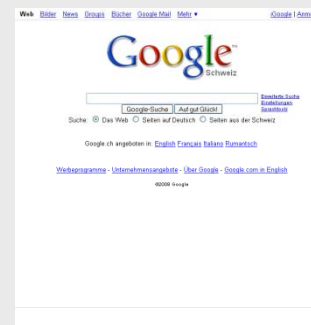
2. Index

Erstellen einer effizienten Datenstruktur für die Suche



3. Suchanfrage

Finden passender Dokumente zu einer Benutzeranfrage



4. Resultatliste

Präsentation der Treffer in der richtigen Reihenfolge



1. Daten

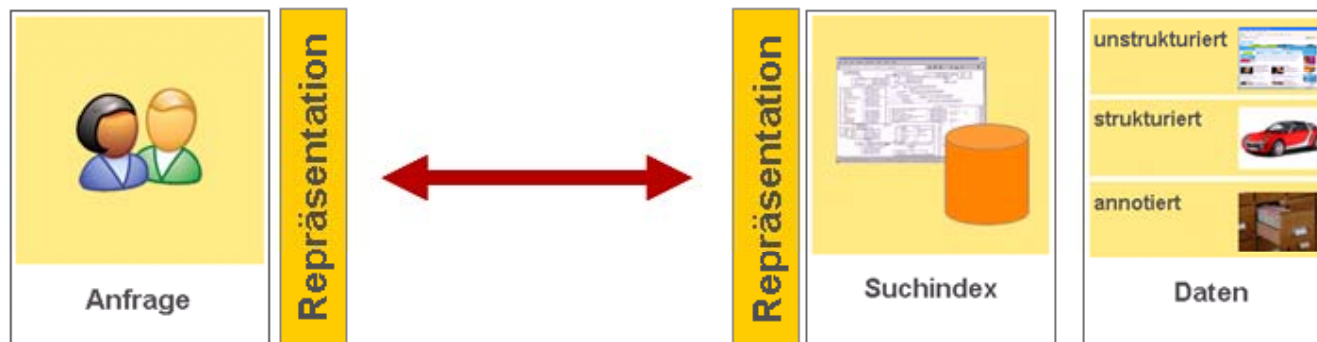
- » “Daten” sind im Kontext einer Websuche
 - Webseiten = [X]HTML Base Pages
 - darin referenzierte, binäre Dokumente (PDF, gif)
- » Ein einzelnes Element muss eineindeutig durch eine URL referenzierbar sein
 - <http://blog.namics.com/seosem/>
 - http://blog.namics.com/images/namics_logo.gif
- » Die Suchmaschine entscheidet, welche Datenformate übernommen werden (z.B. PDF, DOC, Flash/flv etc.)
- » Gesammelt werden die Daten von “Crawlern”, “Robots” “Gatherer” oder “Fetch”

2. Index

- » Dies ist eine Aufgabe, welche die Suchmaschine zu lösen hat (“nicht unser Bier”)
- » Mächtigkeit des Index (Anzahl Features) definiert die verfügbaren Funktionen bei der Suche
- » Technisch ist dieser Schritt hochgradig beherrscht

3. Suchanfrage

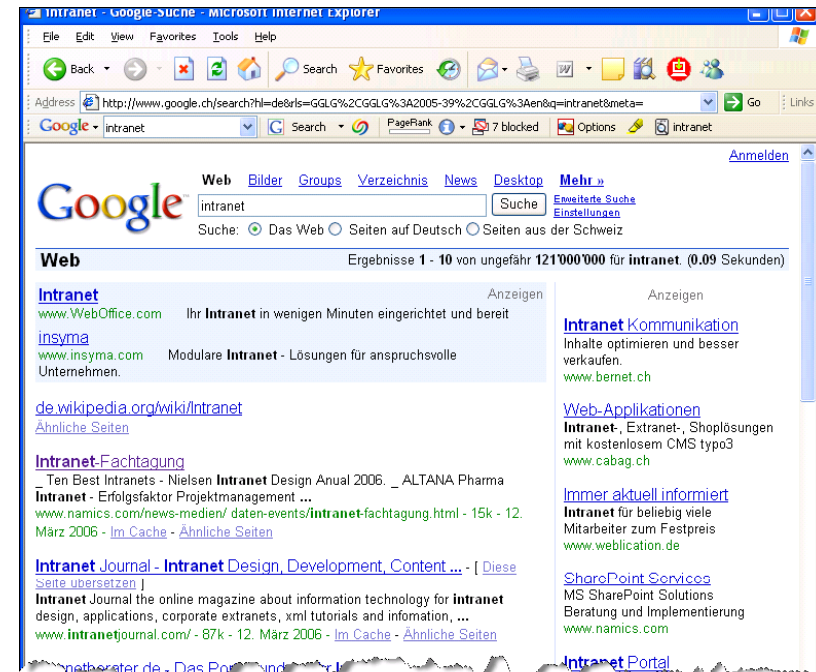
- » Die Anfrage (=Query) ist die textuelle Formulierung eines Informationsbedürfnisses durch einen User



- » Suchmaschine macht einen Wortvergleich mit dem Index (unter Berücksichtigung von Einschränkungen und Steuerungsparametern)
- » Komplex: Suchaktivitäten, Vorwissen des Users, Subjektivität der Relevanz, Synonyme/Homonyme etc.
- » **Suchmaschinenmarketing → Keywordanalyse**

4. Resultatliste

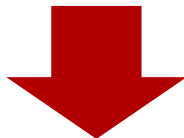
- » «Heiliger Gral» der Suchmaschine
 - 121 Mio. Treffer; wer steht auf Platz 1, 2 und 3?
- » Viele Unwahrheiten
- » Viele unseriöse Anbieter
- » Google-Regeln
 - <http://www.google.com/support/webmasters/bin/answer.py?answer=35769>
 - <http://www.google.com/support/webmasters/bin/answer.py?answer=35291>
- » Suchmaschinenmarketing → “organische Optimierung”



Abgrenzung

- » Finden die Daten ihren Weg nicht in den Index, so können sie später weder gefunden noch beim Ranking optimiert werden...

... der vorliegende Beitrag fokussiert NUR darauf, dass Daten in den Index aufgenommen werden.



