Ambrosianum COLLEGE

# Einführung in die Internetrecherche

## Skript – 4. Sitzung:

## *Internetsuche, Suchmaschinen*

**Stand:** 01.02.2024



## Lernziele dieser Sitzung:

- Grundlagen der Internetsuche kennen
- den Begriff des "invisible web" kennen und Strategien kennen, diesen Effekt zu mindern
- Performanz verschiedener Suchmaschinen kennen; Teil 1 Google

## *1. Internetsuche allgemein*

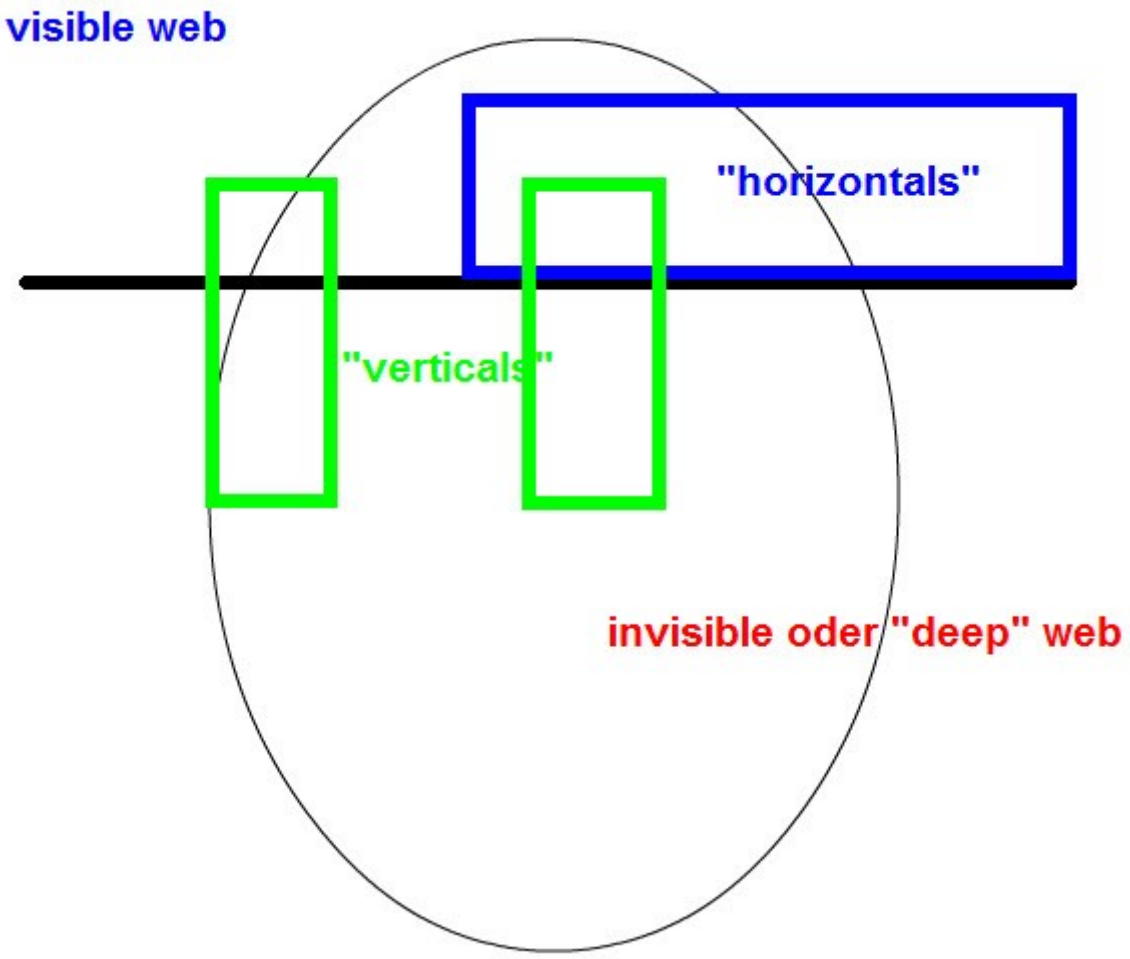
Die relevanten Suchdienste für die Internetsuche sind Suchmaschinen. Linksammlungen und Indices hingegen bieten nur selten eine gute Hilfe und werden deshalb im Kurs nur knapp (s. weiter unten) behandelt. Vergleicht man Suchmaschinen mit den bisher behandelten Suchdiensten (Kataloge und Datenbanken), dann ist das Internet schlecht erschlossen, da Suchmaschinen lediglich eine Stichwortsuche bieten. Man merkt das nur nicht, weil der Datenfundus so groß ist, dass (fast) immer etwas herauskommt. Und ein guter Algorithmus, der bei Suchmaschinen für ein gutes Ranking sorgt, wirkt in die gleiche Richtung. Aber im Sinne einer gezielten Suche, die alles Relevante aus dem Datenfundus herauszuholen vermag, ist die Effizienz einer Internetsuche leider von schlechter Qualität.
Man kann das durch Spezialsuchinstrumente verbessern, ganz ausgleichen kann man es nicht!

