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

namics

Bern, 12. März 2008

Dr. Bernd Langkau, Partner

Menuvorschlag

	Thema	Stichworte
1	Ein Tag im Leben einer Suchmaschine...	• Abgrenzung → Der Beitrag behandelt nur einen Teil davon • Grundlegende Begriffe
2	Definieren und Feststellen der Vollständigkeit	• Wie ist meine Ausgangslage heute? • Sind alle Seiten das Ziel?
3	Technische Anpassungen	• Typische Probleme und Stolpersteine für Suchmaschinen • Bewertung resp. bessere Produktwahl
4	Betrieb	• Überwachen und Erhalten des erreichten Status

Ein Tag im Leben einer Suchmaschine...

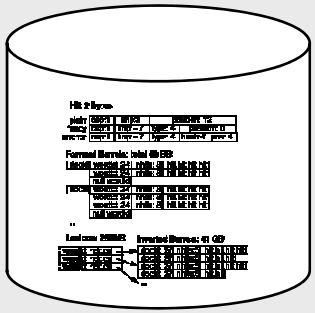
Vier Hauptaufgaben einer Suchfunktion

1. Daten

Akquisition und Speicherung aller Dokumente der Kollektion

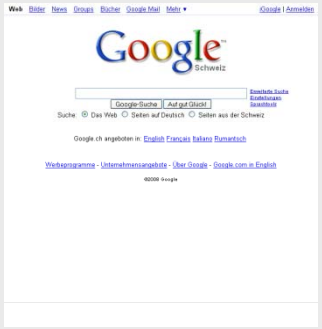
2. Index

Erstellen einer effizienten Datenstruktur für die Suche



3. Such-anfrage

Finden passender Dokumente zu einer Benutzeranfrage



4. Resultat-liste

Präsentation der Treffer in der richtigen Reihenfolge

1. Daten

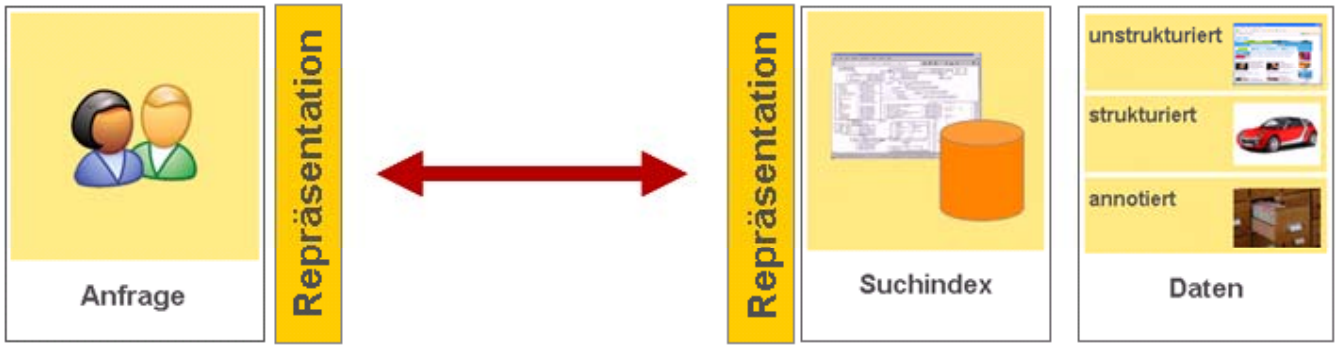
- “Daten” sind im Kontext einer Websuche
  - Webseiten = [X]HTML Base Pages
  - darin referenzierte, binäre Dokumente (PDF, gif)
- Ein einzelnes Element muss eineindeutig durch eine URL referenzierbar sein
  - http://blog.namics.com/seosem/
  - http://blog.namics.com/images/namics_logo.gif
- Die Suchmaschine 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

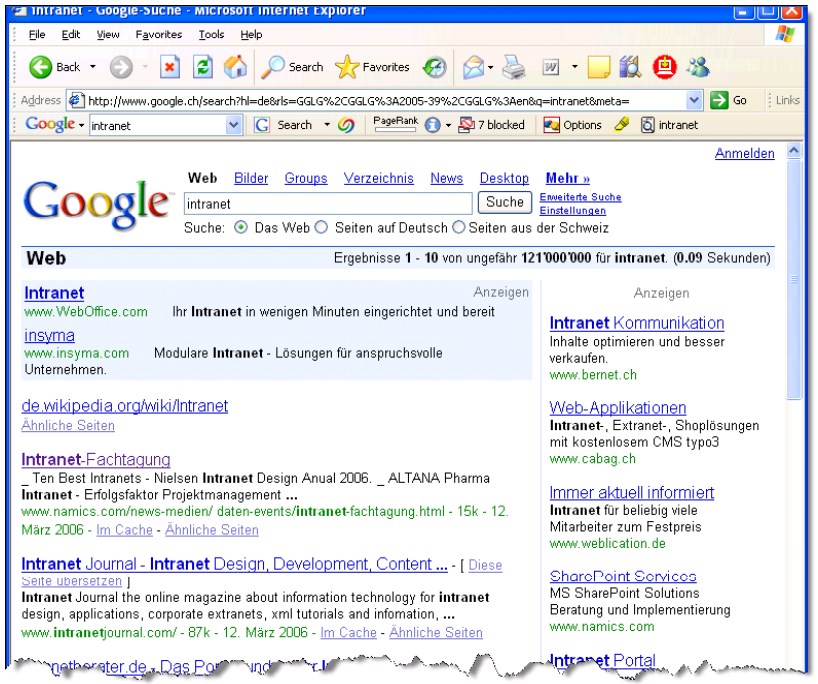
- Die Anfrage (=Query) ist die textuelle Formulierung eines Informationsbedürfnisses durch einen User