Definieren und Feststellen der Vollständigkeit

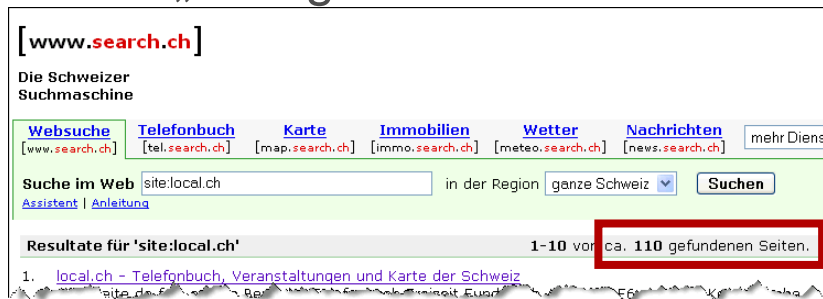


Vollständigkeitsanalyse

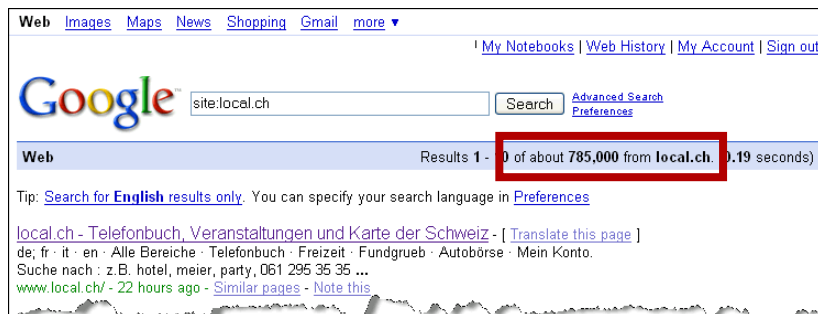
- » Zielsetzung: Alle Seiten des Webangebotes / der Webanwendung sollen vom Suchmaschinen-Crawler besucht (und heruntergeladen) werden
- » Ansatz
 - Was ist im Index: Suchmaschinen mit Query fragen
 - Was ist im Index: Suchmaschinen-Tools
 - Was wird effektiv besucht: Analyse der Access-Logs
- » Was tun?
 1. Wie viele Seiten habe ich: Zählen im Dateisystem, fragen des CMS, fragen der Datenbank etc.
 2. Was sagt Suchmaschine / Access-Log
 3. Quotient rechnen...
 4. Optional: Gruppierung nach Bereichen / Wichtigkeit

Vollständigkeitsanalyse: Was ist im Index (1)? (nicht exakt aber guter Anhaltspunkt)

» Ist das „richtig“



» oder das?

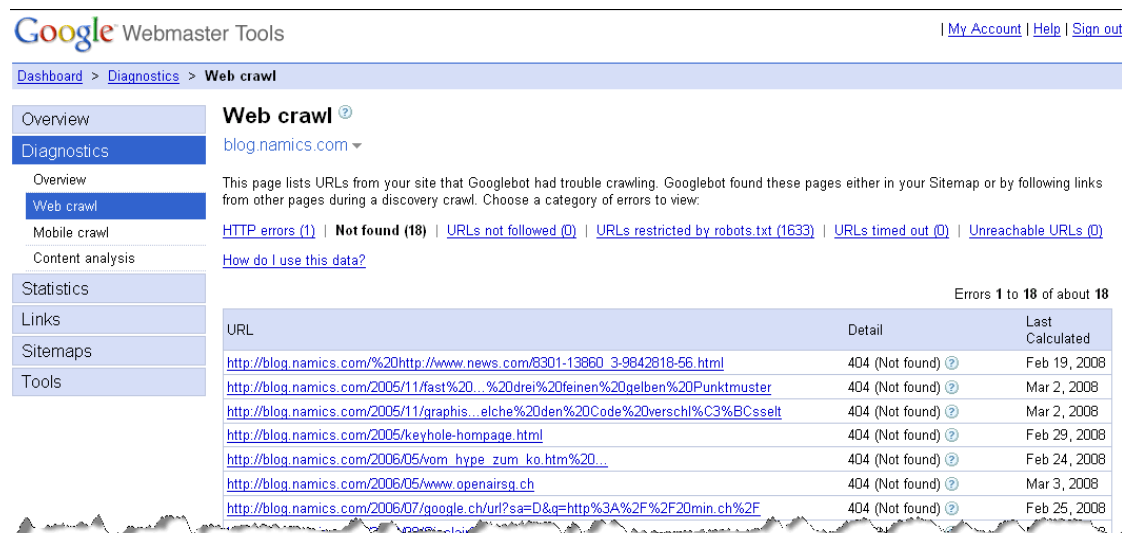


» Query (fast überall): *site:Domänennname*

» Tipp: An das Ende der Liste klicken

Vollständigkeitsanalyse: Was ist im Index (2)?

- » Suchmaschinen bieten für Webmaster Tools gratis an
 - Google Webmaster Central:
<http://www.google.com/webmasters/>
 - Yahoo! Site Explorer:
<http://siteexplorer.search.yahoo.com/>
- » Diese geben Auskunft über die indexierten Seiten (insb. über Crawling-Fehler und eingehenden Links)



Google Webmaster Tools | My Account | Help | Sign out

Dashboard > Diagnostics > Web crawl

Web crawl [?]
blog.namics.com

This page lists URLs from your site that Googlebot had trouble crawling. Googlebot found these pages either in your Sitemap or by following links from other pages during a discovery crawl. Choose a category of errors to view:

HTTP errors (1) | **Not found (18)** | URLs not followed (0) | URLs restricted by robots.txt (1633) | URLs timed out (0) | Unreachable URLs (0)

[How do I use this data?](#)

Errors 1 to 18 of about 18

URL	Detail	Last Calculated
http://blog.namics.com/%20http://www.news.com/8301-13860_3-9842818-56.html	404 (Not found) [?]	Feb 19, 2008
http://blog.namics.com/2005/11/fast%20...%20drei%20feinen%20gelben%20Punktmuster	404 (Not found) [?]	Mar 2, 2008
http://blog.namics.com/2005/11/graphis...elche%20den%20Code%20versch%20C3%BCsselt	404 (Not found) [?]	Mar 2, 2008
http://blog.namics.com/2005/keyhole-hompage.html	404 (Not found) [?]	Feb 29, 2008
http://blog.namics.com/2006/05/vom_hype_zum_ko.htm%20...	404 (Not found) [?]	Feb 24, 2008
http://blog.namics.com/2006/05/www.openairsq.ch	404 (Not found) [?]	Mar 3, 2008
http://blog.namics.com/2006/07/google.ch?url?sa=D&q=http%3A%2F%2F2F20min.ch%2F	404 (Not found) [?]	Feb 25, 2008

Vollständigkeitsanalyse: Was wird vom Crawler besucht? (exakt für GET-Requests)

- » Das Access Log des Server kennt die Antwort!
- » User Agent der grossen Suchmaschinen (raten ist einfach): <http://www.jafsoft.com/searchengines/webbots.html>
- » Aufpassen auf caching Proxies und selten indexierte URIs → lange Periode analysieren

