DRI2020

Rettskilder og informasjonssøking

Søkemotorer:

Troverdighet og synlighet

Gisle Hannemyr
Ifi, høstsemesteret 2014

Opprinnelsen til fritekstsøk

- Vedtak i Pennsylvania en gang på slutten av 1950-tallet om å bytte ut begrepet "retarded child" i diverse lover med det mer politisk korrekte "special child".
- Uoverkommelig å finne alle forekomster manuellt.
- Deler av lovsamlingen var tastet inn på hullkort. John F. Horty fikk utvided databasen til å omfatte hele lovsamlingen med komemntarer.
- Kilde: John F. Horty, "Experience with the Application of Electronic Data Processing Systems in General Law", *Modern Uses of Logic in Law*, December 1960.

The Internet: The Resource discovery problem

- The existence of digital resources on the Internet led to formulation of "The Resource Discovery Problem".
- First formulated by Alan Emtage and Peter Deutsch in *Archie - an Electronic Directory Service for the Internet*[1] (1992)
  - «Before a user can effectively exploit any of the services offered by the Internet community or access any information provided by such services, that user must be aware of both the existence of the service and the host or hosts on which it is available.»

1) Archie was a search engine into ftp-space that pre-dated any web-oriented search engines.

The Resource discovery problem

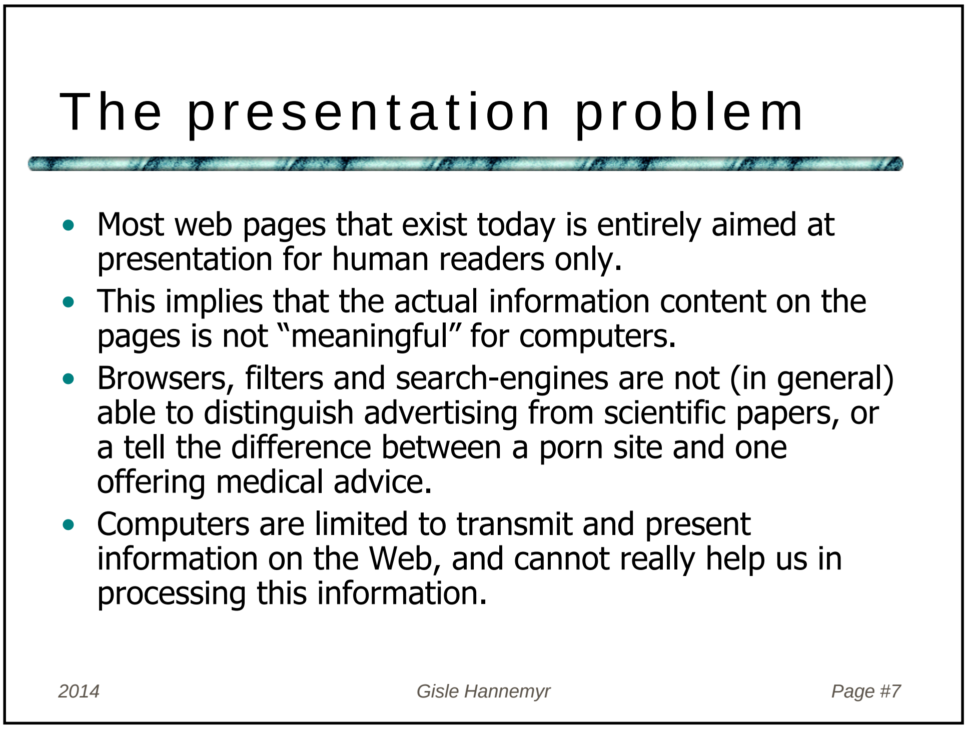
- So the resorce discovery problem encompasses not only to establish the *existence* and *location* of resources, but:
  - If the discovery process yields pointers to several alternative resources, some means to qualify them and to identify the resource or resoures that provides the "best fit" for the problem at hand.
  - Means by the which the user can assess the quality, relevance, topicality, significance and suitability of a given resource.

A Resource According to RFC 2396 (Uniform Resource Identifier: URI)

- A resource is anything that has identity:
  - Familiar examples include an electronic document, an image, a service (e.g., «today's weather report for Los Angeles»), and a collection of other resources.
  - Resources of primary interest are those that are retrievable from the Internet. For those resources the URI is also an unique address that can be used to locate the resource (an Uniform Resource Locator).
  - Not all resources are network «retrievable»; e.g., human beings, corporations, and bound books in a library can also be considered resources.

Uniform Resource Identifier: URI

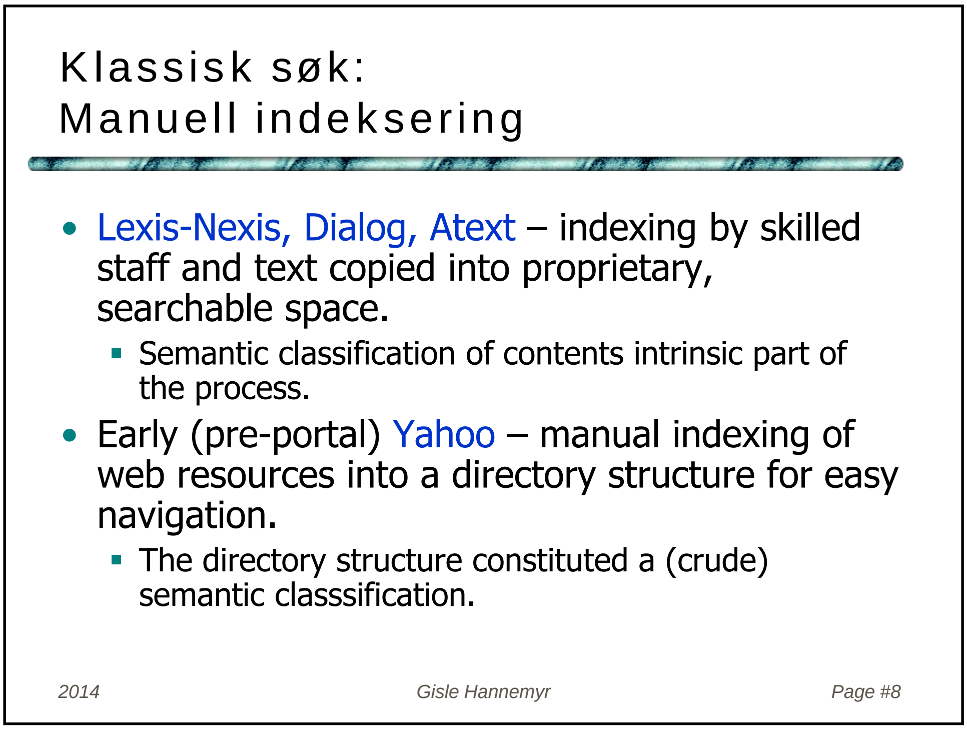
- Much more elusive than a database key, an ISBN-number, or a Dewey identifier:
  - The resource is the conceptual mapping to an entity or set of entities, not necessarily the entity which corresponds to that mapping at any particular instance in time.
  - Thus, a resource can remain constant even when its content - the entities to which it currently corresponds - changes over time, provided that the conceptual mapping is not changed in the process.
- I.e. the URI remains constant even if its meaning changes.



The presentation problem

- Most web pages that exist today is entirely aimed at presentation for human readers only.
- This implies that the actual information content on the pages is not “meaningful” for computers.
- Browsers, filters and search-engines are not (in general) able to distinguish advertising from scientific papers, or a tell the difference between a porn site and one offering medical advice.
- Computers are limited to transmit and present information on the Web, and cannot really help us in processing this information.

2014 *Gisle Hannemyr* *Page #7*



Klassisk søk:
Manuell indeksering

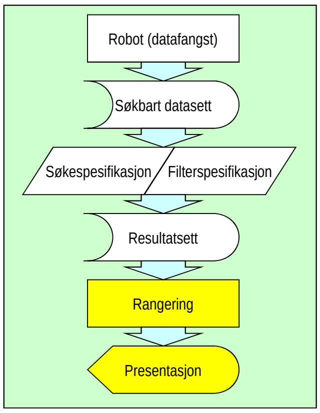
- Lexis-Nexis, Dialog, Atext – indexing by skilled staff and text copied into proprietary, searchable space.
  - Semantic classification of contents intrinsic part of the process.
- Early (pre-portal) Yahoo – manual indexing of web resources into a directory structure for easy navigation.
  - The directory structure constituted a (crude) semantic classsification.

2014 *Gisle Hannemyr* *Page #8*

Internettsøk Robotindeksering

- Med den nåværende størrelsen og vekstraten på verdensveven er menneskelig indeksering av innholdet ikke lenger praktisk mulig.
  - Et eget program, som vanligvis kalles for en «robot» (men som også går under navn som «scooter », «drone», «spider» eller «web crawler») leter rundt på nettet, besøker websider og samler inn data.
  - Når roboten kommer til en ny webside, vil den kopiere alle data som finnes på siden inn i en enorm database som er en del av søkemotoren.
  - Denne databasen gjøres så tilgjengelig for søk gjennom et eller annet brukergrensesnitt og evt. også API/web-services.

Virkemåten til en (ekstern) søkemotor som Google



Figur 1: Anatomien til en typisk Internett søkeportal

Søket etableres gjennom at brukeren angir hva det skal søkes etter (en *søkespesifikasjon*).

Ofte har bruken brukeren muligheten til å avskjære søket gjennom ett eller flere *filtre* som typisk er knyttet til metadata som under datafangsten er syntesert ut fra dataene selv, URLen og/eller HTML-markeringer.

Web Search Engines:

- Differs radically from previous systems in that they:
  - Employ robots/spiders rather than human archivists/cataloguers for data capature.
  - Separation of dataspace and search space.
  - Intially 100% based upon free text search.
  - Core conecpt: URI (Uniform Resource Identifier)

Et problem med fritekst søkemotorer

- Fritekstsøk uten tolking av metadata gir for dårlig kvalifiserte data:
  - Ikke vanskelig å benytte søk til å finne materiale på web *om* Erna Solberg
  - Nærmest umulig å finne materiale der Erna Solberg er *forfatteren.*

Ulike nettbaserte søkemotorer

- Generelle søkemotorer
  - Bing
  - AltaVista
  - Google Web Search
- Metasøkemotorer
  - Ask.com
  - DogPile
  - MetaCrawler
- Mediaorienterte s.m.
  - Google Image Search
  - Google filetype:torrent
  - The Pirate Bay
- Emneorienterte s.m.
  - Google Scholar
  - ISI Web of Science
  - Kulturnettsøk

Pre-internet search vs. Internet search engine usage

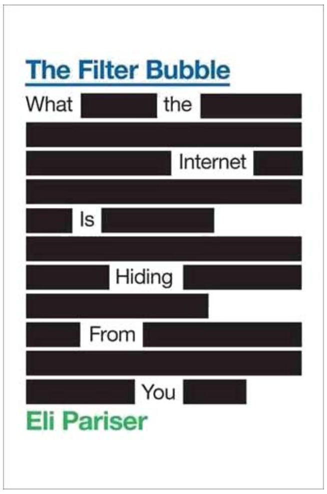
- The classic information services was typically created to cater for the research related needs of professional researchers (e.g. Dialog started its life as an internal service Lockheed aerospace corporation’s library in 1965). When this service became available to external clients in beginning the 1980ies, its typical user was a professional librarian acting on behalf of an academic institution or a paying client using the system for professional research.
- Internet search engines are available at no cost, and the most used search terms are clearly not related to work activities.

What are people looking for on the Internet?

- An analysis of the log of the popular AltaVista search engine conducted in the fall of 1998 yields the following top ten terms: *sex, applet, porno, mp3, chat, warez, yahoo, playboy, xxx, hotmail* (Silverstein et al 1998).
- *Google Insights for Search* (last 12 months, March 2009): *lyrics, you, yahoo, myspace, youtube, games, my, google, facebook, weather.*
- *Google Insights for Search* (last 12 months, March 2011): *facebook en español , facebook español ,4shared , www.facebook.com, face, taringa, twitter, iphone , facebook*

The filter bubble

- I 2005 introduserte Google «personlig søk», der Google benyttet «cookies» til å lage personlige profiler av alle brukere.

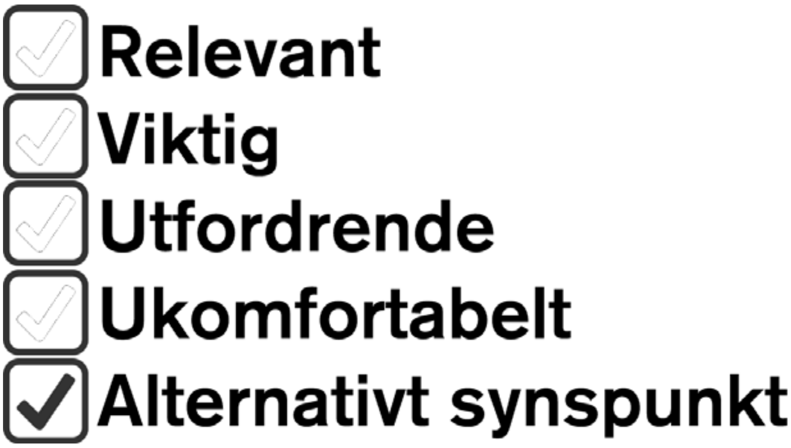


The Filter Bubble

What the Internet Is Hiding From You

Eli Pariser

Et annet søkegrensesnitt



☑ Relevant

☑ Viktig

☑ Utfordrende

☑ Ukomfortabelt

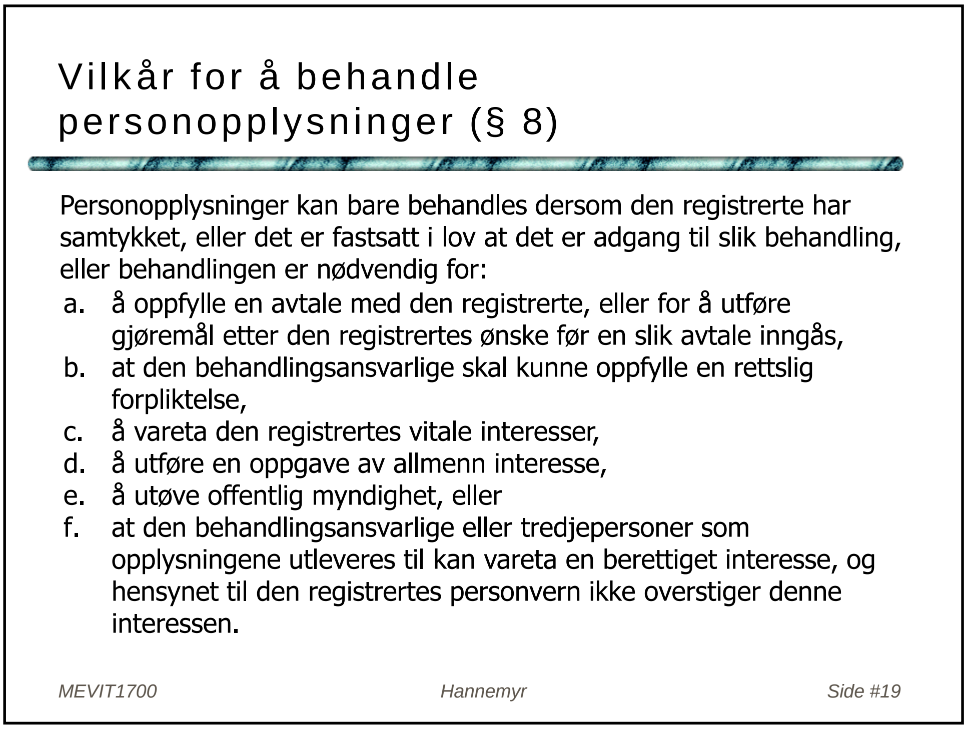
☑ Alternativt synspunkt

2014 Gisle Hannemyr Side #17

Personopplysningsloven

- Når personer indekseres av Google i selskapets søkeindeks er det som skjer en behandling av personopplysninger slik dette er definert i Personopplysningsloven.
- Norsk og europeisk lov krever at det finnes hjemmel for all behandling av personopplysninger.
- Har Google, Yahoo, Facebook, et al noen slik hjemmel?

