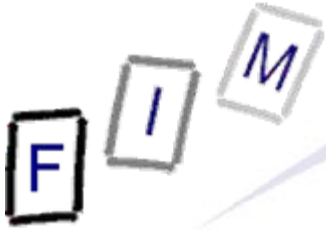
Mag. iur. Dr. techn. Michael Sonntag

Search engines

Methods, advertisements, website integration

Institute for Information Processing and
Microprocessor Technology (FIM)
Johannes Kepler University Linz, Austria

E-Mail: sonntag@fim.uni-linz.ac.at
http://www.fim.uni-linz.ac.at/staff/sonntag.htm



Questions?

Please ask immediately!

Introduction

- How search engines work
  - → Spiders/Bots
  - → Indexing
  - → Ranking
- Search engine spamming
  - → What to avoid to receive no penalties...
- Search engines for own websites
  - → Frontpage
  - → Apache Lucene
  - → Commercial software

Different "search engines"

- Crawlers: Automatically indexing the web
  - → Visits "all" reachable pages in the Internet and indexes them
- Directories: Humans look for/verify interesting pages
  - → Manual classification: Hierarchy of topics needed
  - → High quality: Everything is manually verified
    - » This takes care of a general view only (=page is on the topic it states it is about)
    - » Whether the content is legal, correct, useful, etc. is **NOT** verified!
  - → Slow: Lots of human resources required
    - » Cannot keep up with the growth of the Internet!
  - → Expensive: Because of manual visits and revisits
  - → Very important for special areas!
  - → Now almost no importance for general use
  - → Problem of modifications after review
- Mixed versions

Spiders

- They create the indices of crawler-type search engines
  - → Requires starting points: Entered by owners of webpages
  - → Visits the webpage and indexes it, extracts all links, adds the new links to the list of pages to visit
    - » Exponential growth; massively parallel; extreme internet connection required; with special care distribution possible
    - » This cannot find all links, e.g. links constructed by JavaScript are usually not found (those mentioned in JavaScript might be!)
  - → Regularly revisits all pages for changes
    - » Strategies for timespan until re-indexing exist
    - » Hashmarks/date of last change to avoid unnecessary reindexing
  - → Pages created through forms will not be visited!
    - » Spiders can only "read" ordinary pages
    - » Filling in forms is impossible (what to fill in where?)
  - → Frames and image maps can also cause problems
  - → "Just" dynamically created pages are fine, however!

"robots.txt"

- Allows administrators to forbid indexing/crawling of pages
- This is a single file for the whole server
  - → Must be in the top-level directory!
    - » Exact name: http://<domain name>/robots.txt
  - → Alternative: Specify in Meta-Tags of (each) page
- Robots.txt Format:
  - → "User-agent: " Name of robot to restrict; use "*" for all
  - → "Disallow: " Partial URL which is forbidden to visit
    - » Any URL starting with exactly this string will be omitted
      - – "Disallow: /help" forbids "/help.htm" and "/help/index.htm"
  - → "Allow: " Partial URL which may be visited
    - » Not in original standard!
  - → Visit-time, Request-rate are other new directives
- Most robots follow this standard and respect it!

Robots Meta-Tags

- Can be added into HTML pages as Meta-Tags:
  - → <META NAME="ROBOTS" CONTENT="INDEX,FOLLOW">
    - – Alternative: CONTENT="ALL"
    - » Index page, also handle all linked pages
  - → <META NAME="ROBOTS" CONTENT="NOINDEX,FOLLOW">
    - » Do not index this page, but handle all linked pages
  - → <META NAME="ROBOTS" CONTENT="INDEX,NOFOLLOW">
    - » Index this page, but do not follow any links
  - → <META NAME="ROBOTS" CONTENT="NOINDEX,NOFOLLOW">
    - – Alternative: CONTENT="NONE"
    - » Do not index this page and do not follow any links
  - → Follow: Follow the links in the page. This is not affected by the hierarchy (e.g. pages on levels deeper on the server)!
  - → Non-HTML pages must use robots.txt
    - » No "external" metadata defined!