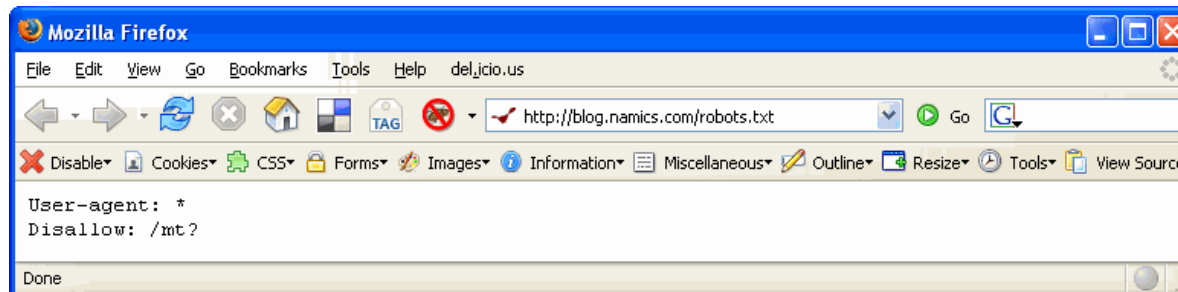
Spider-Besuche				
	Browser	Hits	% von Gesamt	Sitzungen
1	Googlebot	12,201	9.04%	4,417
2	FAST WebCrawler	11,007	8.82%	1,007
3	Cosmos	3,366	2.49%	698
9	KIT-Fireball	304	0.22%	114
10	search.ch V1.4.2 (spiderman@search.ch; http:	70,157	52.02%	92
11	Scooter-vv3.1.2	7,104	5.51%	80

Selektivität (Konzentration der Kräfte + „Müll“)

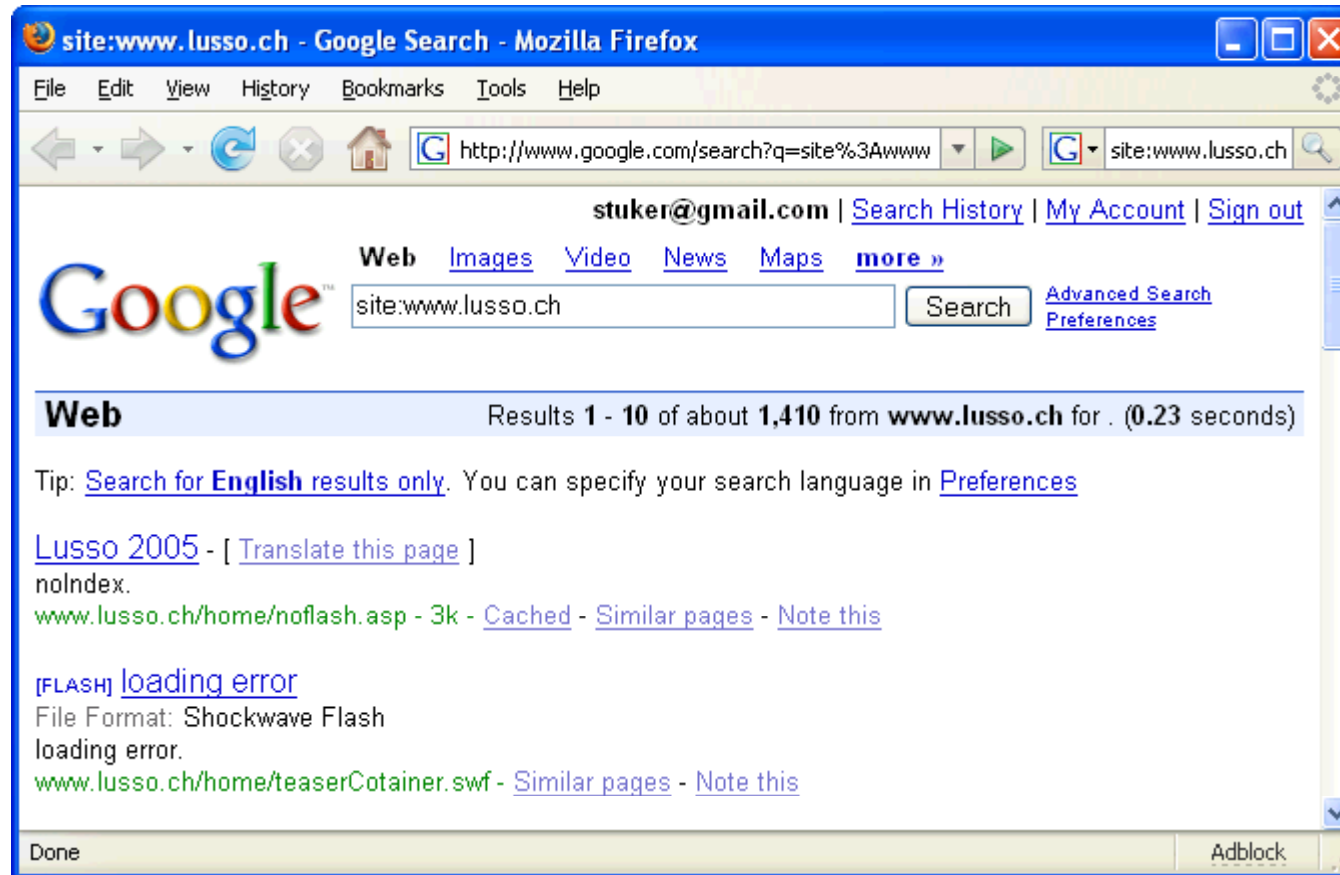
- » Sie wollen nicht alle Seiten in der Suchmaschine!
 - z.B. Login für Autoren, personalisierte Ansichten etc.
- » Zwei Möglichkeiten des Ausschlusses
 - Kennzeichnung im HTML (clientseitig)
(<http://www.robotstxt.org/wc/meta-user.html>)

```
<meta name="robots" content="index, follow">
<meta name="robots" content="noindex, follow">
<meta name="robots" content="index, nofollow">
<meta name="robots" content="noindex, nofollow">
```

- Kennzeichnung auf dem Server in der Datei /robots.txt
(<http://www.robotstxt.org/wc/exclusion-admin.html>)



Selektivität: Will ich die URI drin?



» Tipp: Viel gefährlicher sind insb. Testsysteme!

Zusammenfassung: Vollständigkeit

- » Als Betreiber eines Angebotes, wissen Sie
 - welche Seiten Sie anbieten
 - welche Seiten Sie nicht in der Suchmaschine wollen
- » Die Suchmaschine
 - sagt Ihnen, welche Seiten im ihrem Index vorhanden sind
 - hinterlässt Spuren beim Besuch