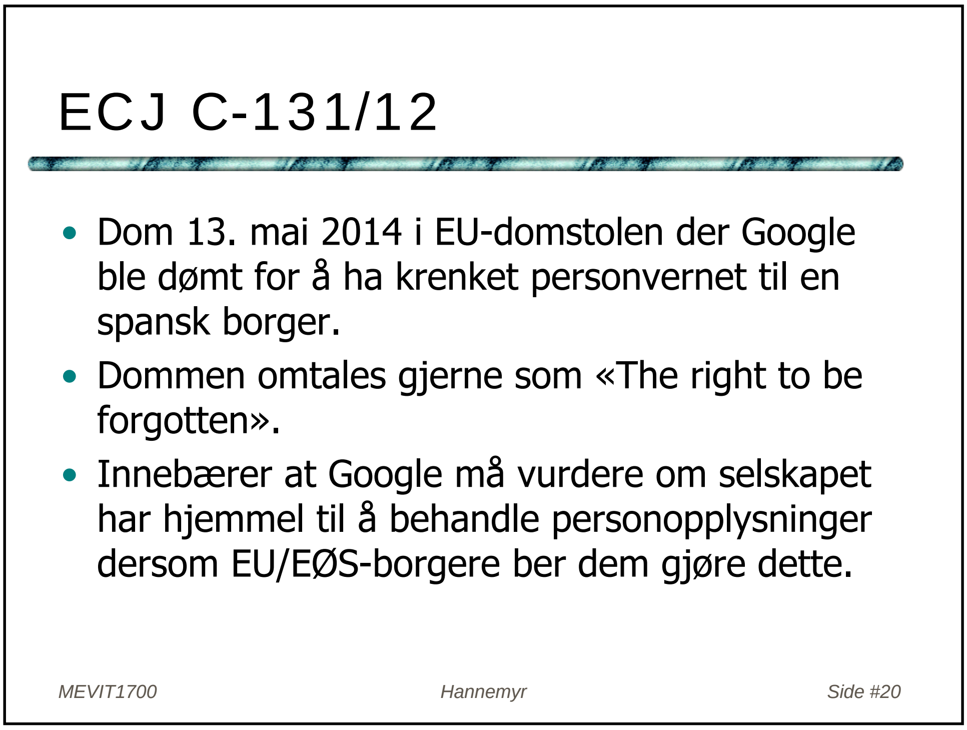
MEVIT1700 Hannemyr Side #18



Vilkår for å behandle personopplysninger (§ 8)

Personopplysninger kan bare behandles dersom den registrerte har samtykket, eller det er fastsatt i lov at det er adgang til slik behandling, eller behandlingen er nødvendig for:

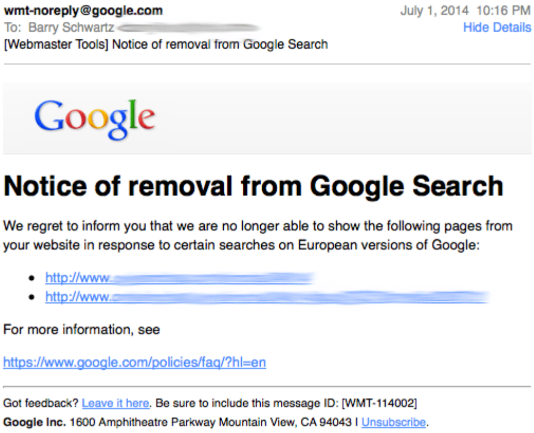
a. å oppfylle en avtale med den registrerte, eller for å utføre gjøremål etter den registrertes ønske før en slik avtale inngås,
b. at den behandlingsansvarlige skal kunne oppfylle en rettslig forpliktelse,
c. å vareta den registrertes vitale interesser,
d. å utføre en oppgave av allmenn interesse,
e. å utøve offentlig myndighet, eller
f. at den behandlingsansvarlige eller tredjepersoner som opplysningene utleveres til kan vareta en berettiget interesse, og hensynet til den registrertes personvern ikke overstiger denne interessen.



ECJ C-131/12

- Dom 13. mai 2014 i EU-domstolen der Google ble dømt for å ha krenket personvernet til en spansk borger.
- Dommen omtales gjerne som «The right to be forgotten».
- Innebærer at Google må vurdere om selskapet har hjemmel til å behandle personopplysninger dersom EU/EØS-borgere ber dem gjøre dette.

Google sender ut slike eposter til alle utgivere ved fjerning



wmt-noreply@google.com July 1, 2014 10:16 PM
To: Barry Schwartz [redacted] Hide Details
[Webmaster Tools] Notice of removal from Google Search

Google

Notice of removal from Google Search

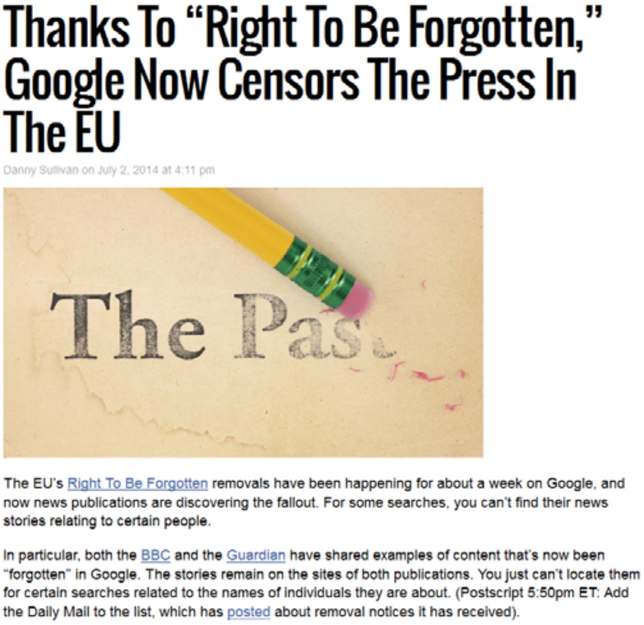
We regret to inform you that we are no longer able to show the following pages from your website in response to certain searches on European versions of Google:

- http://www.[redacted]
- http://www.[redacted]

For more information, see

https://www.google.com/policies/faq/?hl=en

Got feedback? Leave it here. Be sure to include this message ID: [WMT-114002]
Google Inc. 1600 Amphitheatre Parkway Mountain View, CA 94043 I Unsubscribe.



Thanks To "Right To Be Forgotten," Google Now Censors The Press In The EU

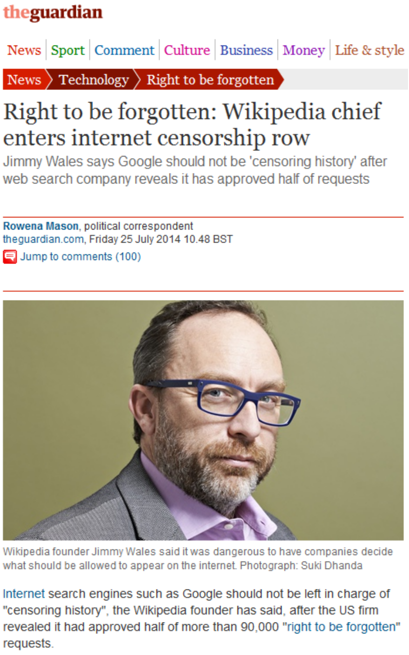
Danny Sullivan on July 2, 2014 at 4:11 pm

The EU's Right To Be Forgotten removals have been happening for about a week on Google, and now news publications are discovering the fallout. For some searches, you can't find their news stories relating to certain people.

In particular, both the BBC and the Guardian have shared examples of content that's now been "forgotten" in Google. The stories remain on the sites of both publications. You just can't locate them for certain searches related to the names of individuals they are about. (Postscript 5:50pm ET: Add the Daily Mail to the list, which has posted about removal notices it has received).

Wikipedias Jimmy Wales om sensur:

Er det nå slik at en SERP (Search Engine Results Page) fra Google er «Historien» om et gitt tema?


theguardian
News | Sport | Comment | Culture | Business | Money | Life & style
News › Technology › Right to be forgotten

Right to be forgotten: Wikipedia chief enters internet censorship row

Jimmy Wales says Google should not be 'censoring history' after web search company reveals it has approved half of requests

Rowena Mason, political correspondent
theguardian.com, Friday 25 July 2014 10.48 BST
Jump to comments (100)

Wikipedia founder Jimmy Wales said it was dangerous to have companies decide what should be allowed to appear on the internet. Photograph: Suki Dhanda

Internet search engines such as Google should not be left in charge of "censoring history", the Wikipedia founder has said, after the US firm revealed it had approved half of more than 90,000 "right to be forgotten" requests.

INF5270

Design av interaktive nettsteder

Informasjonsarkitektur:

Mer om søk

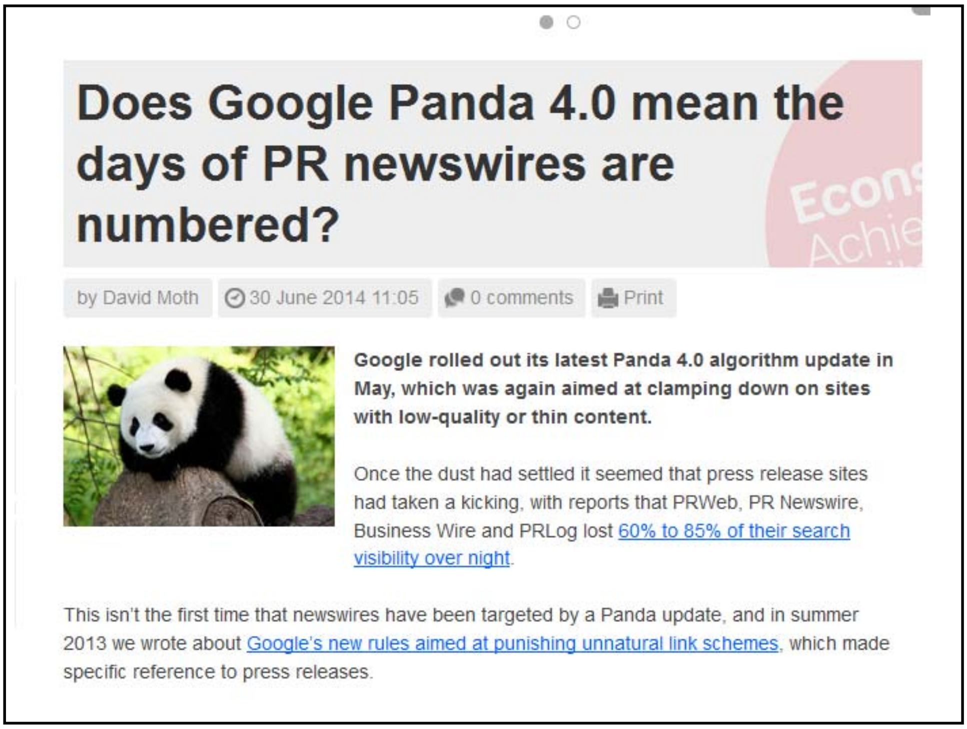
Gisle Hannemyr
Ifi, vårsemesteret 2012

Søkeresultater

- Sortering (rangering)
  - Relevans
  - Kronologisk
  - Alfabetisk
  - Popularitet
  - Score-basert
  - (Betalingsbasert)

Sortering/Rangering

- **Rangering** = 0.30x(**Alder**) + 0.40x(**Nøkkelordtetthet**) + 0.30x(**Nøkkelordenes plassering**).
  - Hver verdi tilordnes en vekt - i dette tilfellet tallet som stå foran parentesene.
  - Bare en illustrasjon. I virkeligheten kan det være flere hundre faktorer.
- **Teori fra sosiale nettverk og biblioteker**. Sammenhenger, sitater og referanser. Direkte og indirekte henvisninger.



Does Google Panda 4.0 mean the days of PR newswires are numbered?

by David Moth | 30 June 2014 11:05 | 0 comments | Print

Google rolled out its latest Panda 4.0 algorithm update in May, which was again aimed at clamping down on sites with low-quality or thin content.

Once the dust had settled it seemed that press release sites had taken a kicking, with reports that PRWeb, PR Newswire, Business Wire and PRLog lost [60% to 85% of their search visibility over night]().

This isn't the first time that newswires have been targeted by a Panda update, and in summer 2013 we wrote about [Google's new rules aimed at punishing unnatural link schemes](), which made specific reference to press releases.

SEO
= Search Engine Optimazation

- Ranking
- Ranking problems
- Black hat methods
- Google and SEO

2014 Gisle Hannemyr Page #28

Ranking algorithms

- Word count (AltaVista)
- PageRank (Google)
  - Hubber & autoriteter (+ mye som er hemmelig)
- Experimental:
  - Semantic mapping
  - Concordance computatuions
  - Semantic web / metadata
  - Webs of trust, quality ranking

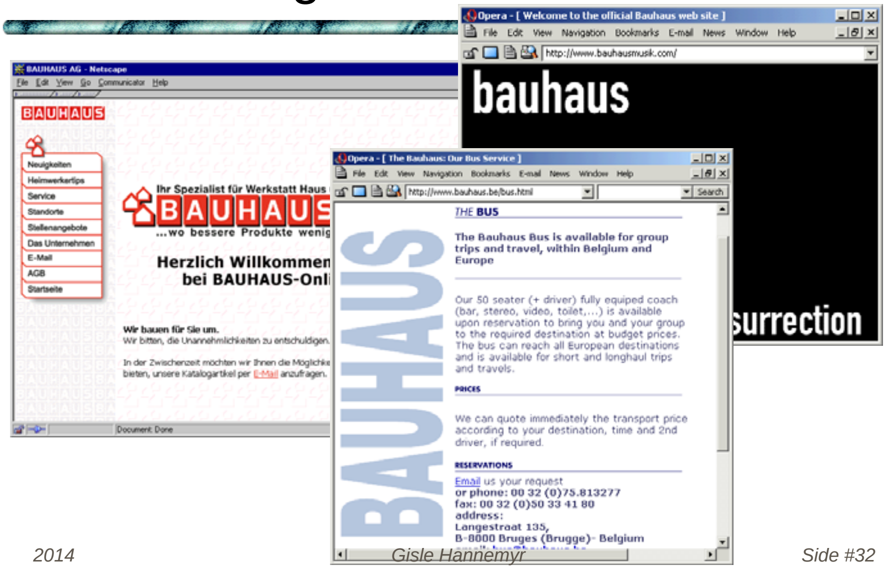
Ranking: Some problems

- Intrinsic problems:
  - Early Voter Problem (Link cardinality, Google love)
  - Culture bias (Weighted link cardinality)
- Black hat SEO:
  - Portal pages (pages designed for SEO)
  - Link spamming and link exchanges (link cardinality)
  - Word spamming (Word count)

Fra intervju med en lege om søk i INSPEC vs. Internett-søk etter medisinsk informasjon.