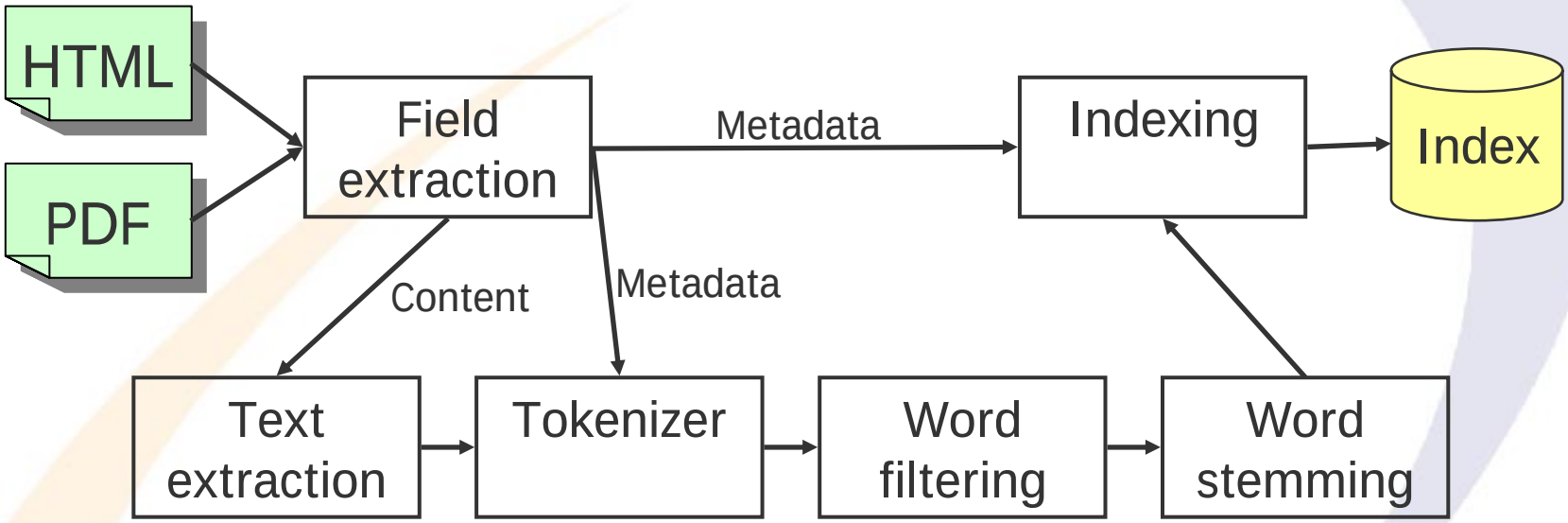
Indexing

- Indexing = Extracting some content and storing it
  - → Assigning the word(s) to the page under which it will be found later on when users are searching
- Uses similar techniques as handling actual queries
  - → Stopword lists: What words do not contribute to the meaning
    - » Examples: a, an, in, the, we, you, do, and, ...
  - → Word stemming: Creating a canonical form
    - » E.g. "words" → "word", "swimming" → "swim", ...
  - → Thesaurus: Words with identical/similar meaning; synonyms
    - » Used probably only for queries!
  - → Capitalization: Mostly ignored (content important, not writing)

Indexing

- Some search engines also index different file types
  - → E.g. Google also indexes PDF, DOC, … files
  - → Multimedia content very rarely indexed (e.g. videos???)
    - » What actually is indexed are descriptions of such content!
      - – One possibility: EXIF metadata
      - – Videos: Textual content (transcripts!)
    - » But advances in image recognition can make this possible

From text to index



- Text extraction: Retrieving the plain content text
- Tokenizer: Splitting up in individual words
- Word filtering: Stop words, lowercase
- Word stemming: Removing suffixes, different forms, etc.
- Field extraction: Identifying separate parts
  - → E.g. text vs. metadata

Page factors

- Word frequency: A word is the more important, the more often it occurs on a page
  - → Also scans for ALT tags of images and words in the URL
  - → Modified according to the location: title, headlines, text,...
    - » Higher on the page = better
  - → Clustering: How many "nearby" pages contain the same word
    - » "Website themes": Related webpages should be linked
  - → Meta-Tags: Might be used as "important", just text, or ignored
  - → Distance between words: When searching for several words
    - » "gadget creator" will match better than "creator for gadgets"
- In-link frequency: How many pages link to this page
  - → Mostly those from different domain names used only!
  - → Might also depend on keywords on that pages
  - → The most important figure currently (→Google!)