- » Zielerreichung ist bekannt resp. lässt sich rechnen

Technische Anpassungen



Drei Ziele

- » Zugänglichkeit
- » Interpretierbarkeit
- » Referenzierbarkeit

Zugänglichkeit für Crawler

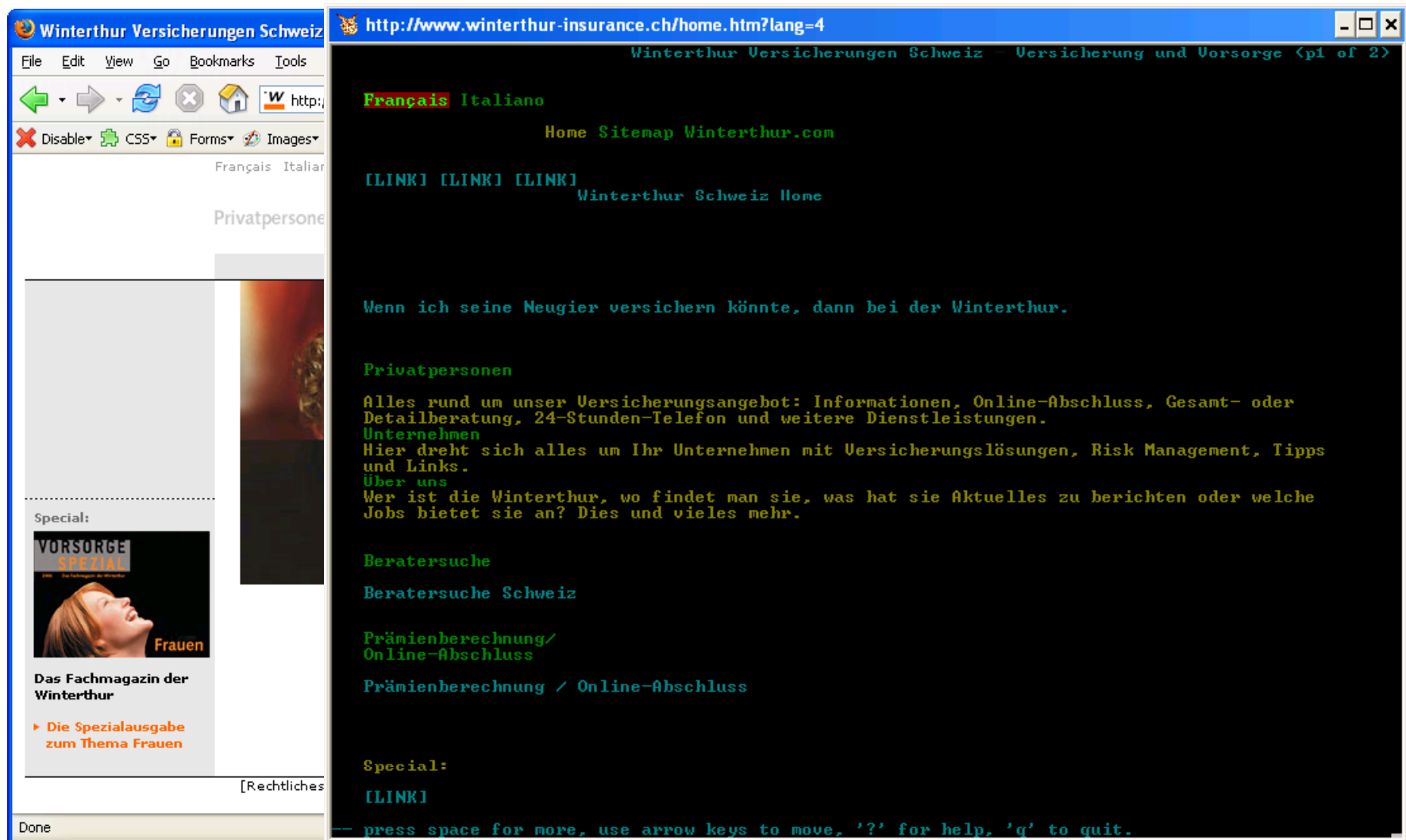
- » Crawler / Spider / Gatherer / Fetch
 - lädt und speichert HTML jeder Seite (Base Page Download)
 - extrahiert alle Links
 - folgt sämtlichen Links rekursiv

- » Was machen Crawler nicht

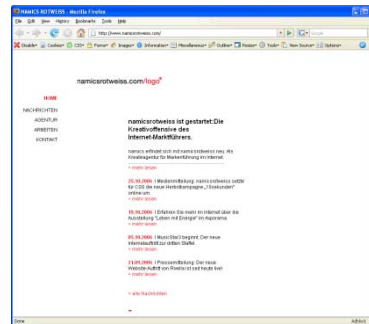
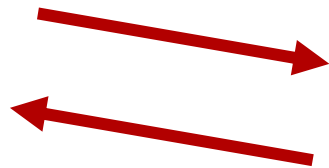
- JavaScript ausführen
- Cookies annehmen
- zu viele Query-Parameter mitnehmen (Faustregel: 2)
- zu viele Redirects folgen
- https Verbindungen folgen

**Faustregeln
(es gibt Ausnahmen)**

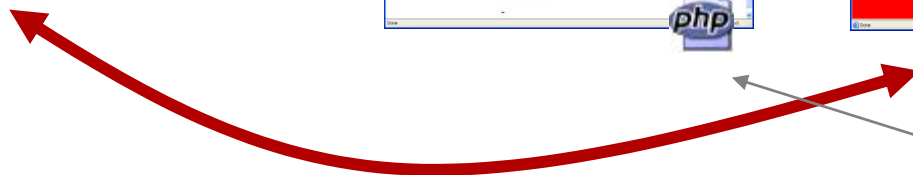
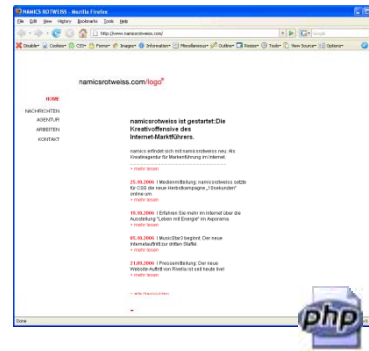
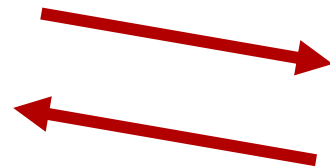
Zugänglichkeit: Test (mit Lynx Browser)



Zugänglichkeit von Flash



Es wird zuerst immer die HTML-Seite geliefert. Darin ist ein JavaScript-Include, welches eine Umleitung nach einem Plugin-Check prüft...