- «Det jeg mener karakteriserer slike systemer [som INSPEC] er den høye grad av struktur på informasjonen, bruk av kontrollerte vokabularer for nøkkelord, og at jeg som bruker har tiltro til de som opererer informasjonsbasene.»
- «Spesielt det siste er viktig. La oss tenke oss at man tok den kontrollerte vokabularet (thesaurus) utviklet for INSPEC i bruk for søking på Internett. **Det ville vært en katastrofe.** Det ville ikke gitt kvalitet i det hele tatt.»

Searching for "Bauhaus"



BAUHAUS AG - Netscape

Ihr Spezialist für Werkstatt Haus

BAUHAUS

...wo bessere Produkte wenig

Herzlich Willkommen
bei BAUHAUS-Onli

Wir bauen für Sie um.
Wir bitten, die Unannehmlichkeiten zu entschuldigen.

In der Zwischenzeit möchten wir Ihnen die Möglichke
bieten, unsere Katalogartikel per E-Mail anzufragen.

Opera - [Welcome to the official Bauhaus web site]

http://www.bauhausmusik.com/

Opera - [The Bauhaus: Our Bus Service]

http://www.bauhaus.be/bus.html

THE **BUS**

The Bauhaus Bus is available for group trips and travel, within Belgium and Europe

Our 50 seater (+ driver) fully equiped coach (bar, stereo, video, toilet,...) is available upon reservation to bring you and your group to the required destination at budget prices. The bus can reach all European destinations and is available for short and longhaul trips and travels.

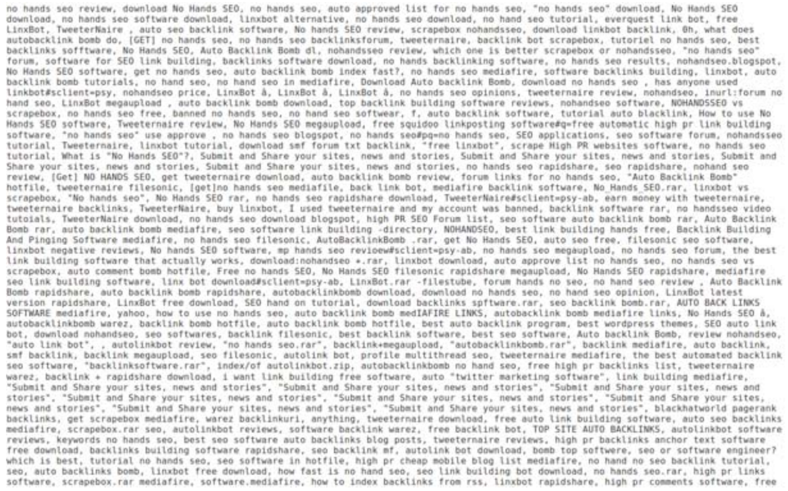
PRICES

We can quote immediately the transport price according to your destination, time and 2nd driver, if required.

RESERVATIONS

Email us your request
or phone: 00 32 (0)75.813277
fax: 00 32 (0)50 33 41 80
address:
Langestraat 135,
B-8000 Bruges (Brugge)- Belgium

Keyword stuffing



no hands seo review, download No Hands SEO, no hands seo, auto approved list for no hands seo, "no hands seo" download, No Hands SEO download, no hands seo software download, linxbot alternative, no hands seo download, no hand seo tutorial, everquest link bot, free LinxBot, TweeterNaire , auto seo backlink software, No hands SEO review, scrapebox nohandsseo, download linkbot backlink, 0h, what does autobacklink bomb do, [GET] no hands seo, no hands seo backlinksforum, tweeternaire, backlink bot scrapebox, tutoriel no hands seo, best backlinks sofftware, No Hands SEO, Auto Backlink Bomb dl, nohandsseo review, which one is better scrapebox or nohandsseo, "no hands seo" forum, software for SEO link building, backlinks software download, no hands backlinking software, no hands seo results, nohandseo.blogspot, No Hands SEO software, get no hands seo, auto backlink bomb index fast?, no hands seo mediafire, software backlinks building, linxbot, auto backlink bomb tutorials, no hand seo, no hand seo in mediafire, Download Auto backlink Bomb, download no hands seo , has anyone used linkbot#sclient=psy, nohandseo price, LinxBot å, LinxBot å, LinxBot å, no hands seo opinions, tweeternaire review, nohandseo, inurl:forum no hand seo, LinxBot megaupload , auto backlink bomb download, top backlink building software reviews, nohandseo software, NOHANDSSEO vs scrapebox, no hands seo free, banned no hands seo, no hand seo softwear, f, auto backlink software, tutorial auto blacklink, How to use No Hands SEO software, Tweeternaire review, No Hands SEO megaupload, free squidoo linkposting software#q=free automatic high pr link building software, "no hands seo" use approve , no hands seo blogspot, no hands seo#pq=no hands seo, SEO applications, seo software forum, nohandsseo tutorial, Tweeternaire, linxbot tutorial, download smf forum txt backlink, "free linxbot", scrape High PR websites software, no hands seo tutorial, What is "No Hands SEO"?, Submit and Share your sites, news and stories, Submit and Share your sites, news and stories, Submit and Share your sites, news and stories, Submit and Share your sites, news and stories, no hands seo rapidshare, seo rapidshare, nohand seo review, [Get] NO HANDS SEO, get tweeternaire download, auto backlink bomb review, forum links for no hands seo, "Auto Backlink Bomb" hotfile, tweeternaire filesonic, [get]no hands seo mediafile, back link bot, mediafire backlink software, No_Hands_SEO.rar, linxbot vs scrapebox, "No hands seo", No Hands SEO rar, no hands seo rapidshare download, TweeterNaire#sclient=psy-ab, earn money with tweeternaire, tweeternaire backlinks, TweeterNaire, buy linxbot, I used tweeternaire and my account was banned, backlink software rar, no handsseo video tutoials, TweeterNaire download, no hands seo download blogspot, high PR SEO Forum list, seo software auto backlink bomb rar, Auto Backlink Bomb rar, auto backlink bomb mediafire, seo software link building -directory, NOHANDSEO, best link building hands free, Backlink Building And Pinging Software mediafire, no hands seo filesonic, AutoBacklinkBomb .rar, get No Hands SEO, auto seo free, filesonic seo software, linxbot negative reviews, No hands SEO software, mp hands seo revioew#sclient=psy-ab, no hands seo megaupload, no hands seo forum, the best link building software that actually works, download:nohandseo +.rar, linxbot download, auto approve list no hands seo, no hands seo vs scrapebox, auto comment bomb hotfile, Free no hands SEO, No Hands SEO filesonic rapidshare megaupload, No Hands SEO rapidshare, mediafire seo link building software, linx bot download#sclient=psy-ab, LinxBot.rar -filestube, forum hands no seo, no hand seo review , Auto Backlink Bomb rapidshare, auto backlink bomb rapidshare, autobacklinkbomb download, download no hands seo, no hand seo opinion, LinxBot latest version rapidshare, LinxBot free download, SEO hand on tutorial, download backlinks spftware.rar, seo backlink bomb.rar, AUTO BACK LINKS SOFTWARE mediafire, yahoo, how to use no hands seo, auto backlink bomb medIAFIRE LINKS, autobacklink bomb mediafire links, No Hands SEO å, autobacklinkbomb warez, backlink bomb hotfile, auto backlink bomb hotfile, best auto backlink program, best wordpress themes, SEO auto link bot, download nohandseo, seo softwares, backlink filesonic, best backlink software, best seo software, Auto backlink Bomb, review nohandseo, "auto link bot", , autolinkbot review, "no hands seo.rar", backlink+megaupload, "autobacklinkbomb.rar", backlink mediafire, auto backlink, smf backlink, backlink megaupload, seo filesonic, autolink bot, profile multithread seo, tweeternaire mediafire, the best automated backlink seo software, "backlinksoftware.rar", index/of autolinkbot.zip, autobacklinkbomb no hand seo, free high pr backlinks list, tweeternaire warez, backlink + rapidshare download, i want link building free software, auto "twitter marketing software", link building mediafire, "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", blackhatworld pagerank backlinks, get scrapebox mediafire, warez backlinkuri, anything, tweeternaire download, free auto link building software, auto seo backlinks mediafire, scrapebox.rar seo, autolinkbot reviews, software backlink warez, free backlink bot, TOP SITE AUTO BACKLINKS, autolinkbot software reviews, keywords no hands seo, best seo software auto backlinks blog posts, tweeternaire reviews, high pr backlinks anchor text software free download, backlinks building software rapidshare, seo backlink mf, autolink bot download, bomb top softwere, seo or software engineer? which is best, tutorial no hands seo, seo software in hotfile, high pr cheap mobile blog list mediafire, no hand no seo backlink tutorial, seo, auto backlinks bomb, linxbot free download, how fast is no hand seo, seo link building bot download, no hands seo.rar, high pr links software, scrapebox.rar mediafire, software.mediafire, how to index backlinks from rss, linxbot rapidshare, high pr comments software, free

Keyword stuffing (with a too trusting search engine)

Siden gir 66% score – ledsaget av følgende forklarende tekst:

INFORMATION CONTENT IS FOCUSED TOWARDS KEY TOPICS The text appears to be very significant. It should be highly interesting due to high information value. It addresses key issues such as *bauhau, art, architecture, bauhau style, national socialism, architecture movement, nazism, craft movement and craft.* Relevant. Some core concepts such as *architect ludwig mie van der rohe, dessau, germany, fine art, international style, art academy, craftsmanship, craftsman william morri and aesthetic standard* are addressed in an informative way.

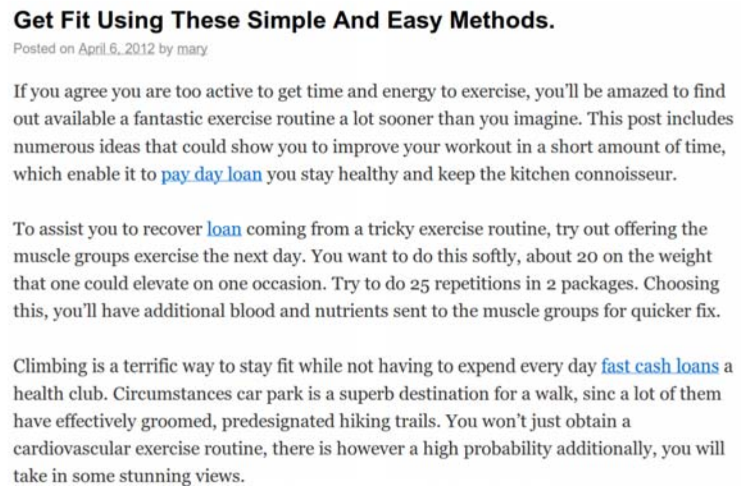
Denne siden er imidlertid bare en online ordliste inneholdene drøyt 45 tusen ord i alfabetisk rekkefølge, inklusive: *art, academy, aesthetic, architect, architecture, bauhaus, craft, craftsman, fine, germany, international, ludwig, morris, movement, national, nazism, socialism, standard, style, van* og *william*.

http://www.enter.net/~butcher/engine.html (404)

Av opphavsmannen selv karakteriseres siden som følger:

This page is completely useless, and is meant only to trick the search engines a bit. I sometimes see pages that have a lot of words that deal with their page down at the bottom, so that search engines hit the page. Well, this is the “Super Macho Man” of search engine trickers.

Link spam
(links unrelated to content)



Get Fit Using These Simple And Easy Methods.

Posted on April 6, 2012 by mary

If you agree you are too active to get time and energy to exercise, you'll be amazed to find out available a fantastic exercise routine a lot sooner than you imagine. This post includes numerous ideas that could show you to improve your workout in a short amount of time, which enable it to pay day loan you stay healthy and keep the kitchen connoisseur.

To assist you to recover loan coming from a tricky exercise routine, try out offering the muscle groups exercise the next day. You want to do this softly, about 20 on the weight that one could elevate on one occasion. Try to do 25 repetitions in 2 packages. Choosing this, you'll have additional blood and nutrients sent to the muscle groups for quicker fix.

Climbing is a terrific way to stay fit while not having to expend every day fast cash loans a health club. Circumstances car park is a superb destination for a walk, sinc a lot of them have effectively groomed, predesignated hiking trails. You won't just obtain a cardiovascular exercise routine, there is however a high probability additionally, you will take in some stunning views.

Some "black hat" methods Google explictly bans

- Blacklettering (invisible text)
- Keyword stuffing
- Link spam (often created through "link exchanges", orphaned blogs, referrerlogs, etc.)
- Cloaking (showing robot and humans different pages)
- Portal pages (SEO optimized page that link to "real page")
- Misuse of metatags (now mostly ignored)

Søkekvalitet

- The fallacy of abundance
- Precision and recall
- Hvordan måle søkekvalitet

The Fallacy of Abundance

- Don Swanson (1960):
  - The fallacy of abundance is the mistake a searcher makes when he uses a large IR system and is able to find *some* useful documents.
  - On a sufficiently large system […] almost any query will retrieve *some* useful documents.
  - The mistake is to think that just because you got some useful documents the IR system is performing well. What you don’t know is how many better [or at least relevant] documents the system missed.

Precision og recall (1)

- I den etterfølgende diskusjon av precision og recall i samband med søk, vil jeg benytte følgende notasjon:
  - Det totale antall dokumenter i samlingen: T.
  - Det totale antall *relevante* dokumenter i samlingen: R.
  - Det totale antall dokumenter i resultatmengden: t.
  - Det totale antall *relevante* dokumenter i resultatmengden: r.
- Hva som ligger i at et dokument er «relevant» drøftes senere.
- «Samlingen» er universet av dokumenter (for eksempel alle dokumenter på web).
- «Resultatmengden» er mengden av dokumenter som betraktes som resultatet av et søk. Det kan for eksempel være ett treff (det første treffet), de første ti treffene, eller samtlige (mange tusen) treff som en typisk internett-søkemotor som Google returnerer.
- I litteraturen om søk defineres det to beregnbare enheter (precision og recall) som kan benyttes for å si noe om kvaliteten på og egenskapene til et gitt søk.