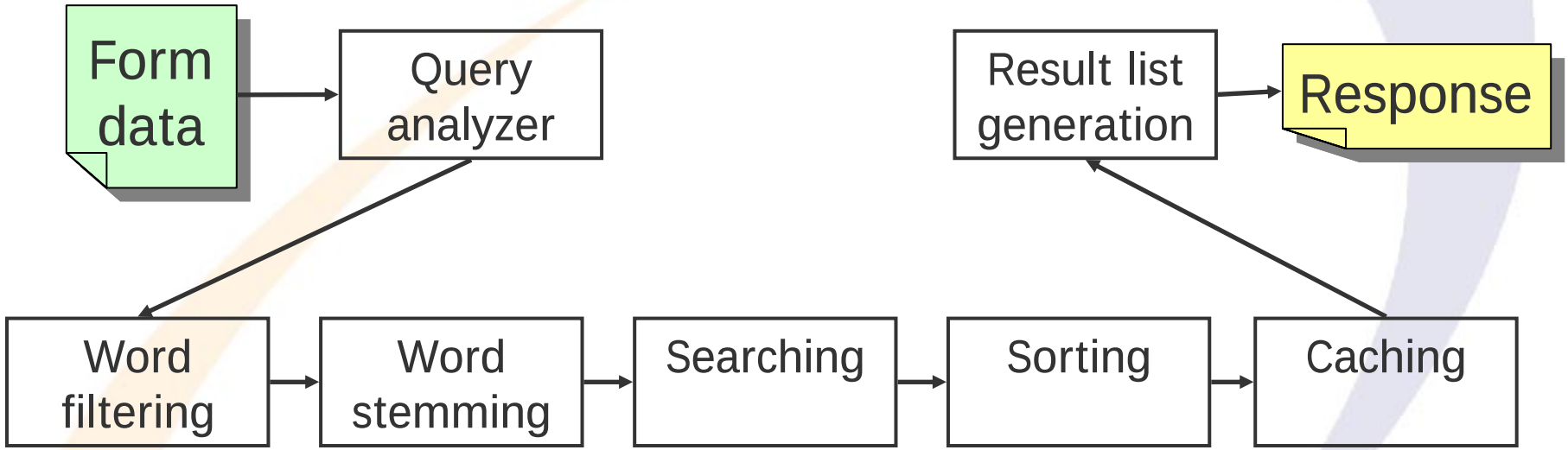
Page factors

- Page design: Load time, frames, HTML conformity, ...
  - → Some elements cannot be handled (well), e.g. Frames
  - → Size of the page (=loadtime) also has influence
  - → HTML conformity is not used directly, but if parsing is not possible or produces problems, the page might be ignored
- Visit frequency: If possible to determine (rare)
  - → How often is the site visited through the SE?
  - → How long till the user clicks on the next search result?
- Payment: Search engines also sell placement
  - → Nowadays only possible with explicit marking (as paid-for)
- Update frequency: Regular updates/changes = "live" site

Differs much between various search engines!

Avoid spamming → this reduces the page value enormously!

Searching



- Query analyzer: Breaking down into individual clauses
  - → Clause: Terms connected by AND, OR, NEAR, ...
- Word filtering: Stop words, lowercase
- Word stemming: Removing suffixes, different forms, etc.
- Caching: For next page or refined searches

Search engine spamming (1)

- Artificially trying to improve the position on the result page
  - → Important: Through unfair practices!
    - » = Deceiving the relevancy algorithm
- Pages decided to use spamming are heavily penalized or excluded completely (there is NO "appeal" procedure!)
  - → Example: The BMW website, which was removed from Google because it used a doorway page!
- Test: Would the technique be used even if there were no search engine around at all?
- Examples for spamming:
  - → Repetition of keywords: "spam, spam, spam, spam"
    - » Both after each other or just excessively
  - → Separate pages for spiders (e.g. by user agent field)
    - » They might try retrieving the page in several ways
  - → Invisible text: white (or light gray) on white
    - » Through font colour, CSS, invisible layers, ...