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

4. Resultatliste

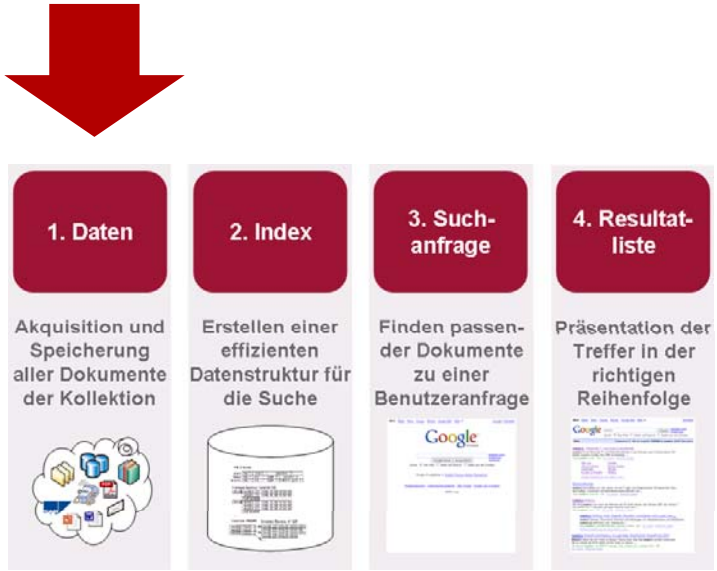
- “Heiliger Gral” der Suchmaschine
  - 121 Mio. Treffer; wer steht auf Platz 1, 2 und 3?
- Viele Unwahrheiten
- Viele unseriöse Anbieter
- Google-Regeln
  - http://www.google.com/support/webmasters/bin/answer.py?answer=35769
  - http://www.google.com/support/webmasters/bin/answer.py?answer=35291
- Suchmaschinenmarketing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

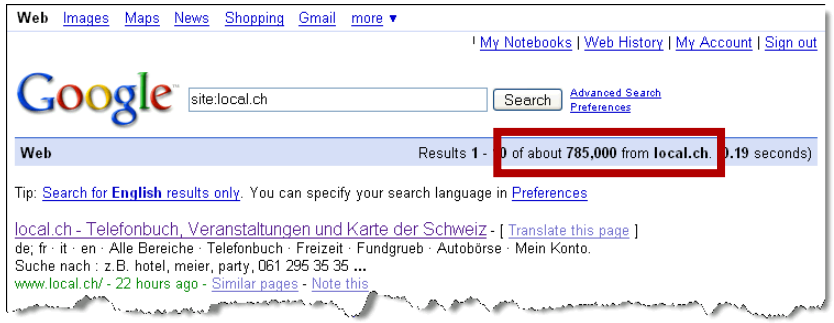
Vollständigkeitsanalyse

- 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“

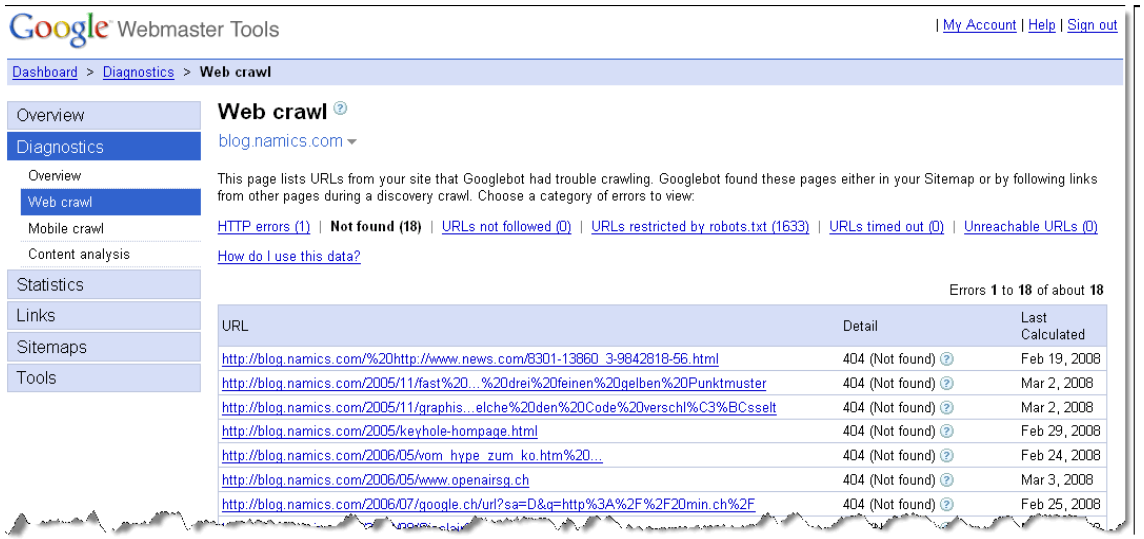
- oder das?