Precision og recall (2)

- **precision = r / t**
  - Maksimal verdi (1) betyr at samtlige treff er relevante.
  - Sier noe om hvor stor *andel* av resultatmengden som er relevant (men ikke noe om hvor mange relevante dokumenter som *ikke* er med i resultatmengden).
  - Vil falle med økende t, så precision-orienterte søkemotorer vil gjerne sette t til en lav verdi og håpe på det beste (jf. Google's "I feel lucky".)
  - Vi trenger ikke kjenne antall relevante dokumenter i samlingen for å beregne *precision*.

Precision og recall (3)

- **recall = r / R**
  - Maksimal verdi (1) betyr at vi har funnet samtlige relevante dokumenter i samlingen.
  - Sier noe om hvor stor andel av relevante dokumenter som er inneholdt i resultatmengden (men ikke noe om hvor mange irrelevante dokumenter vi har fått på kjøpet).
  - Vil øke med økende t, så recall-orienterte søkemotorer vil gjerne sette t til en høy verdi og overlate jobben med å skille klinten fra hveten til brukeren.
  - Trenger å kjenne antall relevante dokumenter i samlingen (R) for å beregne *recall* (noe som ikke er realistisk for store samlinger som web, men det finnes brukbare alt. tilnærminger).

Precision og recall (4)

- **f-score = (2 x precision x recall) / (precision + recall)**
  - En kombinasjon av de to verdiene (med samme t) beregnet etter den statistiske formelen for harmonisk middelverdi kan brukes i analyse av søk.
  - Eksempler:
    - Sett t til en hensiktsmessig verdi (f.eks. 10) og sammenlign alternative søkealgoritmer.
    - Plott f-score som en funksjon av t for en gitt søkealgoritme.
  - Finnes også mange andre statistiske tilnærminger til å beregne verdier som sier noe om kvaliteten på en søkealgoritme.

Som det framgikk: Et godt søk er et som lar oss finne *relevante* dokumenter

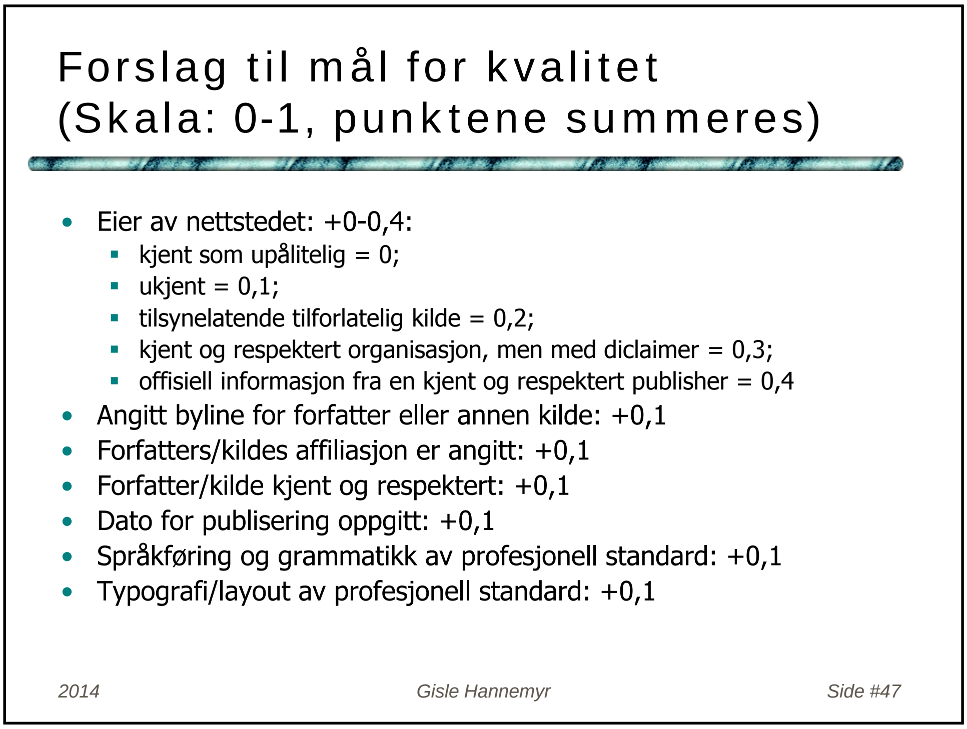
- Relevansbegrepet er svært kompleks. Det opereres i litteraturen med flere flere ulike relevansbegreper.
- Følgende to relevanskriterier er svært mye brukt:
  - Topikalitet (også kalt for innholdsrelevans)
    Topikalitet er et mål for samstemmighetsrelasjonen mellom søkeforespørsel og søkesvar. Topikalitet er *uavhengig av brukerens behov eller situasjon*. Topikalitet skal derfor bedømmes av domene-eksperter, ikke av brukeren.
  - Kvalitet
    Dette er et mål for systemets evne til å rangere *høyverdige ressurser* (definert langs slike definisjoner som lødig innhold, troverdig kilde, relevant genre) foran mindre verdige resursser. Også kvalitet bedømmes best av domeneeksperter.

Tredje mål for relevans

- I en del litteraturen opereres det i tilegg med et tredje mål som vanligvis kalles for *nytteverdi* (utility) eller *subjektiv relevans*.
- Dette er et mål for relasjonen mellom brukerens situasjonsbestemte behov for informasjonsressurser og resulatet av søket.
- Nytteverdi påvirkes både av hva brukeren akter å bruke informasjonen til, og hva slags kunnskaper brukeren har om emnet på forhånd (brukere har for eksempel mer nytte av informasjon som bibringer dem ny informasjon, og mindre nytte av å informasjon som de allerede kjenner til.)
- Nytteverdi må derfor bedømmes av *brukeren* av informasjonen.

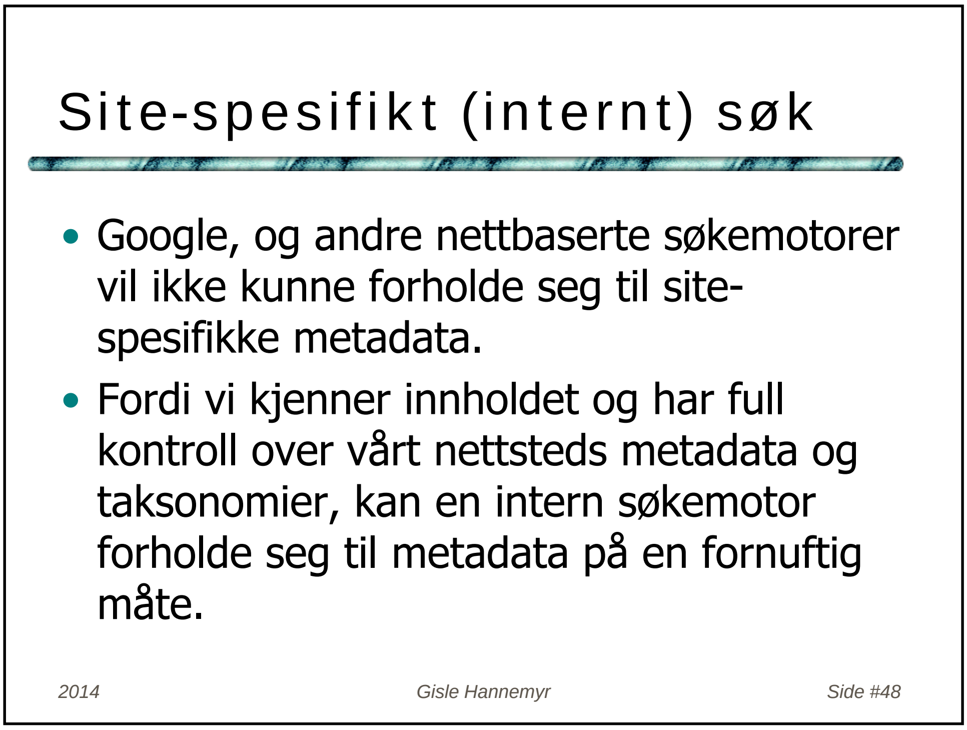
Forslag til mål for topikalitet (Skala: 0-1, gradert skala)

- Siden omhandler et annet emne = 0.
- Siden omhandler emnet mariginalt = 0,1-0,2
- Siden omhandler emnet mariginalt, men gir referanser til eksterne som utdyper emnet = 0,3-0,4
- Siden inneholder en god del informasjon om det emnet det søkes etter = 0,5-0,6
- Siden inneholder mye informasjon om det emnet det søkes etter = 0,7-0,9
- Siden handler om eksakt det emne det søkes etter = 1.0



Forslag til mål for kvalitet (Skala: 0-1, punktene summeres)

- Eier av nettstedet: +0-0,4:
  - kjent som upålitelig = 0;
  - ukjent = 0,1;
  - tilsynelatende tilforlatelig kilde = 0,2;
  - kjent og respektert organisasjon, men med diclaimer = 0,3;
  - offisiell informasjon fra en kjent og respektert publisher = 0,4
- Angitt byline for forfatter eller annen kilde: +0,1
- Forfatters/kildes affiliasjon er angitt: +0,1
- Forfatter/kilde kjent og respektert: +0,1
- Dato for publisering oppgitt: +0,1
- Språkføring og grammatikk av profesjonell standard: +0,1
- Typografi/layout av profesjonell standard: +0,1



Site-spesifikt (internt) søk

- Google, og andre nettbaserte søkemotorer vil ikke kunne forholde seg til site-spesifikke metadata.
- Fordi vi kjenner innholdet og har full kontroll over vårt nettsteds metadata og taksonomier, kan en intern søkemotor forholde seg til metadata på en fornuftig måte.

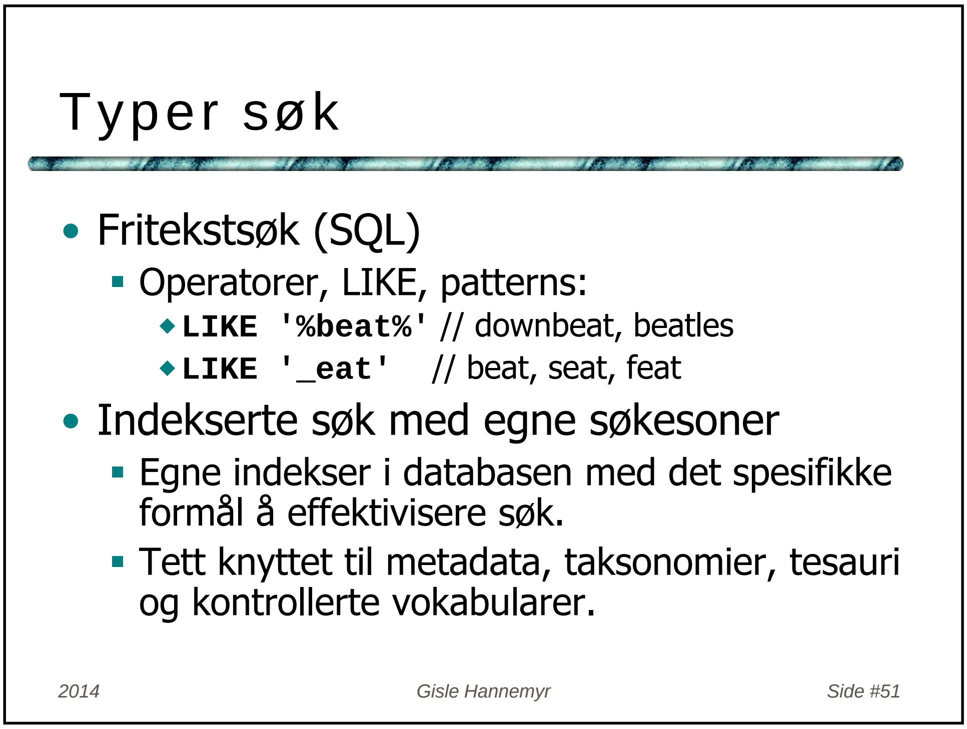
The Dublin Core Element Set v1.1

15 *explicit* predefined metadata elements

Fields	Description
Title	The name given to the resource, usually by the Creator or Publisher.
Subject	A comma-separated list of keywords describing subject.
Description	Free text description of the content of the resource.
Source	A resource identifier pointing to a resource from which the present resource is derived.
Relation	A resource identifier pointing to a second, related resource (e.g. URI).
Language	RFC 1766 tag identifying the language of the intellectual content of the resource.
Coverage	The spatial or temporal characteristics of the intellectual content of the resource.
Creator	Author (can be a person, an organisation or a service).
Publisher	The entity responsible for making the resource available in its present form.
Contributor	An entity not listed as Creator who has made intellectual contributions to the resource.
Rights	A rights management statement, or an resource identifier that links to such a statement.
Date	ISO8601-type date associated with the creation or availability of the resource.
Type	The nature or genre of the content of the resource.
Format	The digital or physical manifestation of the resource (e.g. Mime media types)
Identifier	A string or number used to uniquely identify the resource (e.g. URI or ISBN)

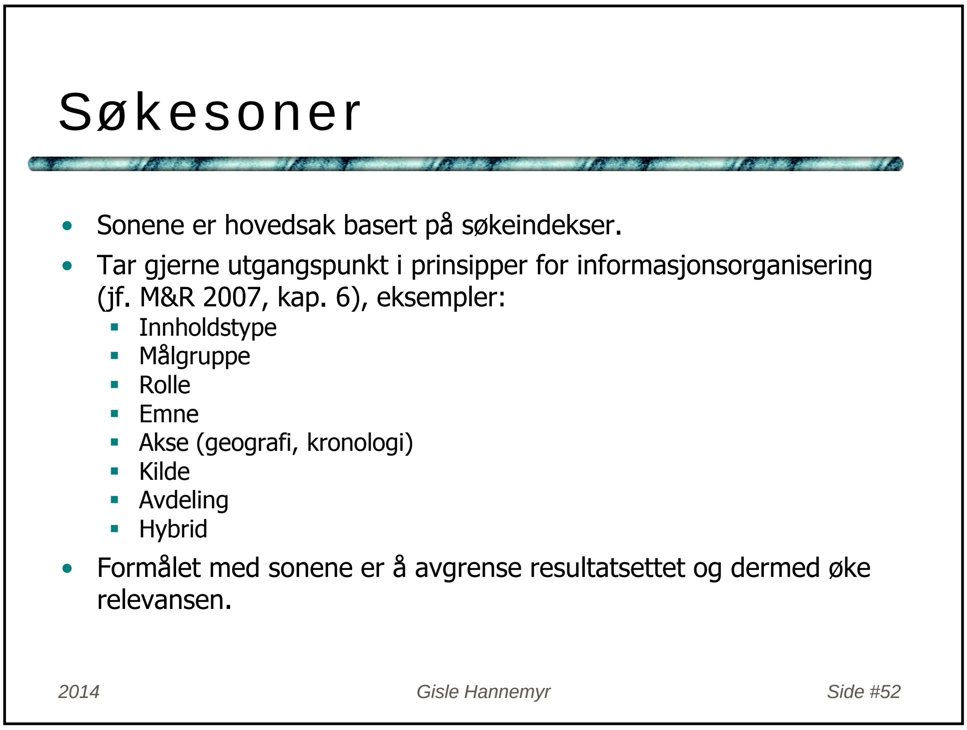
Søke-syntaks

- Enkelt (enkel søkeboks der man taster inn det eller de nøkkelord man vil søke på, med en implisitt logisk "OG" mellom ord.)
- Avansert (benyttes av svært få brukere):
  - Logiske operatorer (OG ELLER IKKE)
  - + - (AltaVista brukte denne notasjonen)
  - "Alle ordene", "ingen av ordene", "eksakt frase", etc.