Search engine spamming (2)

- More spamming examples:
  - → Misusing tags: Difficulty: what is spam and what is not?
    - » noframes, noscript, longdesc,... tags for spam content
    - » DC metadata the same
  - → Very small and very long text: "Nearly" invisible!
  - → Identical pages linked to each other or mirror sites
    - » One page accessible through several URLs
    - » To create themes or as link farms (see below)
  - → Excessive submissions (submission of URLs to crawl)
    - » Be careful with submission programs!
  - → Meta refresh tags/300 error codes/JavaScript
    - » E.g. <body onMouseOver="eval(unescape('.....'))">
    - » Used to present something other to the spider (initial page) than to the user (page redirected to);if required→server side redirects
  - → Code swapping: One page for index, later change content

Search engine spamming (3)

- More spamming examples:
  - → Cloaking: Returning different pages according to domain name and/or IP of the requester
    - » IP addresses/names of search engine spiders are known
  - → Link farms: Network of pages under different domain names
    - » Sole purpose: Creating external links through heavy cross links
      - – Graph theory used to determine them (closed group of heavily interconnected sites with almost no external links)
  - → Irrelevant keywords: No connection to text (e.g. "sex")
    - » E.g. in meta tag but not in text; just to attract traffic
  - → Meta refresh tags: Automatically moving to another page
    - » Used to present something other to the spider (initial page) than to the user (page redirected to)
    - » Use server side redirects if necessary
  - → Doorway pages, machine generated page loops, WIKI links,...

Semantic Web

- The idea is to improve the web through metadata
  - → Describing the content in more detail, according to more properties, and relating them to each other
  - → Machine understandable information on the content
    - » See images and videos as example
    - » Danger of a new spam method!
- Allow searching not only for keywords, but also for media types, authors, related sites
  - → Nevertheless, some parts are already possible through "conventional" search engines!
    - » The advantage would be in better certainty
  - → The result would also be **provably** correct!
    - » But only as long as both rules and base data are correct!
- Might be useful, but is still not picking up
  - → Requires site owners to add this metadata to their pages

Commercial search engine services

- Pay for inclusion/submission: Pay to get listed
  - → Available for both search engines and directories
    - » More important for directories, however!
  - → Usually a flat fee, depending on the speed for inclusion
  - → May depend on the content; may be recurring or once
    - » E.g. Yahoo: Ordinary site: US$ 299, "Adult content": US$ 600
  - → Usually there is no content/ranking/review/... guarantee
    - » Solely that it will be processed within a short time (7 days)!
- Pay for placement: Certain placement guaranteed
  - → Now commonly paid per click: Each time a user clicks on link
    - » Previously (before dot-com crash): Pay-per-view (=displaying!)
  - → Separate from "ordinary" links: Else possibly legal problems

Pay per click (PPC)

- Advantages:
  - → Low risk: Only real services (=visits) are paid for
  - → Targeted visitors: Most campaigns match ads to search words
  - → Measurable result: Usually tracking available
    - » To determine whether the visitor actually bought something
  - → Total budget can be set
- Problems:
  - → Too much/too low success: Prediction difficult
  - → Requires exact knowledge of how much a visitor to the site is worth to allow sensible bidding on terms
  - → Click fraud: Automatic software or humans (whole offshore-centers!) do nothing but clicking on "paid for" links
    - » Affiliate programs making money through this
    - » Exhausting a competitors budget

Google AdWords

- Paid placement; will show up separately on right hand side
- Cost per Click (CPC); daily upper limit can be set
  - → CPC is variable within a user specified range
    - » If range to low it will not show up!
    - » Similar to bidding: The highest bidder will show up
      - – Low bidders will also show up, but only rarely: High CTR improves
- Ranking: Based on CPC and click-through rate (CTR)
  - → Ads not clicked on will get lower!
- Online performance reports
- Targeting by language and country possible
  - → Reduces "ad competition" and enhances click rate
  - → Negative keywords possible to avoid unwanted showings
- Possibility of showing it under (Google-determined) related keywords → Legally very dangerous!

Yahoo! Search Marketing

- Previously: Overture
  - → Powers Yahoo, MSN, Altavista, AllTheWeb, ...
- SearchSubmit: Paid inclusion (fast review and inclusion)
  - → "Quality review process": Probably by experts (good assignment of keywords/categories) and favourably
    - » List of exclusion still applies (e.g. online gambling)
  - → Pair per URL, i.e. homepage and subpages are separate
  - → Pages are re-crawled every 48 hours
  - → Positioning in result by relevance: No "moving up" or "top"!
  - → Costs: Annual subscription (US$ 49/URL; 2-10: 29, 11+: 10)
    - » Additional pay per click (US$ 0,15/0,3 / click)
- SponsoredSearch: Paid listing (sponsored results list)
  - → Position determined by bidding
  - → Pricing not available (demo: US$ 0,59; minimum: US$ 0,10)
    - » US$ 20 minimum per month in Overtures discretion (AGB) ???