- Query (fast überall): *site:Domänenname*
- 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)



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

- 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

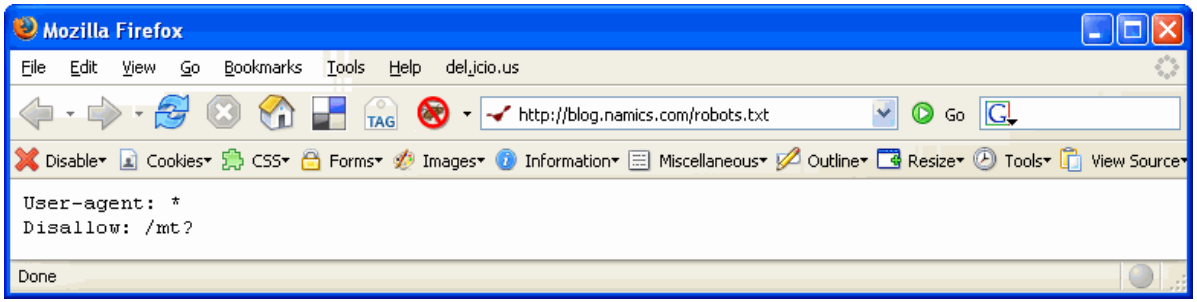
	Browser	Hits	% von Gesamt	Sitzungen
1	Googlebot	12,201	9.04%	4,417
2	FAST WebCrawler	11,807	8.82%	1,807
3	Cosmos	3,366	2.49%	698
9	KIT-Fireball	304	0.22%	114
10	search.ch V1.4.2 (spiderman@search.ch; http:	70,157	52.02%	92
11	Scooter-W3.1.2	7,164	5.31%	80

Selektivität (Konzentration der Kräfte + „Müll“)