Typer søk

- Fritekstsøk (SQL)
  - Operatorer, LIKE, patterns:
    - **`LIKE '%beat%'`** // downbeat, beatles
    - **`LIKE '_eat'`** // beat, seat, feat
- Indekserte søk med egne søkesoner
  - Egne indekser i databasen med det spesifikke formål å effektivisere søk.
  - Tett knyttet til metadata, taksonomier, tesauri og kontrollerte vokabularer.

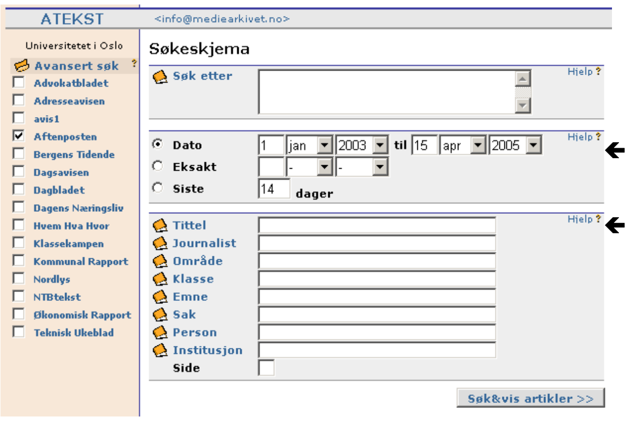


Søkesoner

- Sonene er hovedsak basert på søkeindekser.
- Tar gjerne utgangspunkt i prinsipper for informasjonsorganisering (jf. M&R 2007, kap. 6), eksempler:
  - Innholdstype
  - Målgruppe
  - Rolle
  - Emne
  - Akse (geografi, kronologi)
  - Kilde
  - Avdeling
  - Hybrid
- Formålet med sonene er å avgrense resultatsettet og dermed øke relevansen.

Eksempel: Søkesoner

Publikasjon ➔



ATEKST <info@mediearkivet.no>

Universitetet i Oslo

Avansert søk ?

- ☐ Advokatbladet
- ☐ Adresseavisen
- ☐ avis1
- ☑ Aftenposten
- ☐ Bergens Tidende
- ☐ Dagsavisen
- ☐ Dagbladet
- ☐ Dagens Næringsliv
- ☐ Hvem Hva Hvor
- ☐ Klassekampen
- ☐ Kommunal Rapport
- ☐ Nordlys
- ☐ NTBtekst
- ☐ Økonomisk Rapport
- ☐ Teknisk Ukeblad

Søkeskjema

Søk etter Hjelp ?

● Dato 1 jan 2003 til 15 apr 2005 Hjelp ?
○ Eksakt - -
○ Siste 14 dager

Hjelp ?

Tittel
Journalist
Område
Klasse
Emne
Sak
Person
Institusjon
Side

Søk&vis artikler >>

🡐 Kronologisk

🡐 Hybrid

Brukerens informasjonsbehov

- "Perfect catch": Problemet har et *eksakt svar* og brukeren vet når hun har funnet det.
- "Lobster trapping": Problemet kan løses ved at det finnes et *tilfredsstillende* svar.
- "Indiscriminate driftnetting": Problemet er vanskelig, og vi ønsker et *komplett* svar.

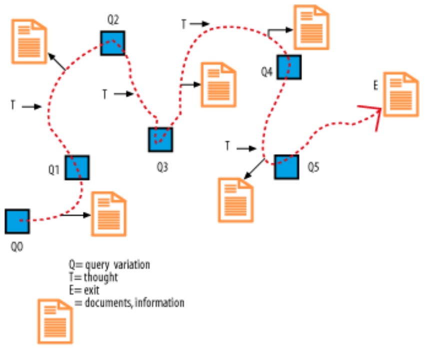
Informasjonssøkeradferd

- Måter som brukere benytter for å skaffe seg informasjon vha. ett nettsted.
  - **Skum** (browse) nettstedet ved å studere menyer, navn og følge hyperlenker – tar utgangspunkt i nettstedets navigasjonsstruktur.
  - **Søk** (via søkefelt) etter relevante termer (dekkes senere).
  - **Spør** (via epost eller chat) en person om hjelp til å finne informasjonen.
- Brukere kan betrakte dette som alternative strategier, eller de kan benytte en strategi der alle tre brukes i kombinasjon.

Berrypicking model (M. Bates)

- Bruker formulerer sitt opprinnelige spørsmål.
- Navigerer iterativt gjennom informasjonssystemet vha. de midler for navigasjon som står til rådighet, og akkumulerer biter av informasjon (berries) underveis.
- Spørsmålet re-formuleres hele tiden under denne prosessen:
  - Spørsmålet spisses fordi brukeren gjennom de biter som finnes blir i stand til å forstå bedre hva slags informasjon han/hun er ute etter.
  - Spørsmålet presiseres fordi fordi brukeren gjennom de biter som finnes blir i stand til å forstå bedre hva slags type informasjon det aktuelle informasjonssystemet rommer.

Berrypicking model



From: Marcia J. Bates: *The Design of Browsing and Berrypicking Techniques for the Online Search Interface;* Online Review 13:5, 1989.
http://www.gseis.ucla.edu/faculty/bates/berrypicking.html

Berrypicking som tre-traversering

- Brukere går "frem" (dvs. følger lenker dypere inn i nettstedet) dersom siden tolkes som et løfte om å lede til mer relevant innhold.
- Kommer de til sider som er irrelevante "rygger" tilbake til siste side som de oppfattet som relevant, og så går "frem" til et nytt sub-tre.

Pearlgrowing model

- Brukeren starter med ett dokument som passer rimelig godt med det han/hun ønsker og bruker dette som basis for å finne informasjon som “ligner”.
- Eksempler på denne modellen:
  - “Similar pages” (Google)
  - Semantisk nett match (Corporum)
  - Citation search (bibliografidatabase)

2014 Gisle Hannemyr Side #59

Two-step model

- Søk, deretter skum:
  - Bruk søkefunksjon til å identifisere subsites eller nettsteder som man kan anta inneholder relevant informasjon. Disse skummes deretter for å finne relevant informasjon.
- Skum, deretter søk:
  - Skum til man finner en egnet subsite, avgrens deretter søket til denne. (Er betinget av nettstedet støtter denne type avgrensede søk).

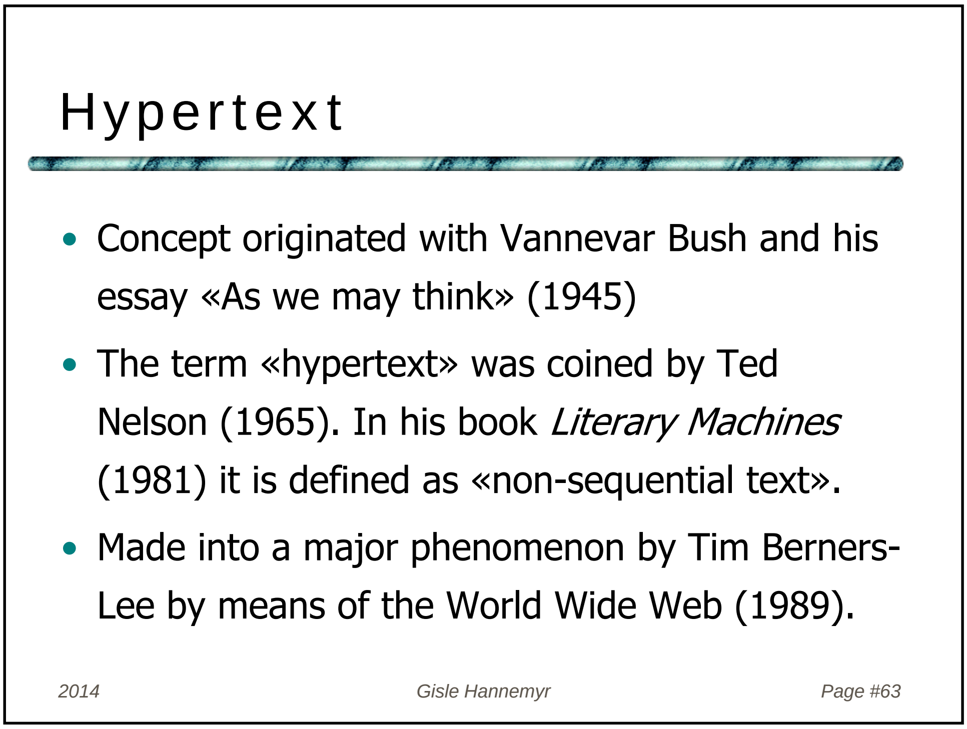
2014 Gisle Hannemyr Side #60

The semantic web

- The various technologies introduced in this lecture is often presented under the heading “the semantic web”.
- Concept coined in May 2001 – in an article in Scientific American by Tim Berners-Lee, James Hendler and Ora Lassila.
- … but let us first go back to the very roots of hypertext.

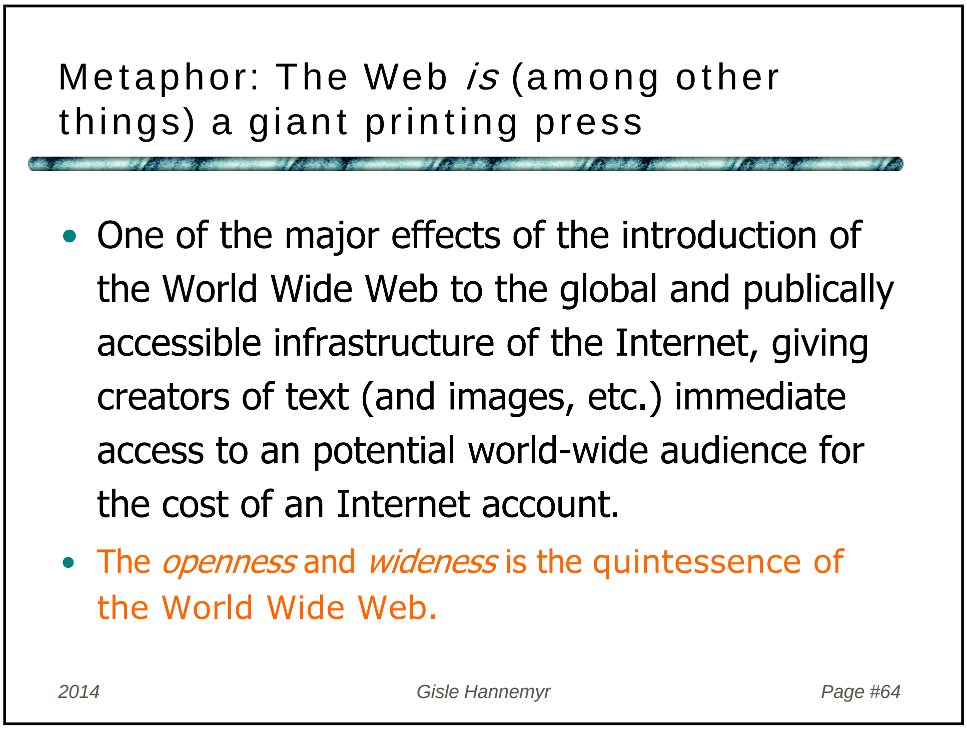
In the beginning.
“As we may think” (1945)

[…] publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships. (Vannevar Bush, 1945)



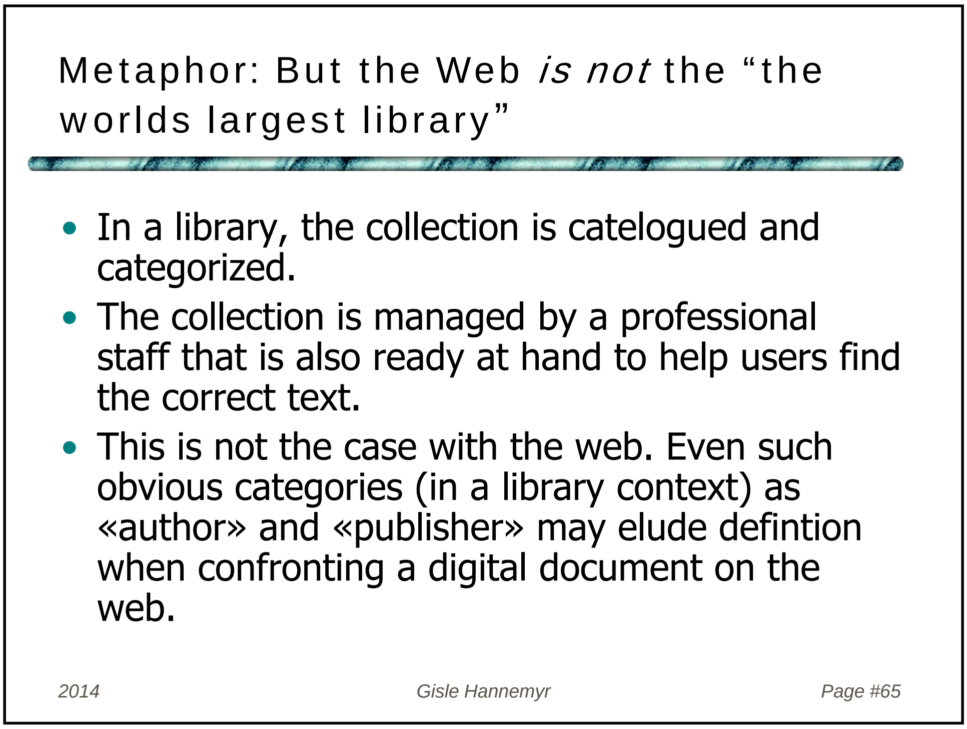
Hypertext

- Concept originated with Vannevar Bush and his essay «As we may think» (1945)
- The term «hypertext» was coined by Ted Nelson (1965). In his book *Literary Machines* (1981) it is defined as «non-sequential text».
- Made into a major phenomenon by Tim Berners-Lee by means of the World Wide Web (1989).