? JavaScript == ja
? FlashPlugin == ja



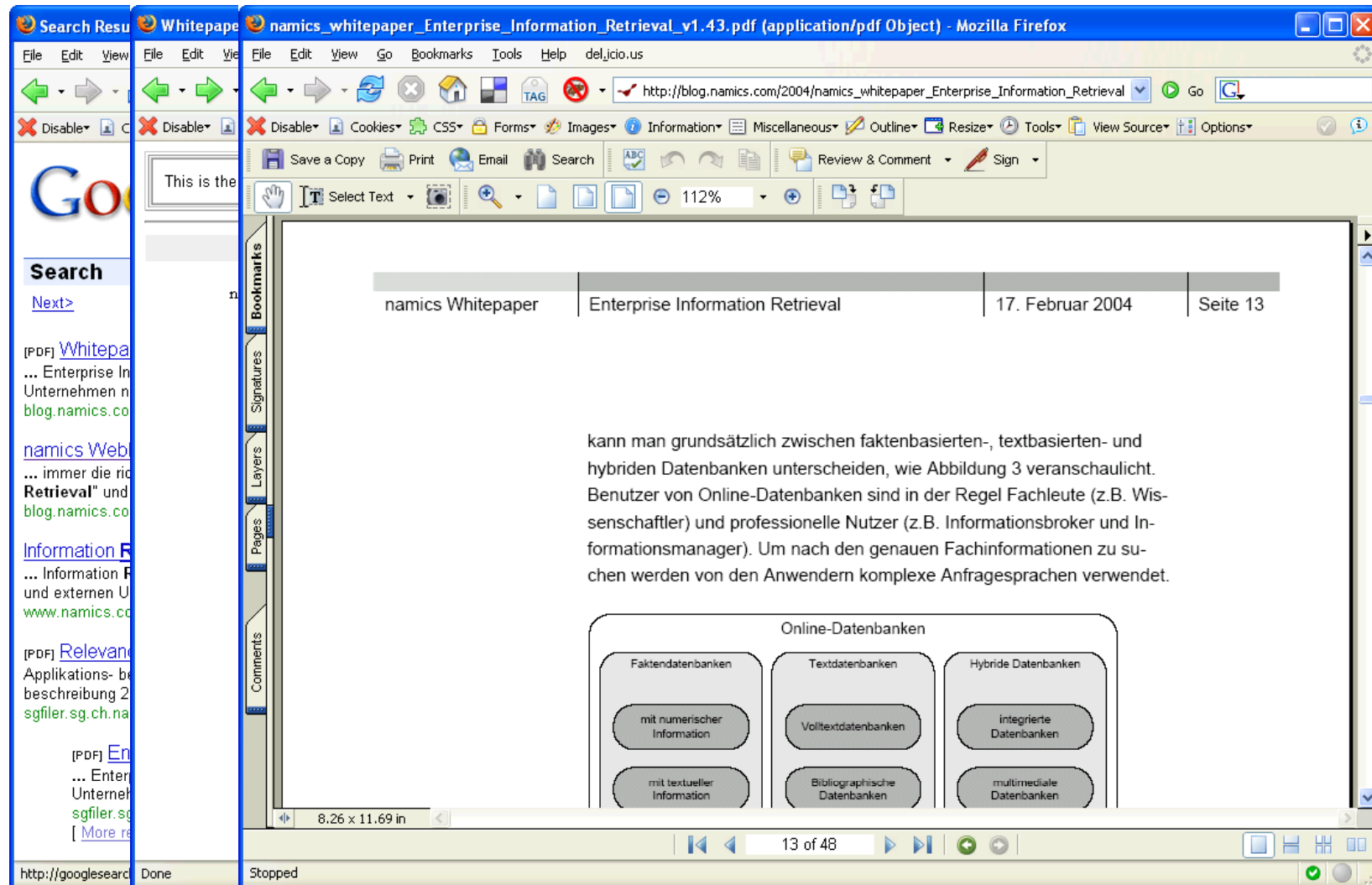
Zugänglichkeit: Tipps

- » Aus der Vollständigkeitsanalyse wissen Sie, welche Seiten nicht besucht werden → korrigieren
- » Browsern Sie Ihr Angebot mit einem Textbrowser (Lynx) und vergleichen Sie mit dem visuellen Angebot
 - Suchmaschine == Lynx
- » Achten Sie v.a. auf Unterschiede bei Cookies (z.B. Login), JavaScript (z.B. AJAX) und Flash
- » Zudem: Login-Seiten und Formulare
- » Crawler können Queryparameter mitnehmen (machen es aber häufig nicht wegen der Gefahr einer zirkulären Referenz)
 - nicht so: <http://www.internet-briefing.ch/index.cfm?page=110909&cfid=4738768&cftoken=78506571>
 - aber so: <http://www.internet-briefing.ch/110909/index.cfm>
 - Oder besser: <http://www.internet-briefing.ch/The-Best-of-Internet>

Interpretierbarkeit

- » Grundregel: Suchmaschinen interpretieren NUR (X)HTML-Code (und am liebsten validen Code)
- » Wenn Ihnen etwas anderes erzählt wird, glauben Sie nur der Grundregel
- » Ausnahmen (je nach Suchmaschine)
 - PDF (wird nach HTML konvertiert)
 - Office: Powerpoint, Word, Excel (wird nach HTML konvertiert)
 - ...
 - Macromedia Flash

Interpretierbarkeit: Von binären Formaten (indexiert)



The screenshot shows a Mozilla Firefox browser window displaying a PDF document. The address bar shows the URL: http://blog.namics.com/2004/namics_whitepaper_Enterprise_Information_Retrieval. The document content includes a header table and a diagram of online databases.

namics Whitepaper	Enterprise Information Retrieval	17. Februar 2004	Seite 13
-------------------	----------------------------------	------------------	----------

kann man grundsätzlich zwischen faktenbasierten-, textbasierten- und hybriden Datenbanken unterscheiden, wie Abbildung 3 veranschaulicht. Benutzer von Online-Datenbanken sind in der Regel Fachleute (z.B. Wissenschaftler) und professionelle Nutzer (z.B. Informationsbroker und Informationsmanager). Um nach den genauen Fachinformationen zu suchen werden von den Anwendern komplexe Anfragesprachen verwendet.

Online-Datenbanken

<p>Faktdatenbanken</p> <ul style="list-style-type: none"> mit numerischer Information mit textueller Information 	<p>Textdatenbanken</p> <ul style="list-style-type: none"> Volltextdatenbanken Bibliographische Datenbanken 	<p>Hybride Datenbanken</p> <ul style="list-style-type: none"> integrierte Datenbanken multimediale Datenbanken
--	--	--

Interpretierbarkeit: Von binären Formaten (nicht indexiert)