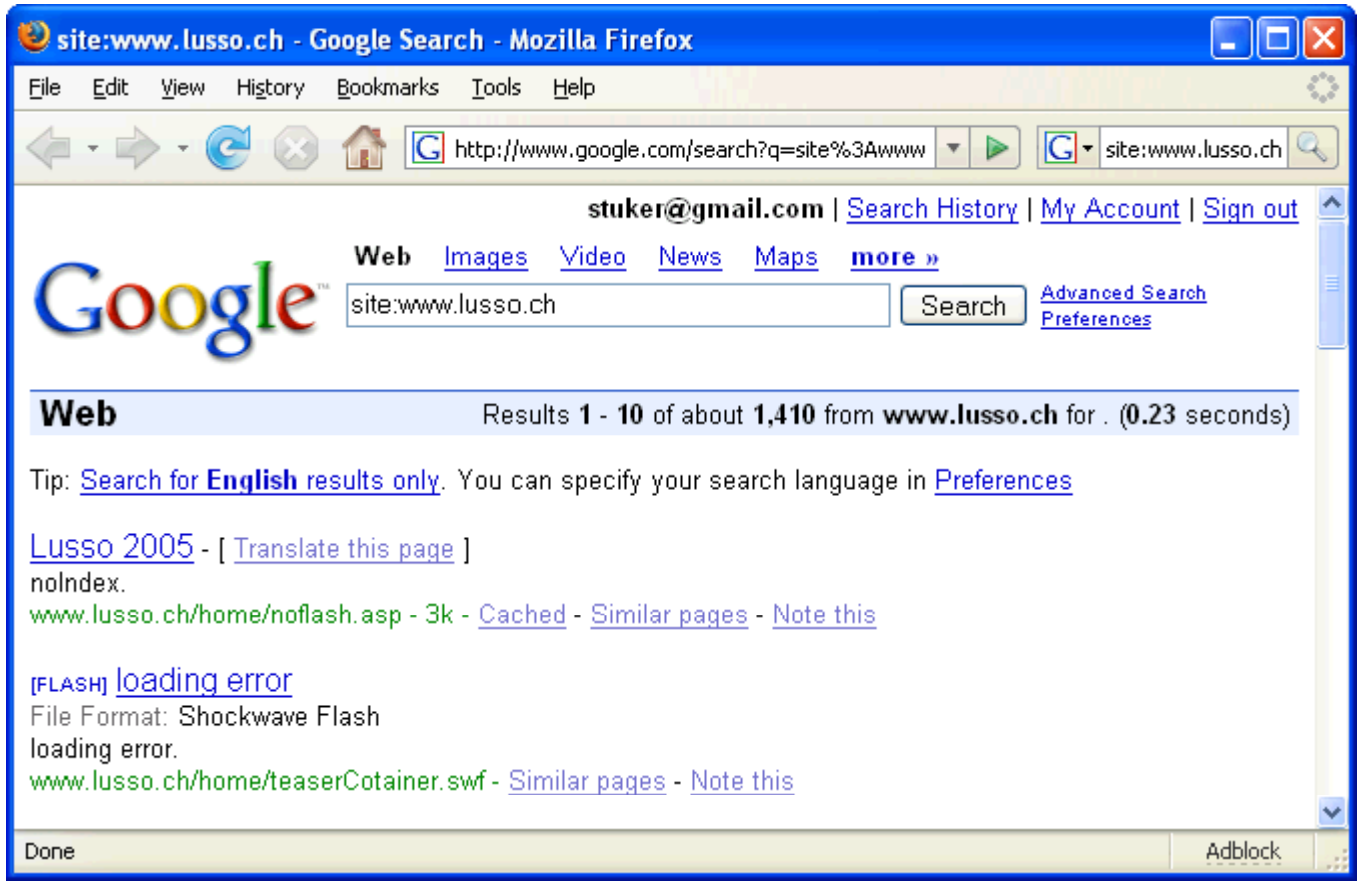
- Sie wollen nicht alle Seiten in der Suchmaschine!
  - z.B. Login für Autoren, personalisierte Ansichten etc.
- Zwei Möglichkeiten des Ausschlusses
  - Kennzeichnung im HTML (clientseitig)
    (http://www.robotstxt.org/wc/meta-user.html)

```
<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

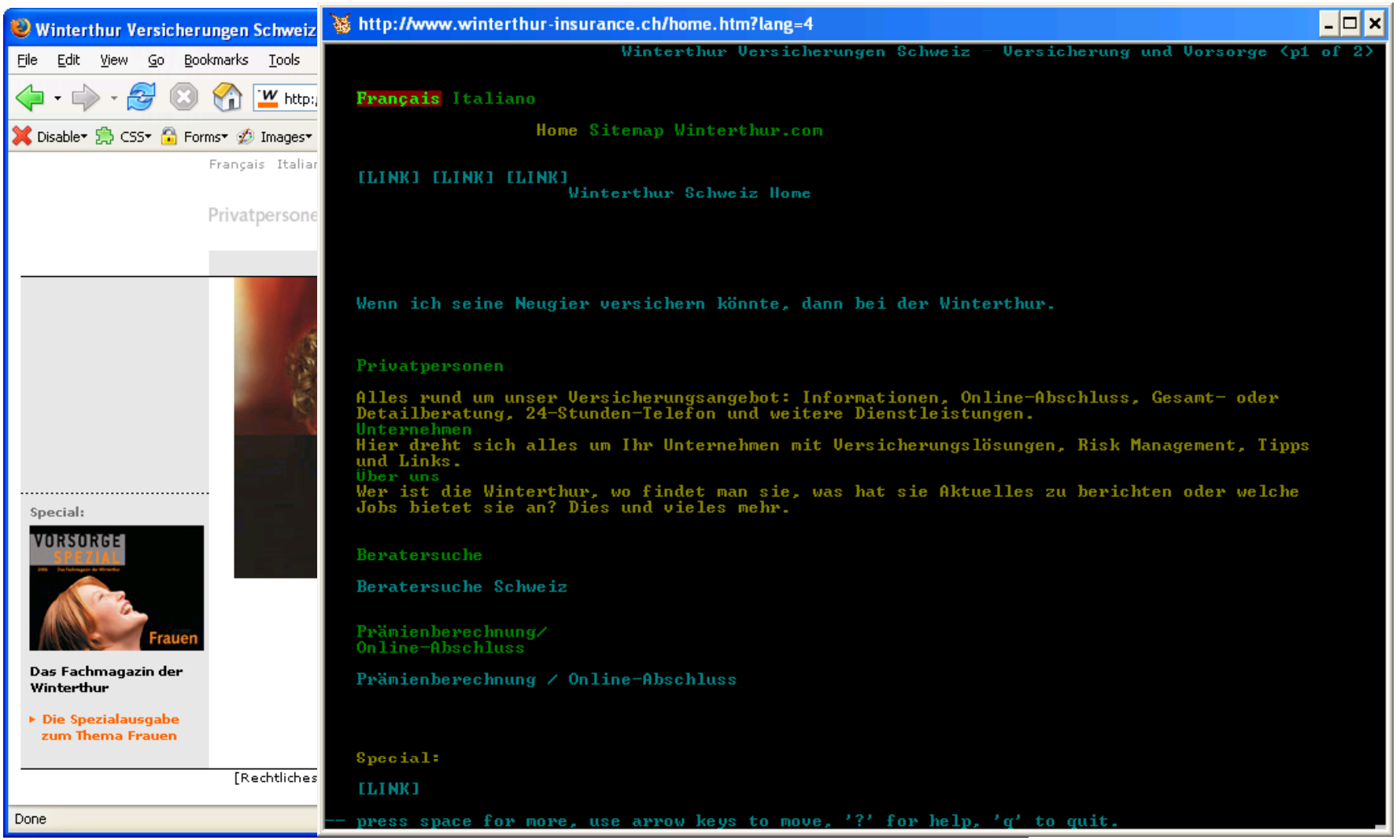
- Zugänglichkeit
- Interpretierbarkeit
- Referenzierbarkeit

Zugänglichkeit für Crawler

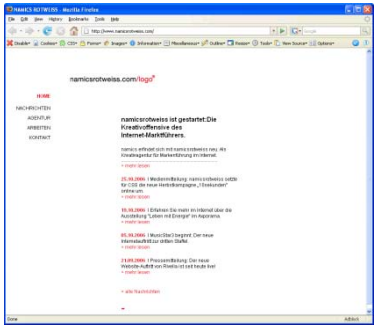
- Crawler / Spider / Gatherer / Fetch