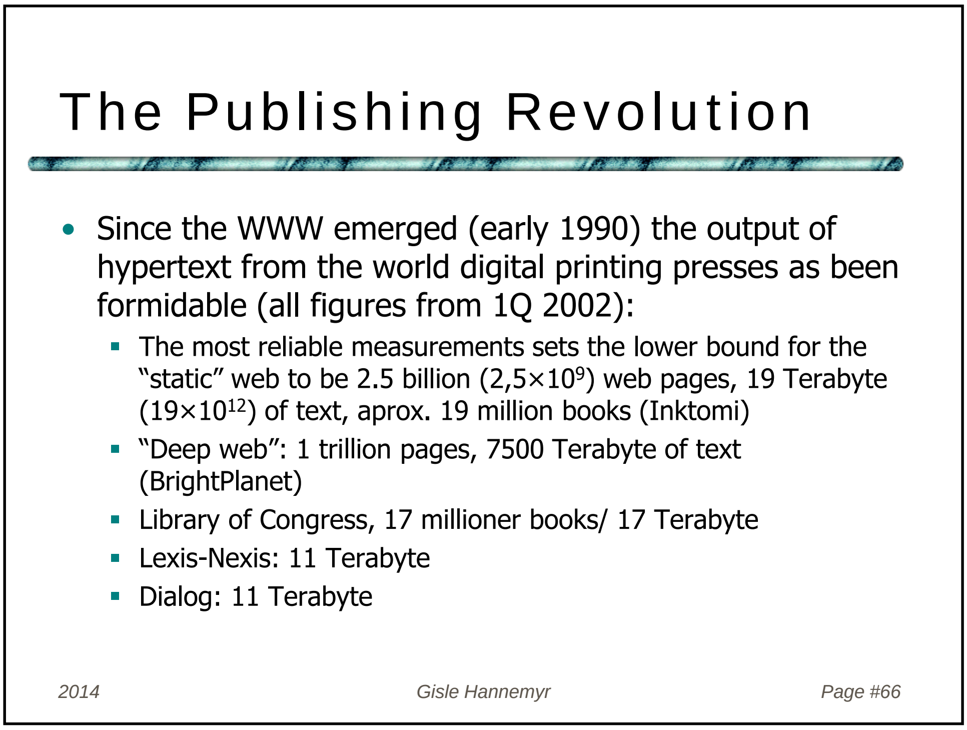
Metaphor: The Web *is* (among other things) a giant printing press

- One of the major effects of the introduction of the World Wide Web to the global and publically accessible infrastructure of the Internet, giving creators of text (and images, etc.) immediate access to an potential world-wide audience for the cost of an Internet account.
- The *openness* and *wideness* is the quintessence of the World Wide Web.



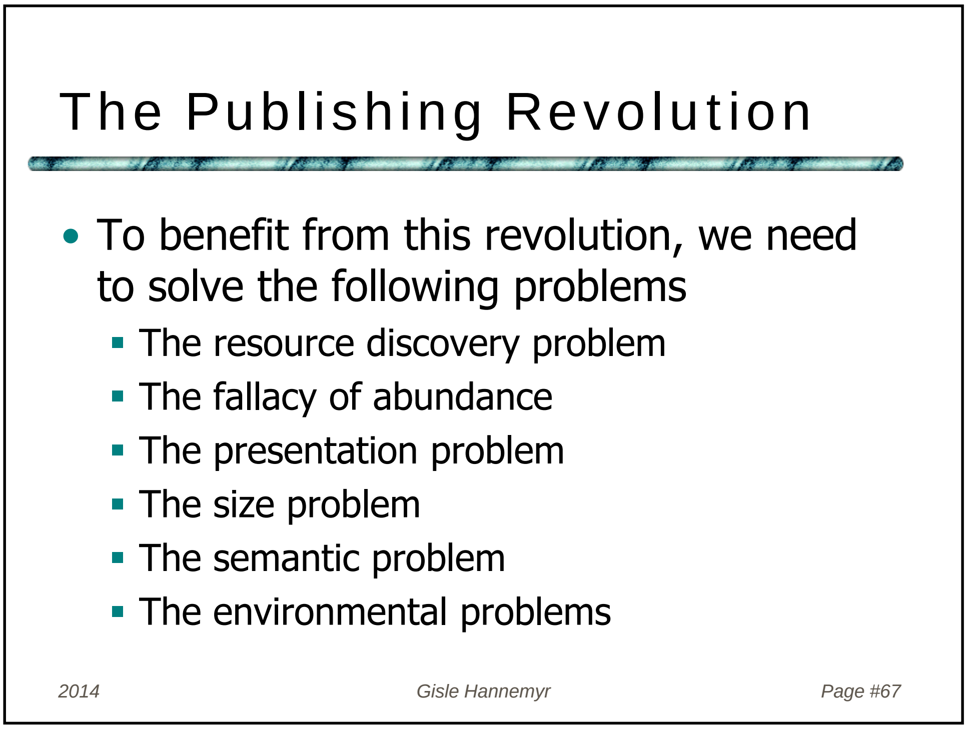
Metaphor: But the Web *is not* the "the worlds largest library"

- In a library, the collection is catelogued and categorized.
- The collection is managed by a professional staff that is also ready at hand to help users find the correct text.
- This is not the case with the web. Even such obvious categories (in a library context) as «author» and «publisher» may elude defintion when confronting a digital document on the web.



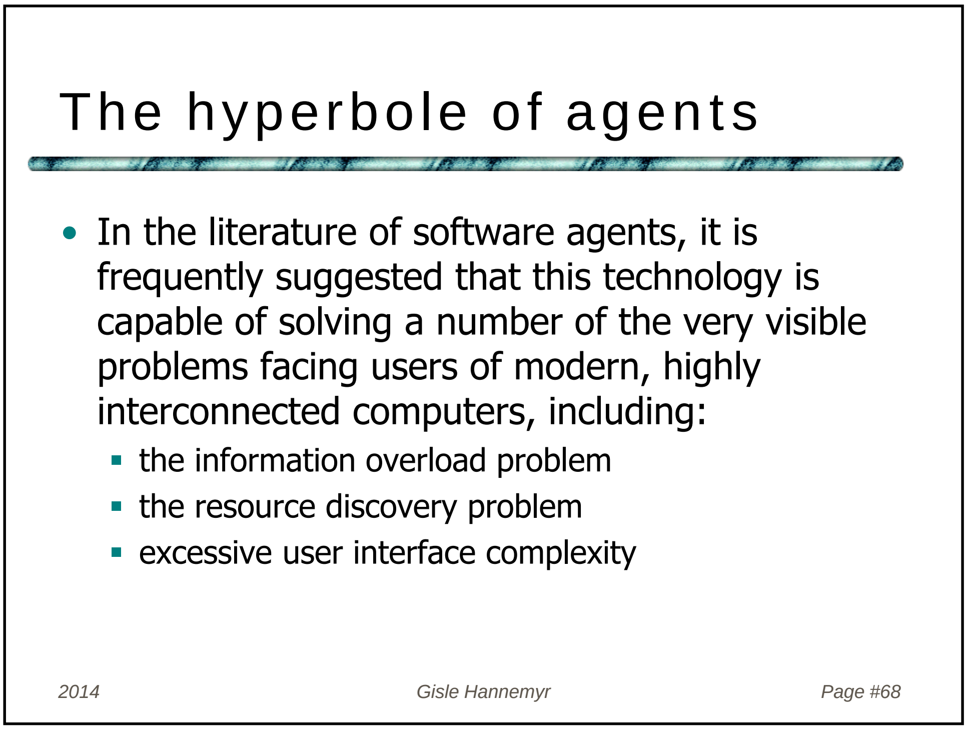
The Publishing Revolution

- Since the WWW emerged (early 1990) the output of hypertext from the world digital printing presses as been formidable (all figures from 1Q 2002):
  - The most reliable measurements sets the lower bound for the "static" web to be 2.5 billion ($2,5\times10^{9}$) web pages, 19 Terabyte (19×10^{12}) of text, aprox. 19 million books (Inktomi)
  - "Deep web": 1 trillion pages, 7500 Terabyte of text (BrightPlanet)
  - Library of Congress, 17 millioner books/ 17 Terabyte
  - Lexis-Nexis: 11 Terabyte
  - Dialog: 11 Terabyte



The Publishing Revolution

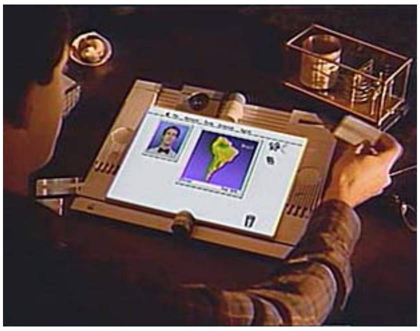
- To benefit from this revolution, we need to solve the following problems
  - The resource discovery problem
  - The fallacy of abundance
  - The presentation problem
  - The size problem
  - The semantic problem
  - The environmental problems



The hyperbole of agents

- In the literature of software agents, it is frequently suggested that this technology is capable of solving a number of the very visible problems facing users of modern, highly interconnected computers, including:
  - the information overload problem
  - the resource discovery problem
  - excessive user interface complexity

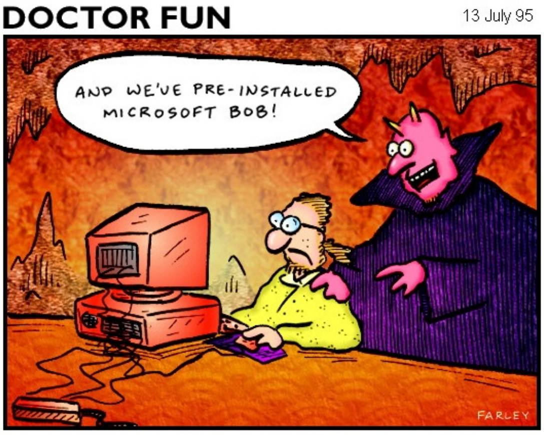
Apple's *The Knowledge Navigator,* featuring "Phil" (1987)



The status of agents

- So far, software agents has not managed to make much sense of electronic knowledge sources.
- The reason is probably that there is very little semantic information available electronically

Why has agents failed?



Unix gurus in hell

Comparing "classic" and Internet resource discovery

- The two worlds:

		Atekst	**Kvasir**	**Intrinsic**
1.	**IRs**	Hosted (IS=IRs)	Non-hosted (IS!=IRs)	Yes
2.	**IRs Structure**	Semantic oriented	Presentation oriented	No
3.	**Genres**	Newspaper articles	Miscellaneous	Yes
4.	**Persistence**	Persistent and in-sync	Transient/Ephemeral	Yes
5.	**Replication**	None	Some	Yes
6.	**Vocabulary**	Controlled	Chaotic	No
7.	**Agency**	Archivists	IR owners	Yes
8.	**Agenda**	«To provide a service»	«To generate hits»	Yes
9.	**Quality Ass.**	Source	Free for all	Partly

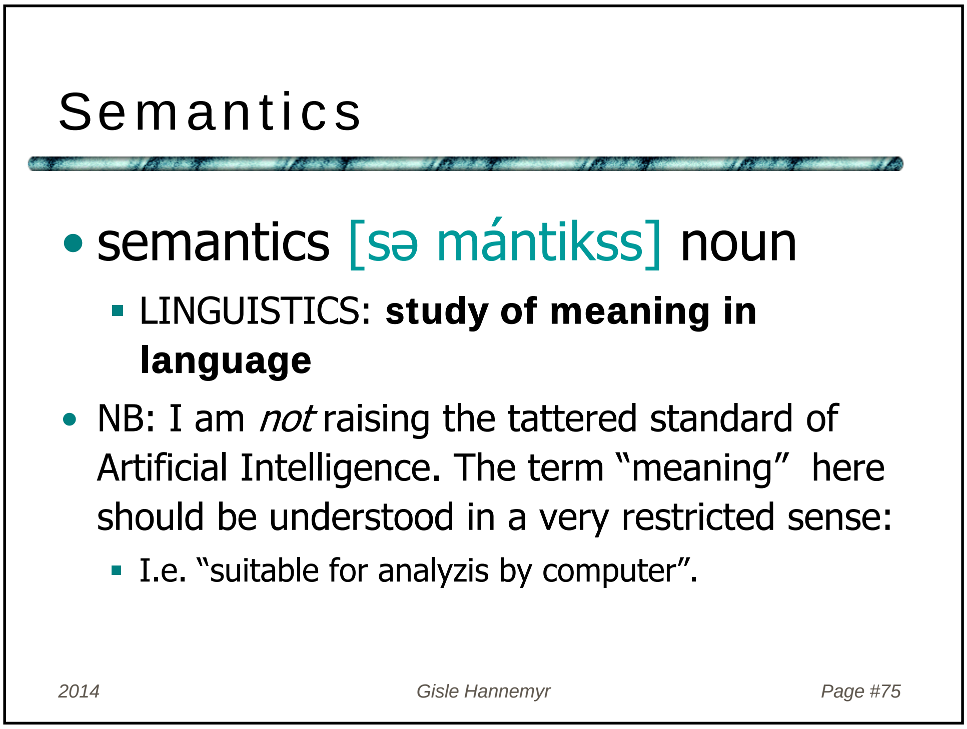
Table 1: *Atekst and Kvasir, summary of characteristics*

Summary of Problems with the Web

- The web's general orientation towards visual presentation.
- The web's free-for-all mixing of genres and formats.
- Deceptive games played by some actors to increase their pages visibility in search engines.
- Inability to handle change events and various types dynamic content (e.g. ephemeral pages, dynamic database selections, version superimpositioning).
- Lack of authentication mechanisms.

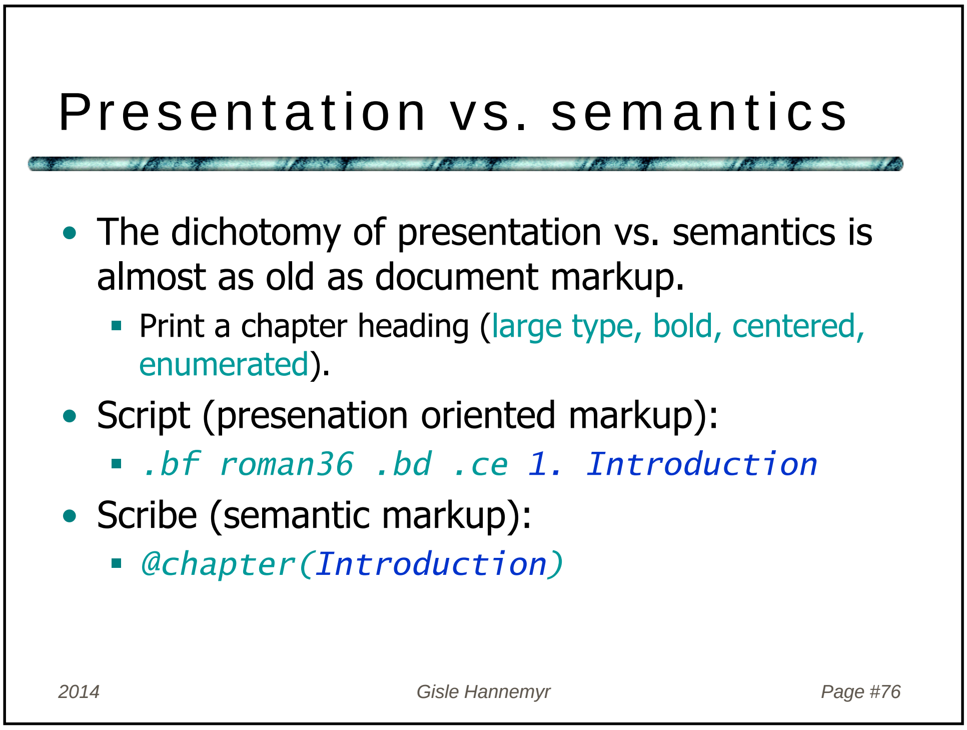
The proposed (partial?) fix: The Semantic Web

- The vision of the Semantic Web aims at creating a Web where information can be “understood” and processed by machines.
- This requires that information is represented in such a way that its *meaning* (i.e. its “semantics” ") is available in a machine-accessible form.