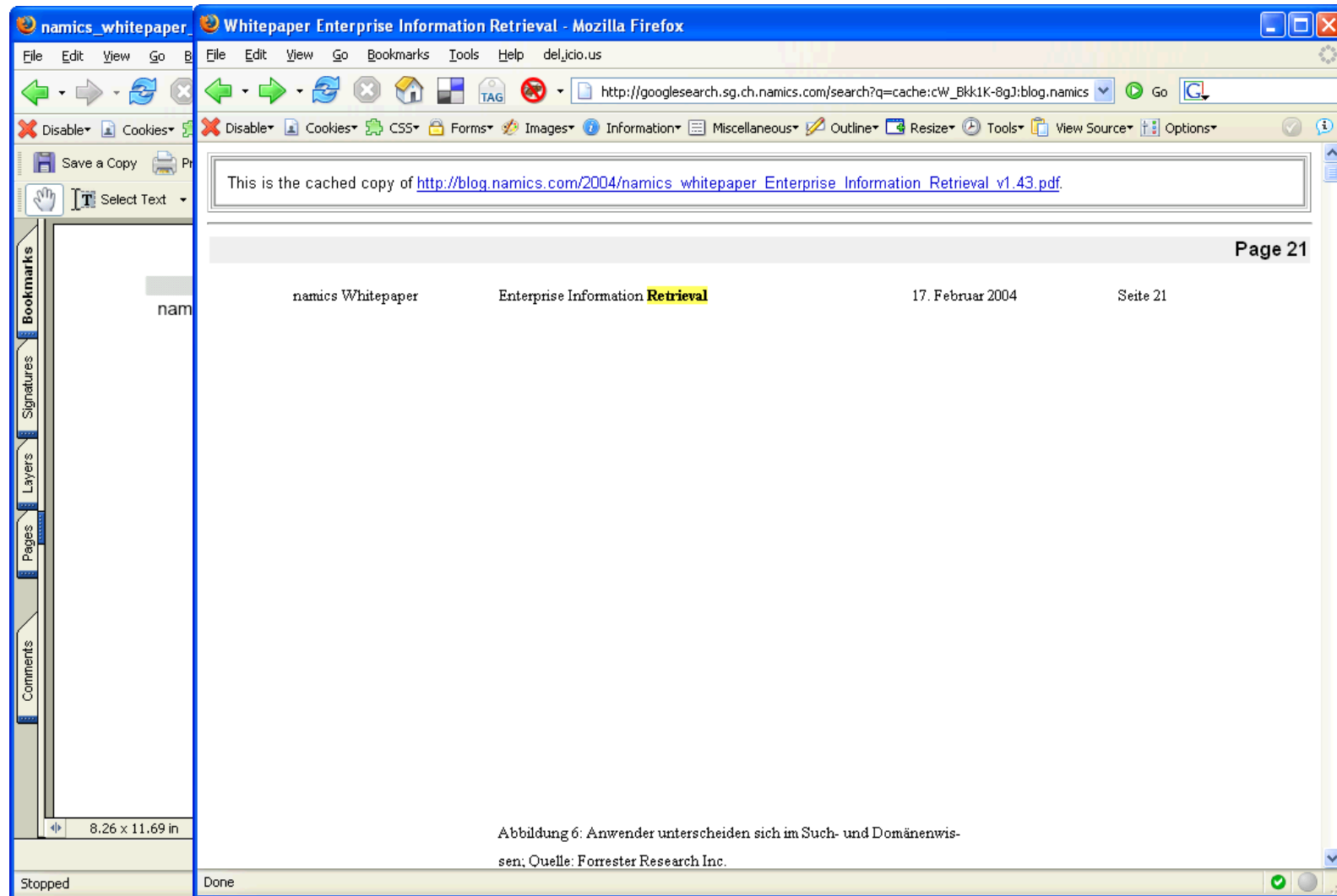


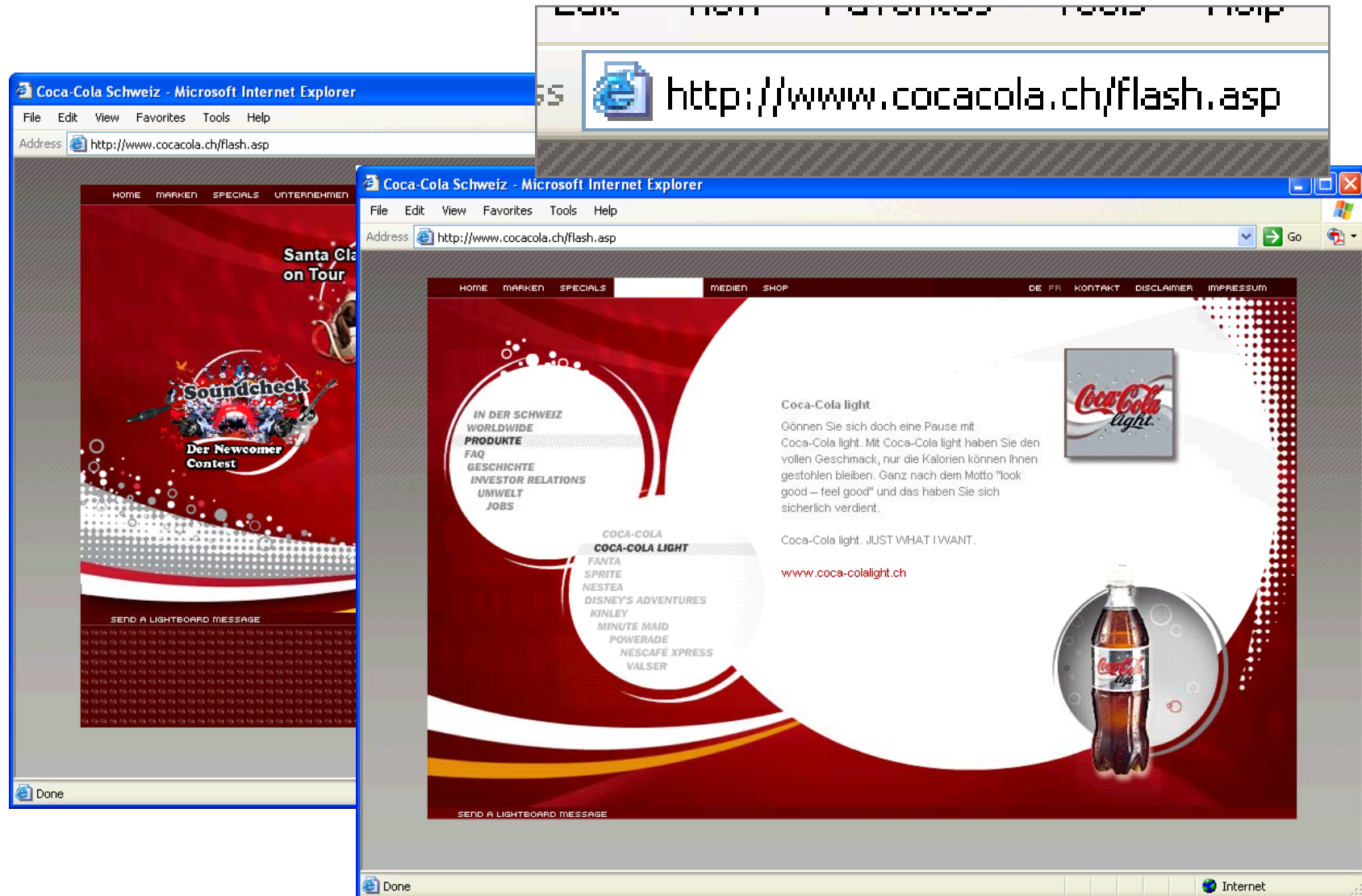
Abbildung 6: Anwender unterscheiden sich im Such- und Domänenwissen; Quelle: Forrester Research Inc.