  - lädt und speichert HTML jeder Seite (Base Page Download)
  - extrahiert alle Links
  - folgt sämtlichen Links rekursiv
- Was machen Crawlers nicht
  - JavaScript ausführen
  - Cookies annehmen
  - zu viele Query-Parameter mitnehmen (Faustregel: 2)
  - zu viele 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

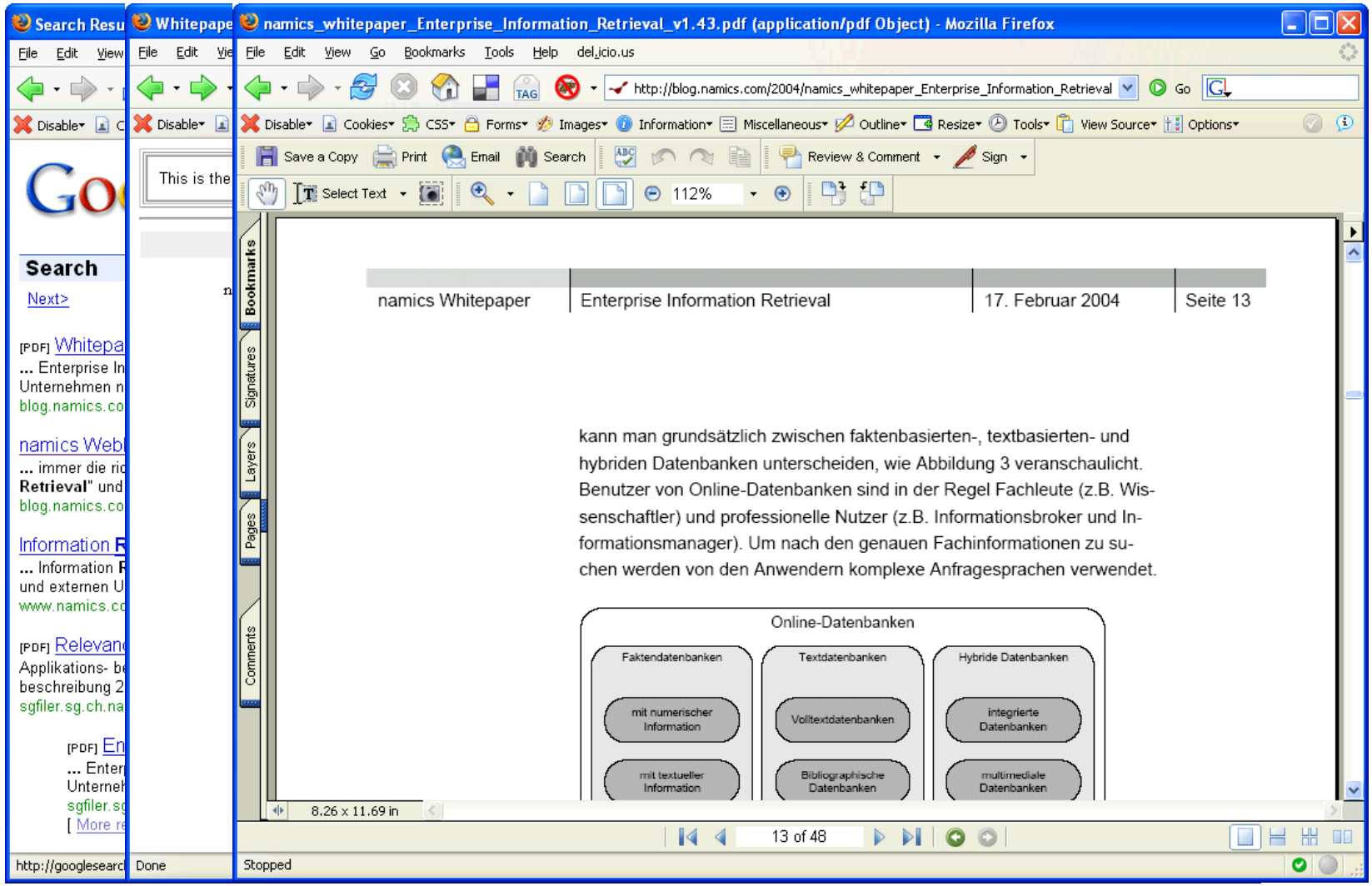
Zugänglichkeit: Tipps

- 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

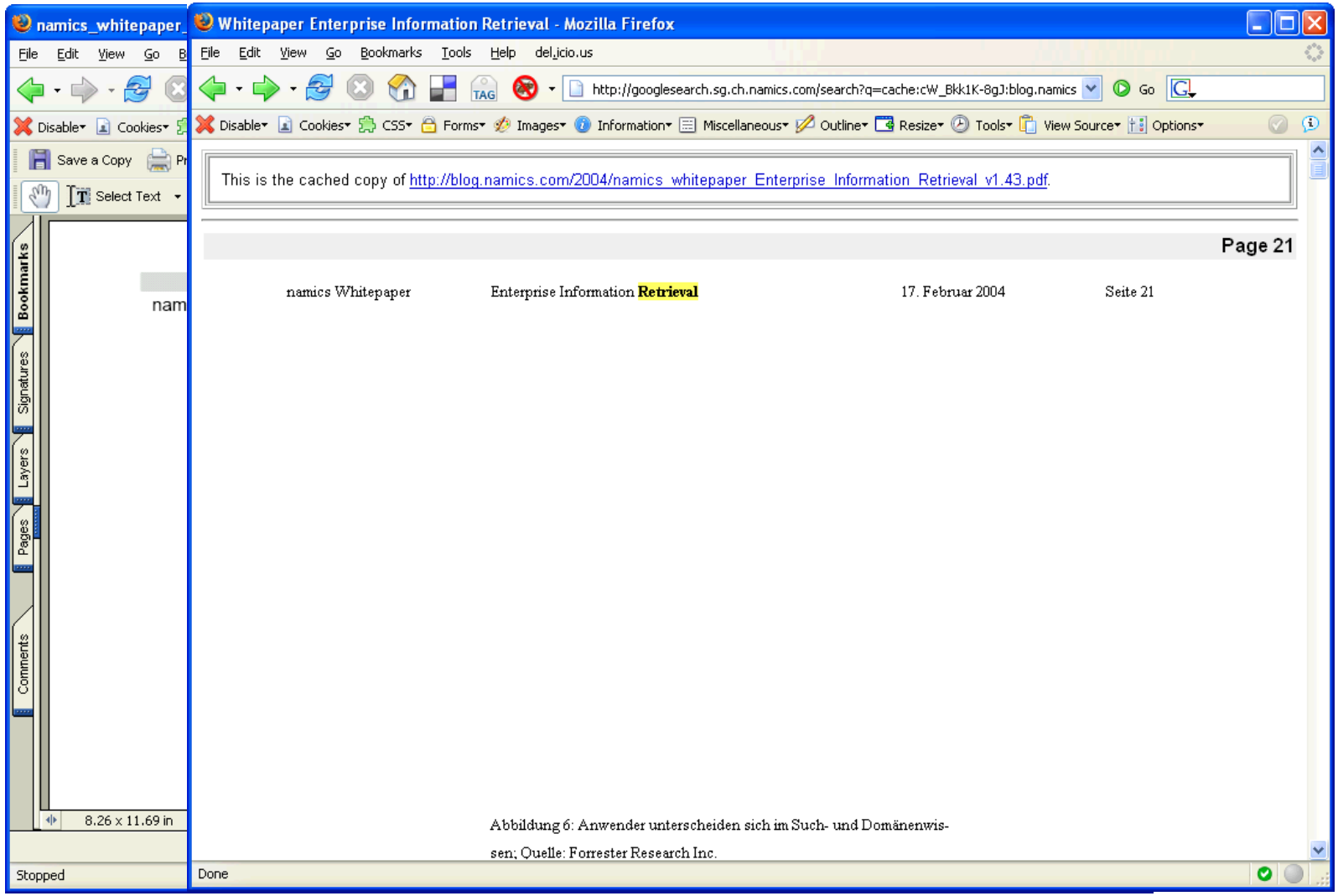
Interpretierbarkeit

- 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)