Semantics

- semantics [sə mántikss] noun
  - LINGUISTICS: **study of meaning in language**
- NB: I am *not* raising the tattered standard of Artificial Intelligence. The term "meaning" here should be understood in a very restricted sense:
  - I.e. "suitable for analyzis by computer".



Presentation vs. semantics

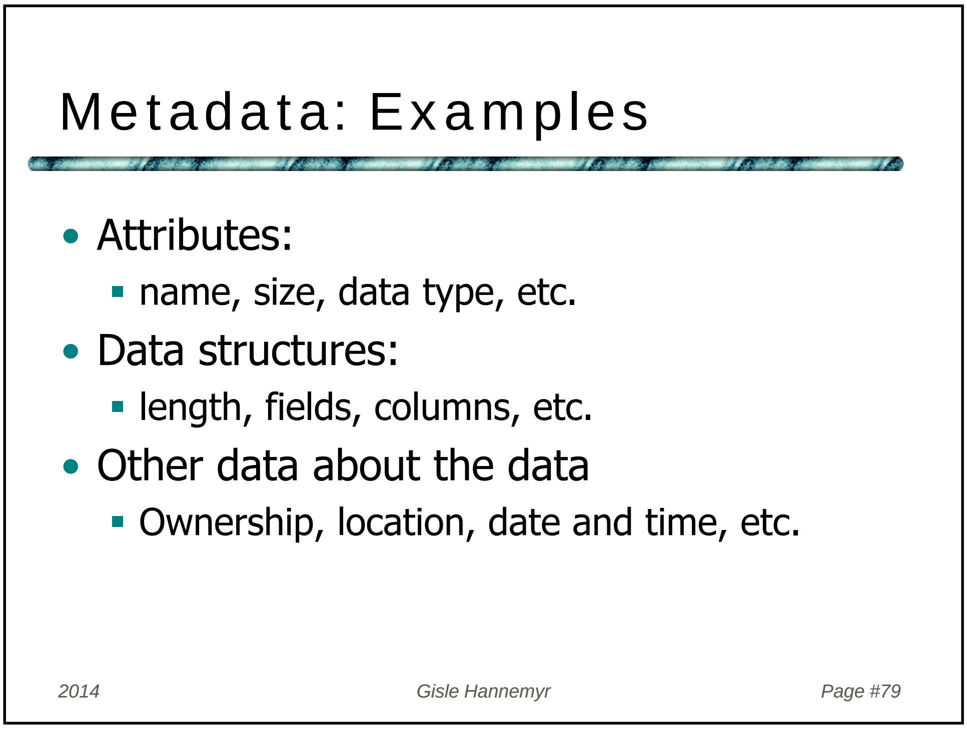
- The dichotomy of presentation vs. semantics is almost as old as document markup.
  - Print a chapter heading (large type, bold, centered, enumerated).
- Script (presenation oriented markup):
  - `.bf roman36 .bd .ce 1. Introduction`
- Scribe (semantic markup):
  - `@chapter(Introduction)`

The key component in the semantic web: Metadata

- Metadata is "data about data". Real life examples of metadata include such things as a library catalogue card (the "data" on the card describes the data contained in the books in the library) or a TV guide (the "data" in it describes the data in the programmes about to be broadcast).
- On the World Wide Web, webmasters may embed metadata in their web pages. So far this has meant using a schema where the Dublin Core model for metadata (Weibel et al 1998) are expressed by means of HTML META-tags (Kunze 1999). This model has a number of problems (such as meta-tag misuse by unscrupulous actors).
- Newer XML-based technologies, such as RDF and XTM allow construction by interested parties of ontologies independent of the web pages they refer to.

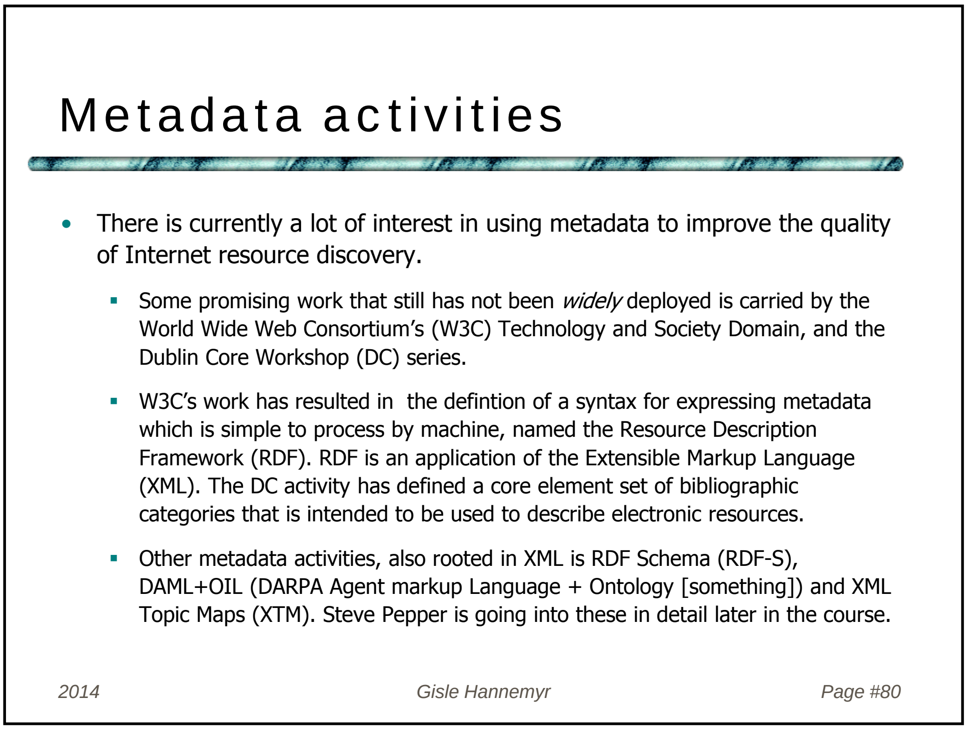
Metadata (data about data)

- *Definitional* data that provides information about or documentation of other data
- May include *descriptive information* about the context, quality, condition and characteristics of the data.
- Added by human help, extracted or synthezised.
- Managed within an application or environment to help us search, navigate and evaluate the data.



Metadata: Examples

- Attributes:
  - name, size, data type, etc.
- Data structures:
  - length, fields, columns, etc.
- Other data about the data
  - Ownership, location, date and time, etc.



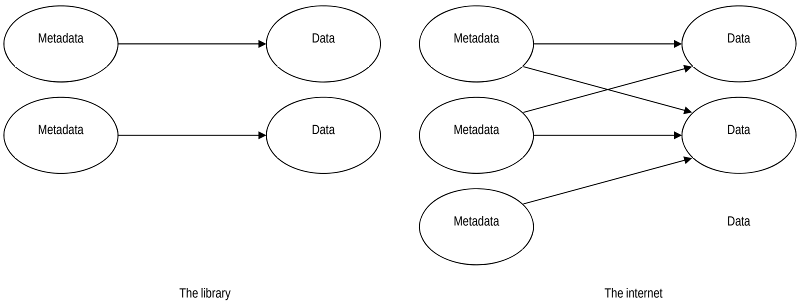
Metadata activities

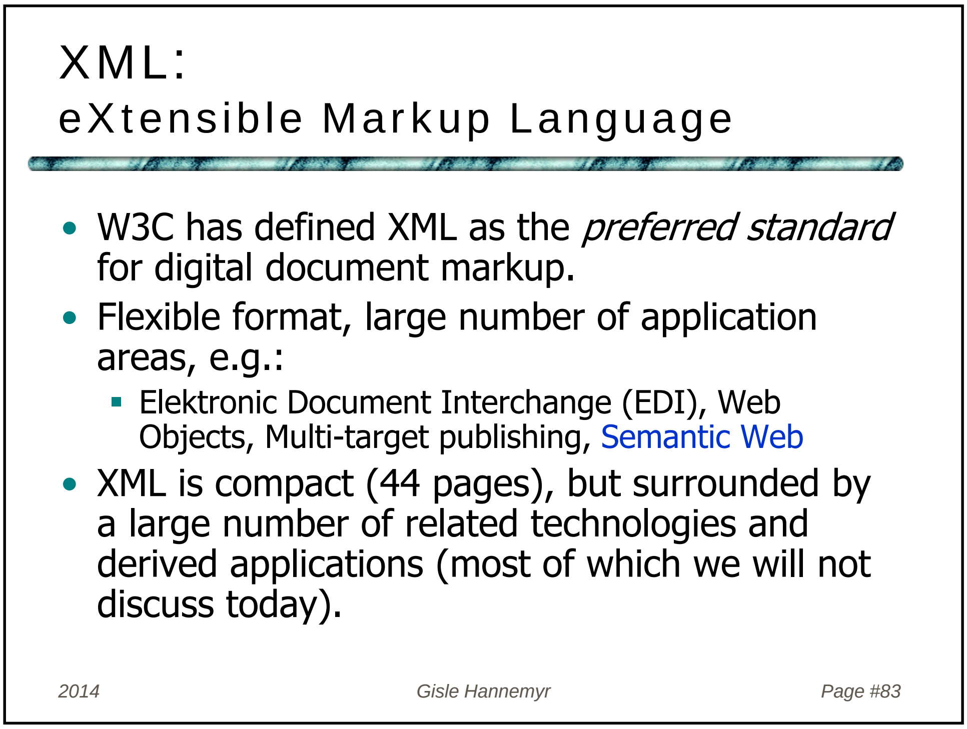
- There is currently a lot of interest in using metadata to improve the quality of Internet resource discovery.
  - Some promising work that still has not been *widely* deployed is carried by the World Wide Web Consortium's (W3C) Technology and Society Domain, and the Dublin Core Workshop (DC) series.
  - W3C's work has resulted in the defintion of a syntax for expressing metadata which is simple to process by machine, named the Resource Description Framework (RDF). RDF is an application of the Extensible Markup Language (XML). The DC activity has defined a core element set of bibliographic categories that is intended to be used to describe electronic resources.
  - Other metadata activities, also rooted in XML is RDF Schema (RDF-S), DAML+OIL (DARPA Agent markup Language + Ontology [something]) and XML Topic Maps (XTM). Steve Pepper is going into these in detail later in the course.

More about Metadata

- Metadata is data that describes and qualifies other data.
- Typical examples of metadata is:
  - important properties of the data (e.g. the name of the creator, and the year of publication),
  - data that is used to locate the data (e.g. the Dewey-code for a library book, and the time and channel for a television program),
  - data that is helpful when searching for data (e.g. a free-text description or a summary of the data, or a list of searchable subject keywords appropriate for the data).

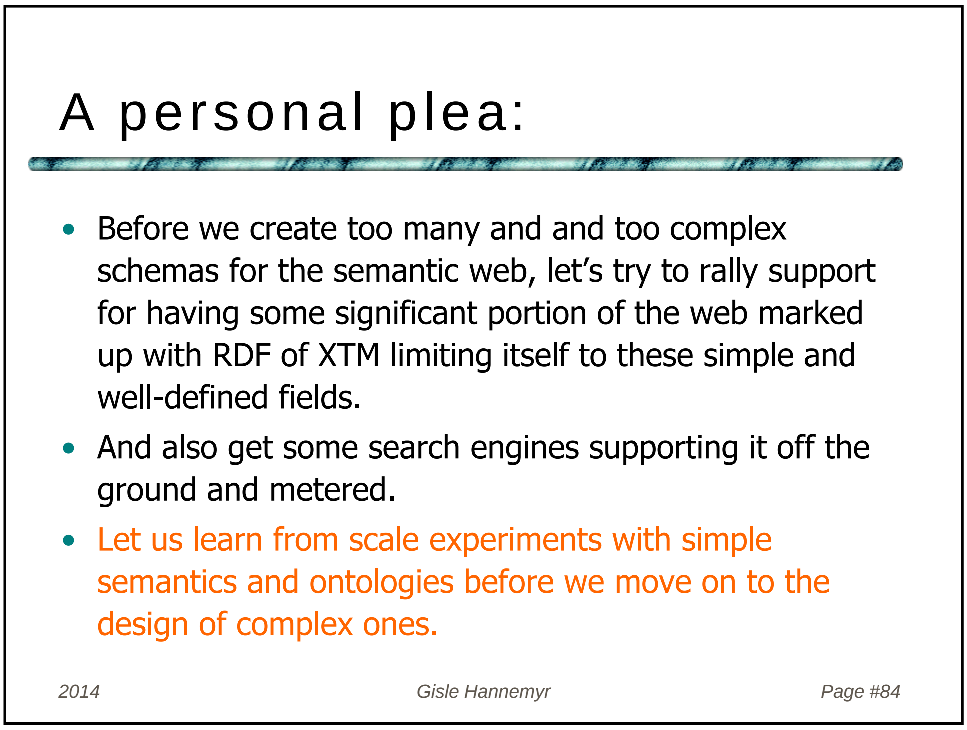
Binding Data to Metadata. Relations may be complex.





XML: eXtensible Markup Language

- W3C has defined XML as the *preferred standard* for digital document markup.
- Flexible format, large number of application areas, e.g.:
  - Elektronic Document Interchange (EDI), Web Objects, Multi-target publishing, Semantic Web
- XML is compact (44 pages), but surrounded by a large number of related technologies and derived applications (most of which we will not discuss today).



A personal plea:

- Before we create too many and and too complex schemas for the semantic web, let's try to rally support for having some significant portion of the web marked up with RDF of XTM limiting itself to these simple and well-defined fields.
- And also get some search engines supporting it off the ground and metered.
- Let us learn from scale experiments with simple semantics and ontologies before we move on to the design of complex ones.

Where are we now?

- Some very powerful frameworks for representing semantics are being created and tools supporting them are becoming available.
  - So far, we don’t see much widespread use of these frameworks on the World Wide Web.
    - We need real-world deployment of these frameworks, followed by evaluation (based upon agreed-upon metrics) on how various schemes perform.
  - The idea behind introducing metadata is not only to enable an IR Owner (e.g. the creator or publisher) to create metadata describing *own* resources but to enable *anyone* (literally) to create ontologies.
    - So far, the implications of this (as far as Digital Rights Managements, deep linking, authentication, management of ontologies, etc.) has not been explored to any extent (e.g. how do we create and autheticate webs of trust?).