Interpretierbarkeit: Tipps

- » Bieten Sie alle Inhalte, welche Sie in den Suchmaschinen finden wollen in (X)HTML an
- » Sind andere Formate nötig, so testen Sie, ob die “gewünschten” Suchmaschinen diese verarbeiten
- » Finden Sie im Zitat einer Trefferliste einen Text, der nicht im HTML-Body vorkommt, so ist dies
 - META Description
 - Open Directory Project (<http://www.dmoz.org>)



Referenzierbarkeit: Problem?



Referenzierbarkeit

» Problem

- für die Suchmaschine ist die URI die eindeutige Referenzierung einer Seite
- oder: Pro einzelne URI ist nur ein Element im Index der Suchmaschine gespeichert

» Grundanforderung des WWW:

<http://www.w3.org/2001/tag/doc/whenToUseGet.html>

- *Assign distinct URIs to distinct resources*
- *A URI owner SHOULD provide representations of the identified resource consistently and predictably*
- Bonus: Die verlustfreie Übertragung einer URI via Telefon ist ein guter Test ;-)

Referenzierbarkeit – Zielsetzung

- » Zielsetzung (was muss funktionieren)
 - Bookmarkfunktion (aus Browser-Menu!)
 - Seite per Mail verschicken
 - Verlinkbarkeit von extern

- » Die komplette URL adressiert immer einen wiederherstellbaren Zustand der Seite
 - keine Session ID
 - keine Zufallszahl o.ä.

Referenzierbarkeit: Tipps

- » “Da führt kein Weg dran vorbei”
 - jede Seite im Index der Suchmaschine muss eine eindeutige URL haben
 - die URL ist immer gültig (keine Timeouts)
- » Diese URL sollte immer kommuniziert werden mit dem Ziel möglichst viele externe Links zu erhalten



Nun noch der letzte Trick...

- » Wenn das Genannte schwierig umzusetzen ist, so gibt es bei vielen Suchmaschinen (inkl. Google) die Möglichkeit sein Angebot aktiv zu übermitteln
 - <http://www.sitemaps.org/>
- » Vorgehen
 1. erstellen einer XML-Datei mit den zu indexierenden URLs
 2. ablegen auf dem Webserver
→ Suchmaschine benachrichtigen
 3. übermitteln an die Suchmaschine (Upload)
- » Ist integriert in Google Webmaster Central und Yahoo! Site Explorer

```

<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.com/</loc>
    <lastmod>2005-01-01</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=12&desc=vacation_hawaii</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=73&desc=vacation_new_zealand</loc>
    <lastmod>2004-12-23</lastmod>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=24&desc=vacation_south_island</loc>

```

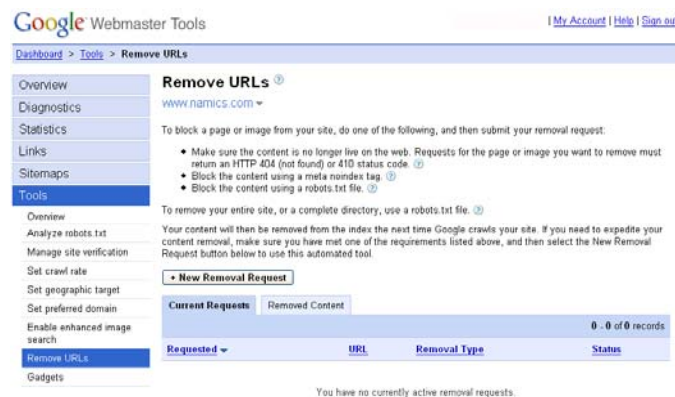

... und die Frage, wie entferne ich eine Seite aus dem Index

» Langfristig entfernen

- Seite mit Passwort schützen
- mit robots.txt oder „META INDEX“ ausschliessen
- auf Abruf einer ungewollten URL mit http 404 (oder http 410) antworten

» Kurzfristig entfernen

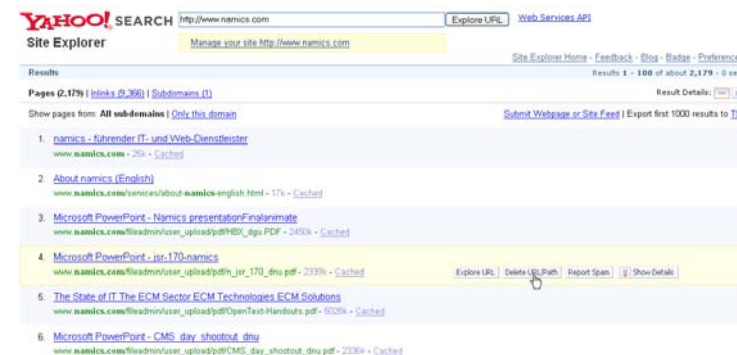
- In Google Webmaster Central und Yahoo! Site Explorer



The screenshot shows the 'Remove URLs' section of Google Webmaster Tools for the domain www.namics.com. It provides instructions on how to block content from the index and includes a table for 'Current Requests'.

Requested	URL	Removal Type	Status
0 - 0 of 0 records			

Below the table, it states: "You have no currently active removal requests."



The screenshot shows the Yahoo! Site Explorer interface for the domain http://www.namics.com. It displays a list of search results for the site, including various pages and documents.

Rank	URL	Cache Status	Actions
1.	namics - fitrender IT- und Web-Dienstleister	www.namics.com - 254 - Cached	
2.	About namics (English)	www.namics.com/services/about-namics-english.html - 174 - Cached	
3.	Microsoft PowerPoint - Namics presentationFinalimagine	www.namics.com/fileadmin/user_upload/pdf/MS_PDF - 24524 - Cached	
4.	Microsoft PowerPoint - ist:170-namics	www.namics.com/fileadmin/user_upload/pdf/ist_170_dnu.pdf - 22324 - Cached	Explore URL Delete URL Report Spam Show Details
5.	The State of IT The ECM Sector ECM Technologies ECM Solutions	www.namics.com/fileadmin/user_upload/pdf/OpenText-Handouts.pdf - 50224 - Cached	
6.	Microsoft PowerPoint - CMS day_shootout_dnu	www.namics.com/fileadmin/user_upload/pdf/CMS_day_shootout_dnu.pdf - 22324 - Cached	

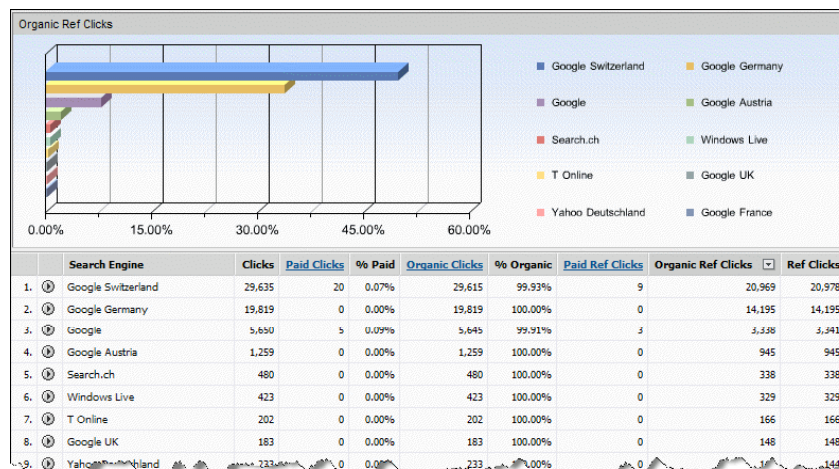
» http://blog.namics.com/2006/03/wie_verschwinde.html