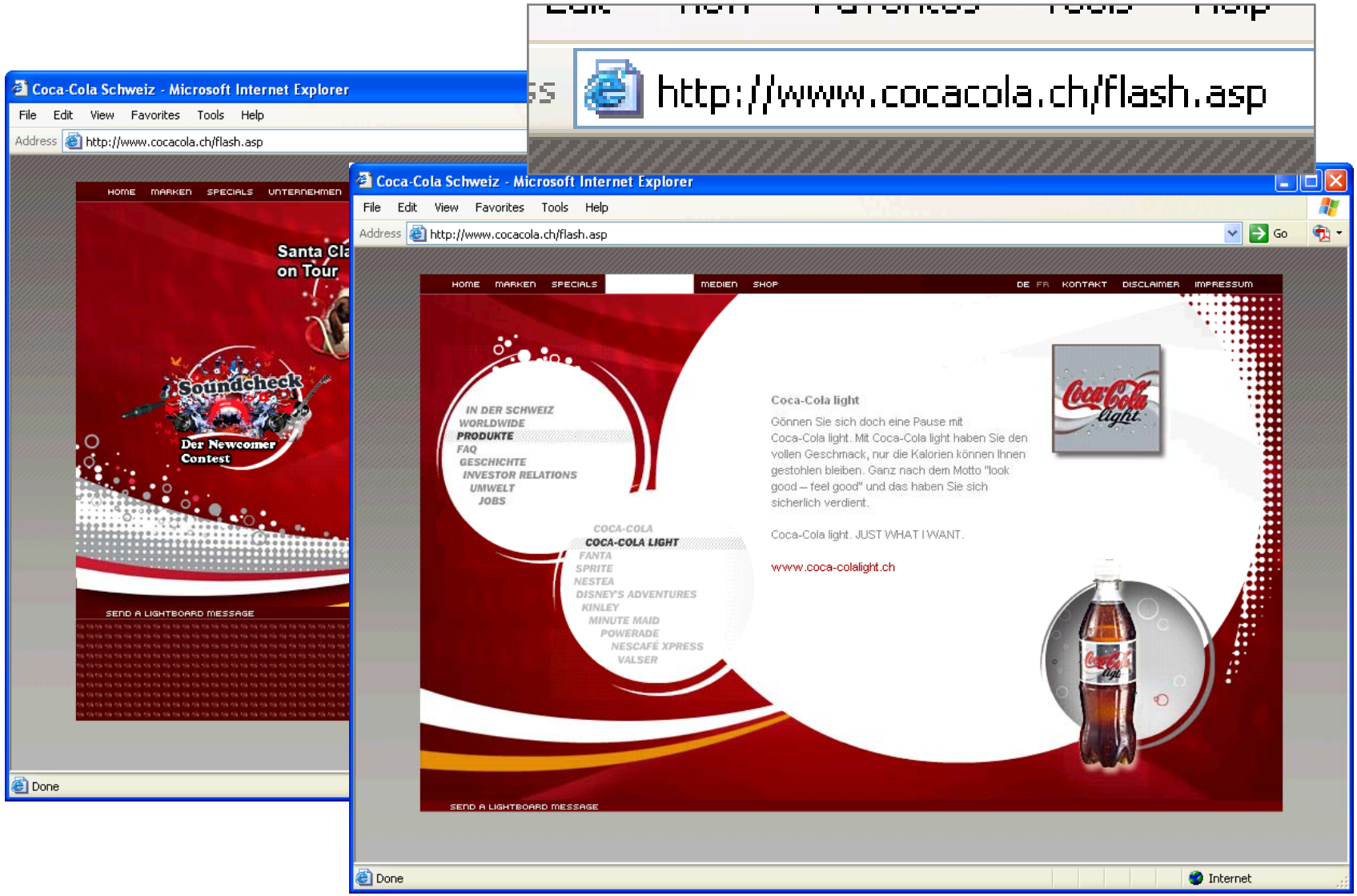
Interpretierbarkeit: Von binären Formaten (nicht indexiert)



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

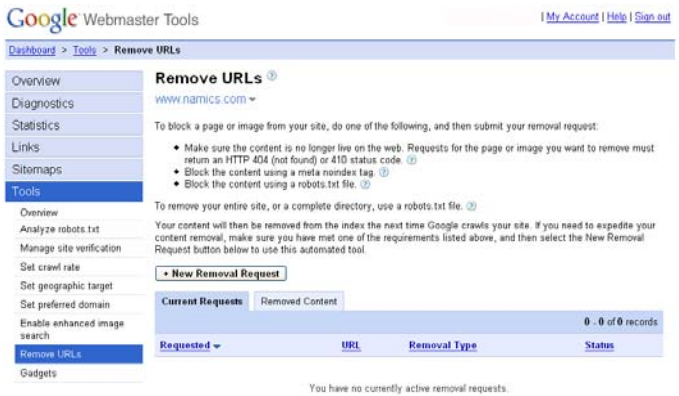
- “Da führt kein Weg dran vorbei”
  - jede Seite im Index der Suchmaschine muss eine eineindeutige URL haben
  - die URL ist immer gültig (keine Timeouts)
- Diese URL sollte immer kommuniziert werden mit dem Ziel möglichste viele externe Links zu erhalten

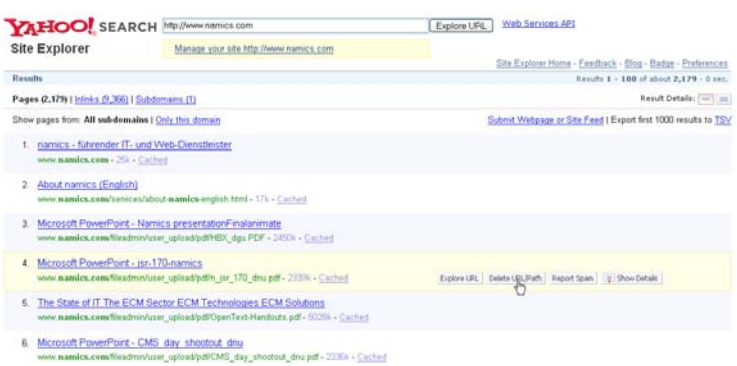
Nun noch der letzte Trick…