Search engine integration

- Local search engine for a single site
  - → Can again be of both kinds
    - » Search engine: Special software required
      - – Automatic update (re-crawling)
      - – Configurable: Visual appearance, options, methods, ...
    - » Directory: Manual creation; no special software needed (CMS)
      - – Regular manual updates required
  - → Usually search engine is used
    - » Directory is the "normal" navigation structure
- Necessity for larger sites
  - → Difficulty: Often special requirements needed
    - » Full-text search engine for documents
    - » Special search engine for product search
    - » Special result display for forums, blogs, Wikis, ...

Features for local search engines (1)

- Language support: Word stemming, stop words, etc.
  - → Also important for user interface (search results)
  - → Stop words: Should be customizable
  - → Spell checking: For mistyped words
- File types supported: PDF, Word, multimedia files, ...
- Configurable spider: Time of day, server load, etc.
  - → Spidering through the web or on the file system level?
  - → Can password-protected pages also be crawled?
  - → Crawling of/support for personalized pages?
- Search options: Boolean search, exact search, wildcards, ...
  - → Quality of search: Difficult to assess, however!
  - → Inclusion, exclusion, "near" matches, phrase matching, synonyms, acronyms, sound matching

Features for local search engines (2)

- Admin configurability: Layout customization, user rights, definition of categories, file extensions to include, description of result items, ...
- User configurability: E.g. Results per page, history of last searches, descriptions shown, sub-searches, etc.
- Reports and statistics:
  - → Top successful queries: What users are most interested in, but cannot find easily
  - → Top unsuccessful queries: What would also be of interest
    - » Or where the search engine failed
  - → Referer: On which page they started to search for something
  - → Top URLs: Which pages are visited most through searching
- Adheres to "robots.txt" specification?

Features for local search engines (3)

- Indexing restrictions: Excluding parts form crawling/indexing
  - → Internal/private pages!
- Relevancy configuration: Weight of individual elements
  - → E.g. if everywhere good metadata is in, this can receive high priority; title tag, links, usage statistics, custom priority, etc.
- Server based, appliance, or local: Where is the engine?
- Additional features:
  - → Automatic site map: Hierarchy/links from where to where
  - → Automatic "What's new" list
  - → Highlighting: Highlight search words in result list and/or actual result pages
- "Add-ons": Free offers usually contain advertisements

Offline search engines

- Used for CD's or DVD's
  - → Cannot run a real program; no installation
- Two main variants:
  - → Run a small program, e.g. an applet (or even Tomcat!)
  - → Use JavaScript
    - » Should be preferred: Faster startup, less compatibility, version and right difficulties!
- In both cases, typically the index is created once and stored on the CD/DVD in a compressed form
  - → A lot of computing power used for spidering and indexing
  - → The result is one/several files stored on the CD
  - → Programs only parse the query, search through those file(s) and present the result

Jakarta Lucene

- Free search engine; pure Java
  - → Open source; freely available
- Features:
  - → Incremental indexing
    - » Indexing only new or changed documents; can remove deleted documents from index
  - → Searching: Boolean and phrase queries, date-range
    - » Field searching (e.g. in "title" or in "text")
    - » Fuzzy queries: Small mistypings can be ignored
  - → Universally usable: Searching for files in directories, web page searches, offline documentation, etc.
    - » No webserver needed; also possible as stand-alone
    - » Cross-platform (Java)
  - → Completely customizable
- Not a product, but a library

Jakarta Lucene: Missing features

- "Plug&Play": Installing, configuring and working site search
  - → Not available: Program needed for indexing, field definition
- Complicated search options: sound (but see "Phonetix"), synonyms, acronyms
- Spell checking not available
- No spider component
  - → Examples contain filesystem spider in basic form
    - » Problems with path differences (webserver ↔ index) possible
  - → "robots.txt" not supported
- Reports/statistics must be manually programmed
- No file types supported: Example contains HTML
  - → Word, PDF, etc. easily added, however!

Not easily deployed, but good idea for special applications!

Search Engine Optimization (SEO)

- Paid services to optimize a website for search engines
  - → Usually also includes submission to many search engines
    - » How man search engines are really of any importance today?
      - – These are very few: They can be "fed" by hand also easily!
    - » Doesn't work to well for directories: Long time without payment; important directories are small and specialized ones, which are probably not covered
  - → Often contains rank guarantees
    - » This is to be taken with very much caution: They really cannot guarantee this, therefore illegal methods or spamming is used
      - – E.g. Link farms, to provide this rank once for a short time
  - → First page/<50: Are users still convinced that everything important is there?
    - » Many can do without top listing
      - – Especially on very generic terms; focus on more specific ones!