Betrieb



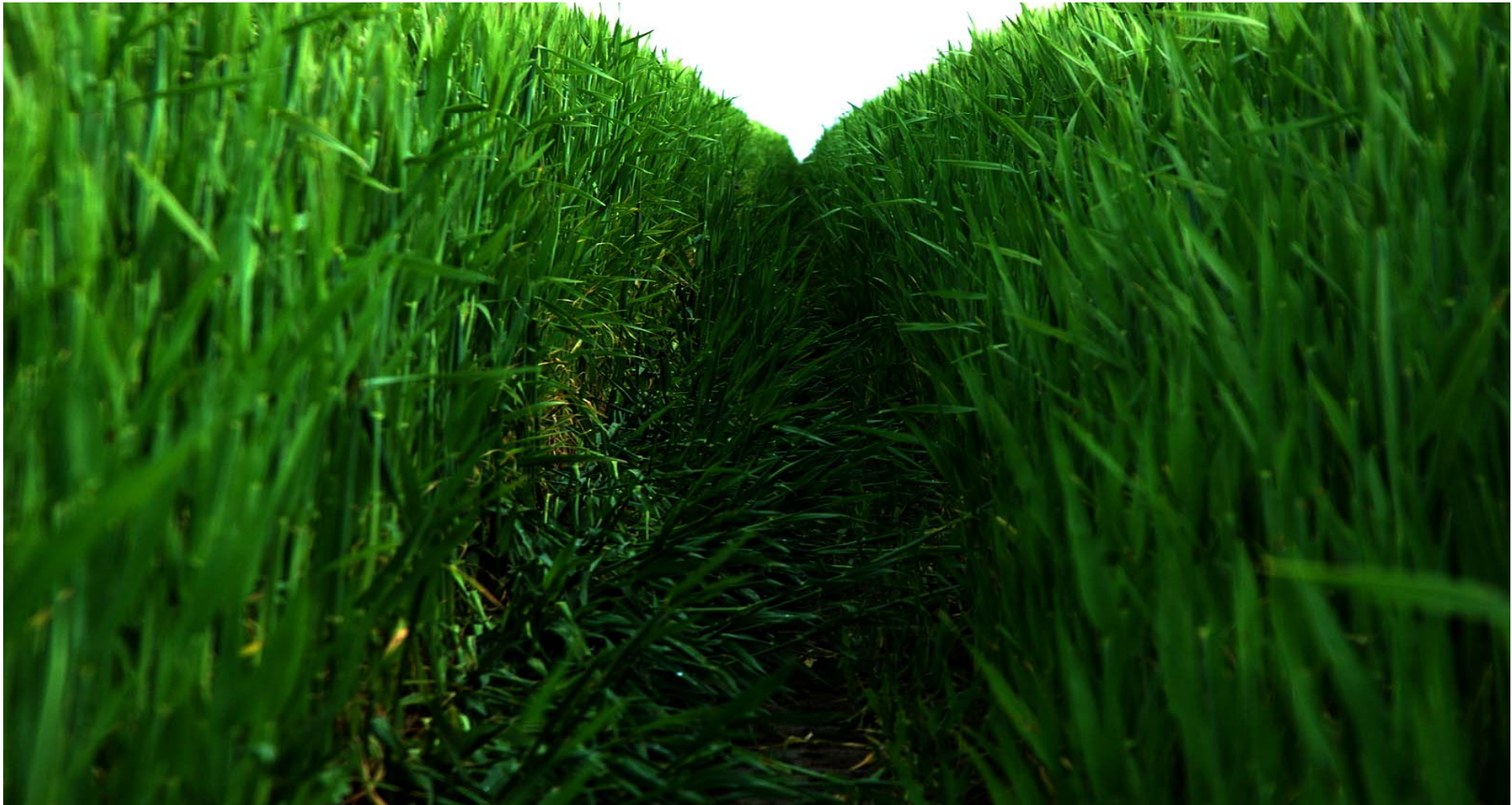
“Alles ist im Fluss”

- » Ihr Angebot und die Suchmaschinen verändern sich
- » Messen Sie den Erfolg Ihrer Platzierung in Suchmaschinen
 - Vollständigkeitsanalyse
 - Referrer von Suchmaschinen: Von welcher Seite auf welche Seite



- » Alle 3 Monate mal anschauen...

Zusammenfassung



Checkliste: Jede Seite, die ich in der Suchmaschine will

1. hat eindeutige URI
2. lässt sich bookmarken (und wieder aufrufen)
3. ist mit `<a href=„...“` in einer HTML-Seite verlinkt
4. lässt sich in Lynx ohne Cookies bedienen
5. sieht in Lynx ohne Java Script gut aus
6. besteht aus validem HTML
7. hat wichtige Keywords häufig und auffällig platziert
8. „?“ , „&“ , „\$“ , „=„ , „+“ , „%“ in der URL -> Vermeiden / Verstecken
9. robots.txt und „META INDEX“ im Griff

Pflicht

Kür

Tools

- » Regeln von Suchmaschinenbetreibern
 - <http://www.google.com/support/webmasters/bin/answer.py?answer=35769>
 - <http://webmaster.search.ch/>
- » Lynx (Textbrowser)
 - <http://lynx.browser.org/>
- » Keine (faulen Tricks)
 - <http://www.google.com/webmasters/seo.html>
- » Informationen zu Websuche und SEO+SEM
 - <http://blog.namics.com/seosem/>
 - <http://www.kso.co.uk/de/tutorial/>
 - <http://searchenginewatch.com/>
 - <http://www.searchengineshowdown.com/>
 - <http://www.searchenginejournal.com/>

Besten Dank für Ihre Aufmerksamkeit.

namics



bernd.langkau@namics.com
<http://blog.namics.com/seosem/>

Eckdaten zu namics

- » Marktführender Schweizer Berater für Online-Anwendungen und E-Business, Präsenz in Deutschland
- » Fokus
 - Strategieberatung für Internet
 - Konzeption und Implementierung nutzergerechter, effizienter und begeisternder Internet-Anwendungen
 - Vermarktung und Lancierung von Online Aktivitäten
 - werblich orientierte Markenkommunikation
- » Zahlen und Fakten
 - gegründet 1995 als Spinoff der Universität St. Gallen
 - 240 Mitarbeiter, Umsatz 2007 CHF 34,2 Mio.
 - Standorte: Bern, Frankfurt, Hamburg, München, St. Gallen, Zug, Zürich