- 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 intergriert in Google Webmaster Central und Yahoo! Site Explorer

… 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



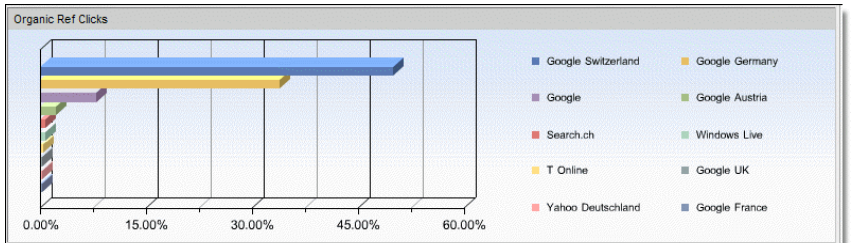


- http://blog.namics.com/2006/03/wie_verschwinde.html

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



		Search Engine	Clicks	Paid Clicks	% Paid	Organic Clicks	% Organic	Paid Ref Clicks	Organic Ref Clicks	Ref Clicks
1.		Google Switzerland	29,635	20	0.07%	29,615	99.93%	9	20,969	20,978
2.		Google Germany	19,819	0	0.00%	19,819	100.00%	0	14,195	14,195
3.		Google	5,650	5	0.09%	5,645	99.91%	3	3,338	3,341
4.		Google Austria	1,259	0	0.00%	1,259	100.00%	0	945	945
5.		Search.ch	480	0	0.00%	480	100.00%	0	338	338
6.		Windows Live	423	0	0.00%	423	100.00%	0	329	329
7.		T Online	202	0	0.00%	202	100.00%	0	166	166
8.		Google UK	183	0	0.00%	183	100.00%	0	148	148
9.		Yahoo Deutschland	233	0	0.00%	233	100.00%	0	144	144

- 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

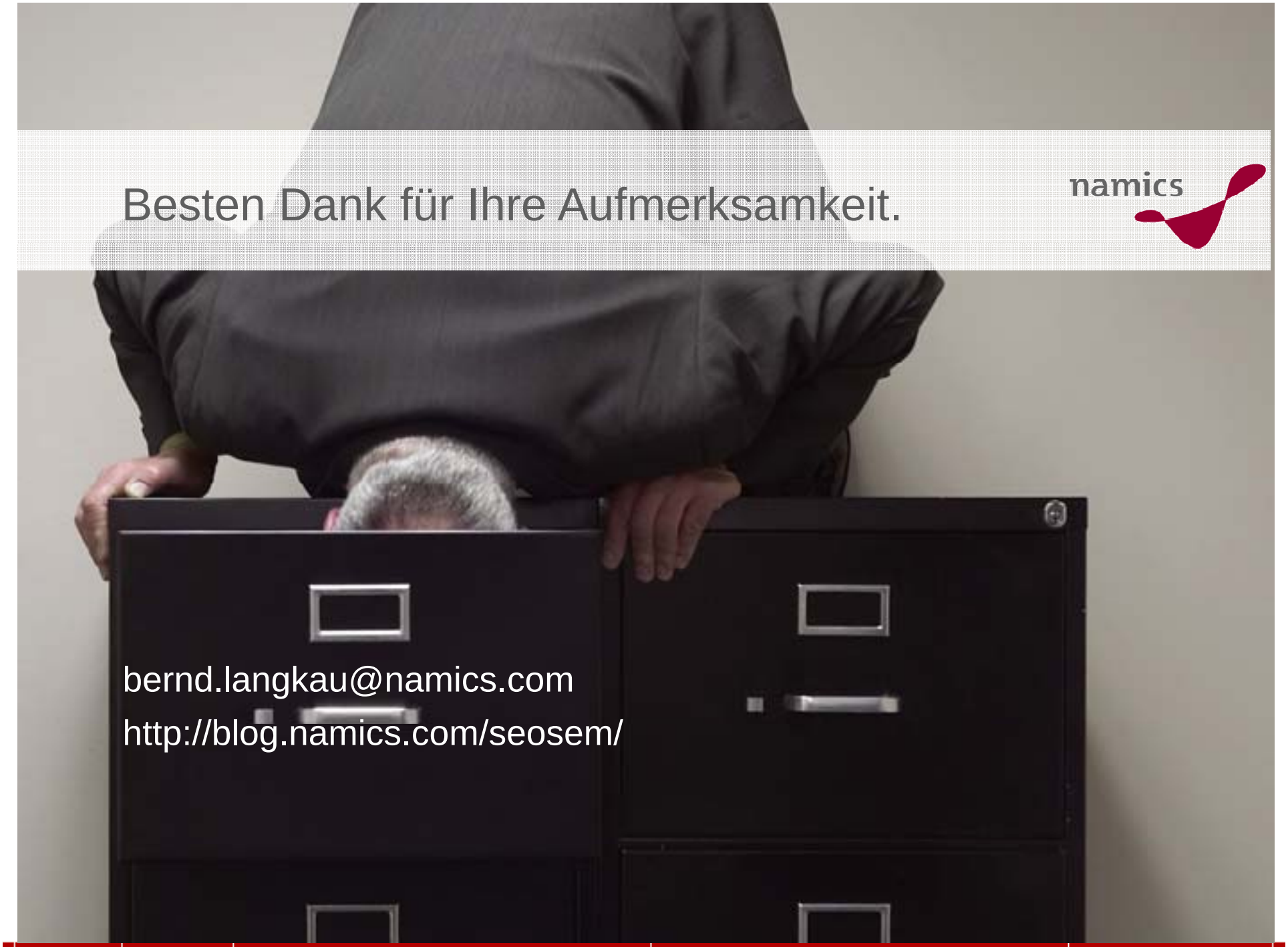
Pflicht

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.



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