Promoting your website
My thoughts

- Avoid all pitfalls and search engine spamming
  - → Use tools to verify suitability of webpages
  - → Keywords, keyword density; HTML verification; style guides;...
- Focus on specific terms and phrases as keywords
  - → Avoid "sex", "car", "computers"; use "used porsche cars", ...
- Provide valuable information to visitors
  - → FAQ, hints, comparisons, ...: This will get you external links
- Submit to the 5 top most crawlers and the 10 most important (for your business!) directories
- Wait a long time to get listed before re-submitting
  - → Currently some engines take up to two month
- Do not depend on visitors: Focus on customers!
  - → E.g. avoid ad selling
  - → Use other advertisement avenues too, if possible

Legal aspects

- Search engine "optimization" can lead to legal problems
- Examples are:
  - → Very general terms, e.g. "divorceattorney.com"
    - » Channeling users away from competitors
    - » Decisions mixed: Sometimes allowed (no dominant position), and sometimes not
      - – Depends also on content: Does it claim to be "the only one"?
  - → Unrelated words in meta-tags
    - » E.g. using "legal studies" for a website selling robes for judges
      - – Probably allowed as long as no channeling takes place
  - → Trademarks: Using "foreign" trademarks in meta-tags
  - → Liability for links to illegal content
  - → Search engines and copyright

Discussion according to EU (and Austrian) law!

Trademarks

- Using trademarks, service marks, common law marks, etc. of competitors on the own site (or in own meta-tags)
  - » This applies only to commercial websites
  - → Depends on whether there is a legally acceptable reason for inclusion on the webpages
    - » Example: Allowed product comparison, other suppliers, selling those products, ...
  - → In general this is dangerous/forbidden
    - » Trademark law: Illegal "use" of a mark
    - » Competition law: "Freeriding" on others fame
    - » Competition law: Customer diversion
- Search engines: Results may contain the trademark
  - → Can even contain links to infringing sites; see below
- Legal status not completely decided
  - → But German BGH: Foreign trademarks are forbidden

Search engine liability

- Liability for a link itself depends on several elements
  - → Directories provide pre-verification
    - » See link liability below!
  - → Paid for links are selected; depending on selling method knowledge may or may not exist by default
    - » One of the exceptions from the no-liability clause:
      - – "Selection or modification of content"
    - » Hosting: Full liability as soon as knowledge of illegal content is present; illegality must be clear even for laymen
  - → Automatically gathered and presented links are privileged
    - » Only for foreign content (site search engines → full liability)
    - » This even applies if knowledge of illegal content is present!
      - – Exception: Knowingly furthering the illegal activity (Austria!)
- No obligation to actively search for anything illegal!

Link liability

- General liability of links; also applies to directories
  - → Any time a link is consciously created
- Liability by default if the content linked to is "integrated"
  - → = made to own content (link just to avoid re-typing it)
  - → Depends on the context of the link
- No liability for foreign information if
  - → no knowledge of illegality of content
    - » If something illegal is in a directory, knowledge will exist through pre-approval process!
    - » "Illegality" must be obvious to laymen; suspicions insufficient
  - → and if knowledge attained, immediate steps taken to remove link
    - » This will be quite fast as the Internet is a fast medium!
- No obligation to actively search for anything illegal!
  - → Germany: On "controversial" topics, screening necessary

Search engines and copyright
Image thumbnail archives

- Google also indexes images
  - → This in itself is no problem
- For preview, a smaller version of the image is created and stored locally; this is shown alongside the link
  - → Reducing an image's size is a modification of the original picture and therefore requires the owners permission
    - » Unless the content is no longer discernible
  - → Therefore this practice is forbidden!
- Similarity: Extracting the description of the webpage and showing it with the link
  - → This is however a privileged use ("small citation")
    - » Additionally this is the intended use for the meta tag!
- Currently ongoing litigation; final decision pending!

http://www.jurpc.de/rechtspr/20040146.pdf

Literature

- Robot Exclusion Standard:
  http://www.conman.org/people/spc/robots2.html
  http://www.robotstxt.org/
- Jared M. Spool: Why On-Site Searching Stinks
  http://www.uie.com/articles/search_stinks/