Case Study:

Selection and Navigation with Search Engines – from Marcel Machill & Carsten Welp: *Wegweizer im Netz* (2003)

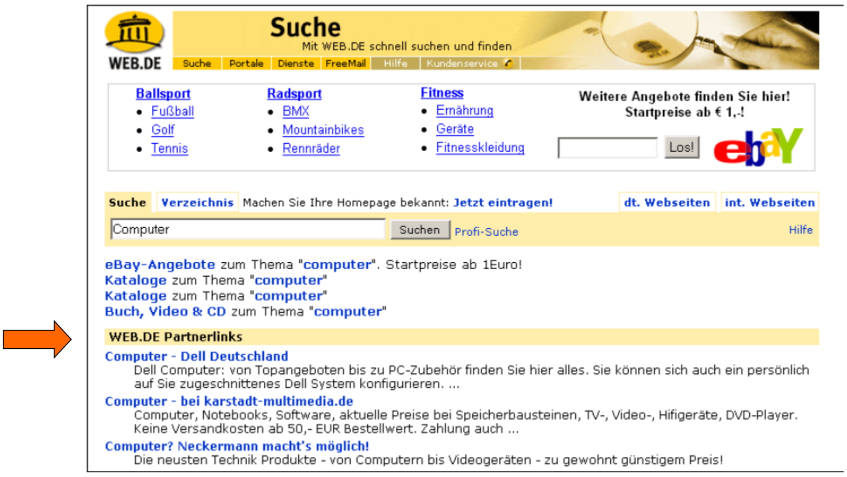
Research questions

- Competence and Search Success:
  - How successful are users with different search tasks?
  - How do they formulate their search queries?
  - How and when are search queries revised?
- Selection and Navigation:
  - How do users navigate through hit lists?
  - Which hits are activated?
  - How are hits evaluated?
  - Any differences between user groups and types of searches?
  - Do invalid, problematic or illegal hits appear?

Action:
Usability Lab Experiment

- N=150 sittings, of which 22 were children
  - Carried out between Nov. 2002 and Feb. 2003.
  - For youth and adults:
    - A search for images,
    - four research and four retrieval tasks,
    - four search tasks per sitting.
  - For children:
    - one search and one retrieval task.
  - Double video recordings (screen-cam and experimentees),
    Recording of decision-making processes.
  - Content analysis (user-oriented selectivity analysis) of navigation actions and of facial expressions/gestures.
  - Evaluation on several different levels of abstraction.

Example: Users confused by sponsored links named “partner links”


WEB.DE
Suche
Mit WEB.DE schnell suchen und finden
Suche | Portale | Dienste | FreeMail | Hilfe | Kundenservice

Ballsport
- Fußball
- Golf
- Tennis

Radsport
- BMX
- Mountainbikes
- Rennräder

Fitness
- Ernährung
- Geräte
- Fitnesskleidung

Weitere Angebote finden Sie hier!
Startpreise ab € 1,-!

Los!

Suche | **Verzeichnis** | Machen Sie Ihre Homepage bekannt: **Jetzt eintragen!** | **dt. Webseiten** | **int. Webseiten**

Computer | Suchen | Profi-Suche | Hilfe

eBay-Angebote zum Thema "**computer**". Startpreise ab 1Euro!
Kataloge zum Thema "**computer**"
Kataloge zum Thema "**computer**"
Buch, Video & CD zum Thema "**computer**"

WEB.DE Partnerlinks

Computer - Dell Deutschland
Dell Computer: von Topangeboten bis zu PC-Zubehör finden Sie hier alles. Sie können sich auch ein persönlich auf Sie zugeschnittenes Dell System konfigurieren. ...

Computer - bei karstadt-multimedia.de
Computer, Notebooks, Software, aktuelle Preise bei Speicherbausteinen, TV-, Video-, Hifigeräte, DVD-Player. Keine Versandkosten ab 50,- EUR Bestellwert. Zahlung auch ...

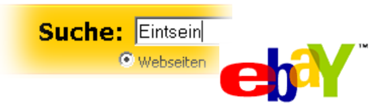
Computer? Neckermann macht's möglich!
Die neusten Technik Produkte - von Computern bis Videogeräten - zu gewohnt günstigem Preis!

Typical Errors

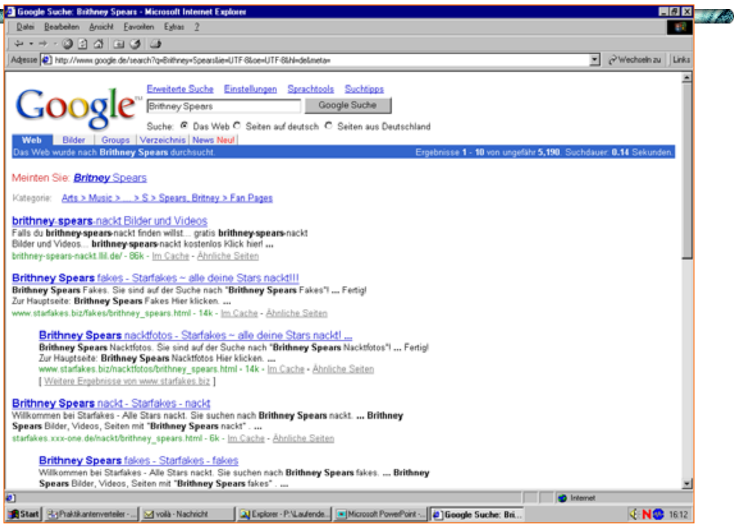
Table: Entry Errors at the Search

		Ratio of queries...		
	N	**Total**	**Novices**	**Experts**
Typing errors (sent)	123	9,2%	7,2%	10,9%
Typing errors (corrected before sending)	222	16,6%	16,4%	16,8%
Typing Errors (in spite of correction sent)	25	1,9%	1,3%	2,4%
Queries in a wrong search window (e.g. ebay)	27	2,0%	2,5%	1,6%

Query in a wrong search window


Suche: Eintsein
○ Webseiten

Typing error and protection of minors: Example: “Brithney Spears”


Google Suche: Brithney Spears - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Adresse http://www.google.de/search?q=Brithney+Spears&ie=UTF-8&oe=UTF-8&hl=de&meta=
Wechseln zu
Links
Google
Erweiterte Suche Einstellungen Sprachtools Suchtipps
Google Suche
Web Bilder Groups Verzeichnis News Neu!
Internet
Start Praktikantenverteiler - voilà - Nachricht Explorer - P:\Laufende... Microsoft PowerPoint -... Google Suche: Bri...
16:12

Brithney Spears

Suche: ● Das Web ○ Seiten auf deutsch ○ Seiten aus Deutschland

Das Web wurde nach **Brithney Spears** durchsucht. Ergebnisse **1** - **10** von ungefähr **5,190**. Suchdauer: **0.14** Sekunden.

Meinten Sie: ***Britney*** Spears

Kategorie: Arts > Music > ... > S > Spears, Britney > Fan Pages

brithney-spears-nackt Bilder und Videos
Falls du **brithney-spears**-nackt finden willst... gratis **brithney-spears**-nackt Bilder und Videos... **brithney-spears**-nackt kostenlos Klick hier! **...**
brithney-spears-nackt.llll.de/ - 86k - Im Cache - Ähnliche Seiten

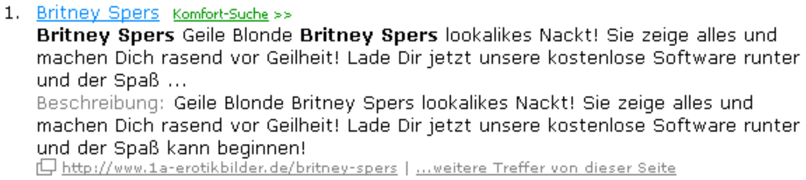
Brithney Spears fakes - Starfakes ~ alle deine Stars nackt!!!
Brithney Spears Fakes. Sie sind auf der Suche nach "**Brithney Spears** Fakes"! **...** Fertig! Zur Hauptseite: **Brithney Spears** Fakes Hier klicken. **...**
www.starfakes.biz/fakes/brithney_spears.html - 14k - Im Cache - Ähnliche Seiten

Brithney Spears nacktfotos - Starfakes ~ alle deine Stars nackt! ...
Brithney Spears Nacktfotos. Sie sind auf der Suche nach "**Brithney Spears** Nacktfotos"! **...** Fertig! Zur Hauptseite: **Brithney Spears** Nacktfotos Hier klicken. **...**
www.starfakes.biz/nacktfotos/brithney_spears.html - 14k - Im Cache - Ähnliche Seiten
[Weitere Ergebnisse von www.starfakes.biz]

Brithney Spears nackt - Starfakes - nackt
Willkommen bei Starfakes - Alle Stars nackt. Sie suchen nach **Brithney Spears** nackt. **... Brithney Spears** Bilder, Videos, Seiten mit "**Brithney Spears** nackt" . **...**
starfakes.xxx-one.de/nackt/brithney_spears.html - 6k - Im Cache - Ähnliche Seiten

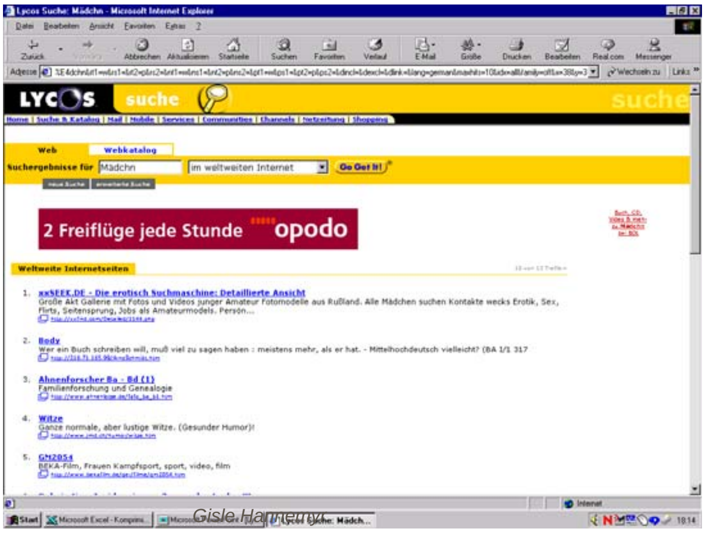
Brithney Spears fakes - Starfakes - fakes
Willkommen bei Starfakes - Alle Stars nackt. Sie suchen nach **Brithney Spears** fakes. **... Brithney Spears** Bilder, Videos, Seiten mit "**Brithney Spears** fakes" . **...**

Hit No 1 at Lycos (search item “Britney Spers”)



1. Britney Spers
Britney Spers Geile Blonde **Britney Spers** lookalikes Nackt! Sie zeige alles und machen Dich rasend vor Geilheit! Lade Dir jetzt unsere kostenlose Software runter und der Spaß ...
Beschreibung: Geile Blonde Britney Spers lookalikes Nackt! Sie zeige alles und machen Dich rasend vor Geilheit! Lade Dir jetzt unsere kostenlose Software runter und der Spaß kann beginnen!
http://www.1a-erotikbilder.de/britney-spers | ...weitere Treffer von dieser Seite

Further results prejudicial to young persons Lycos: “Mädchn“ (=“girl“), first hit



Lycos Suche: Mädchn - Microsoft Internet Explorer

Suchergebnisse für Mädchn im weltweiten Internet Go Get It!

2 Freiflüge jede Stunde opodo

Weltweite Internetseiten

1. xxSEEK.DE - Die erotisch Suchmaschine: Detaillierte Ansicht
Große Akt Gallerie mit Fotos und Videos junger Amateur Fotomodelle aus Rußland. Alle Mädchen suchen Kontakte wecks Erotik, Sex, Flirts, Seitensprung, Jobs als Amateurmodels. Person...

2. Body
Wer ein Buch schreiben will, muß viel zu sagen haben : meistens mehr, als er hat. - Mittelhochdeutsch vielleicht? (BA 1/1 317

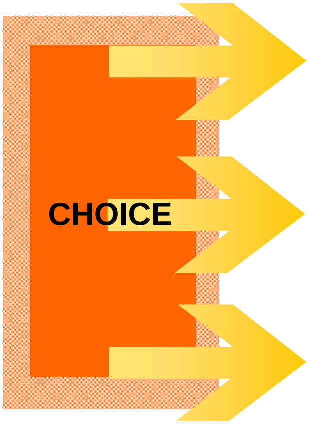
3. Ahnenforscher Ba - Bd (1)
Familienforschung und Genealogie

4. Witze
Ganze normale, aber lustige Witze. (Gesunder Humor)!

5. GM2054
BEKA-Film, Frauen Kampfsport, sport, video, film

Pornographic / erotic content

Method

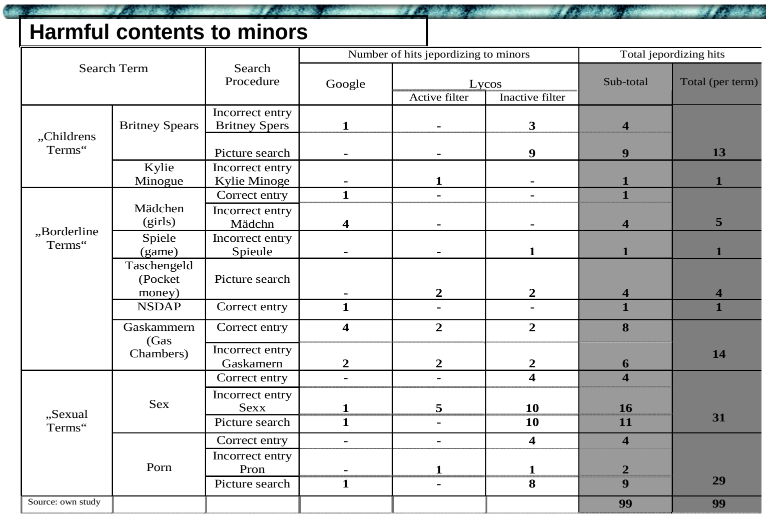


Search engines that are used to the greatest extent by young people under the age of 16: “Google” and “Lycos”.

Search terms out of three categories
1) “children’s terms“: e.g. homework, Britney Spears
2) “borderline terms“: e.g. girls, gas chambers
3) “sex terms“: Sex and Porn

Three respective four different search modi
1) correct spelling
2) misspelling
3) image search
4) in the case of ”Lycos”, the additionally activated or inactivated filter

Number of hits leading to harmful content



Harmful contents to minors

Search Term		Search Procedure	Number of hits jepordizing to minors			Total jepordizing hits	
			Google	Lycos		Sub-total	Total (per term)
				Active filter	Inactive filter		
„Childrens Terms“	Britney Spears	Incorrect entry Britney Spers	1	-	3	4	13
		Picture search	-	-	9	9	
	Kylie Minogue	Incorrect entry Kylie Minoge	-	1	-	1	1
„Borderline Terms“	Mädchen (girls)	Correct entry	1	-	-	1	5
		Incorrect entry Mädchn	4	-	-	4	
	Spiele (game)	Incorrect entry Spieule	-	-	1	1	1
	Taschengeld (Pocket money)	Picture search	-	2	2	4	4
	NSDAP	Correct entry	1	-	-	1	1
	Gaskammern (Gas Chambers)	Correct entry	4	2	2	8	14
		Incorrect entry Gaskamern	2	2	2	6	
„Sexual Terms“	Sex	Correct entry	-	-	4	4	31
		Incorrect entry Sexx	1	5	10	16	
		Picture search	1	-	10	11	
	Porn	Correct entry	-	-	4	4	29
		Incorrect entry Pron	-	1	1	2	
		Picture search	1	-	8	9	
Source: own study						99	99