Bei der Internetsuche ist oft vom „visible“ und „invisible Web“ die Rede. Das sichtbare Netz ist dasjenige, das durch Suchmaschinen recherchierbar ist, wohingegen das unsichtbare Netz Inhalte enthält, die nicht oder kaum erschlossen sind. Es ist kaum zugänglich,

- weil die Robots der Suchmaschinen durch die robots.txt-Datei ausgeschlossen sind, beispielsweise bei Aggregatoren oder kommerziellen Seiten,
- weil die Inhalte lizensiert sind und nur per Passwort oder IP-Range zugänglich sind,
- weil rechtliche Ausschlüsse (Urheberrecht, Leistungsschutzrecht, "Recht auf Vergessen") bestehen,
- weil die Inhalte hinter User-Interfaces liegen und von dort aus erst aufgerufen werden müssen (Kataloge, Datenbanken etc.)

- weil Admins Fehler gemacht haben und Seiten beispielsweise unverbunden für sich bestehen,
- weil Inhalte in anderen Formaten als Text/HTML enthalten sind: Bilder, Videos, Flash etc.
- weil Inhalte dynamisch generiert werden oder
- weil Inhalte aktualisiert wurden oder Echtzeitinhalte sind.

Aber auch das sichtbare Netz ist unterschiedlich gut erschlossen, da nur ein Kern an viel besuchten und gut vernetzen Seiten gut erschlossen ist, ebenso Inhalte, auf die von diesem Kern aus verwiesen wird. Allerdings existieren auch Seiten, die nur auf den Kern verweisen, aber auf die nicht verwiesen wird, weder vom Kern noch von den anderen, von Suchmaschinen erschlossenen Seiten. Dieser Teil des Netzes ist schlecht erschlossen!




Suchmaschinen, die den sichtbaren Teil erschließen, werden „Horizontals“ genannt, wohingegen Spezialsuchmaschinen „Verticals“ genannt werden.

## 2. Allgemeine Suchmaschinen

Was man mit dem Wort "Suchmaschinen" bezeichnet, ist im Grunde viererlei: erstens sogenannte "robots", oder "spider", Programme, welche im Internet nach Dokumenten suchen und diese in geeigneter Weise indexieren, zweitens riesige Datenbanken, in denen die Indexierungen von WWW-Dokumenten und diese selbst gespeichert sind, ein „Indexer“, welcher die Dokumente erschließt und aufbereitet und viertens eine Software, die Abfragen in bestimmter Struktur erlaubt und Ergebnisse in einem bestimmten Ranking ausgibt.

Wie umfangreich der Datenbestand einer solchen Suchmaschine ist, hängt natürlich von der Anzahl der indexierten Seiten ab. Es gibt keine Suchmaschine, die *alles* indexiert hat. Außerdem sind die Datenbestände der Suchmaschinen nicht gleich. Daher sollten Sie unbedingt nicht eine, sondern mehrere oder auch so genannte "Meta-Suchmaschinen" benutzen, die mehrere Suchmaschinen simultan abfragen.
In manchen Gebieten (z.B. Wissenschaft, Zeitungsartikel, Bilder) gibt es auch *Spezialsuchmaschinen (s.o., „horizontals“)*, die einen spezifischen Datenbestand besser erschließen als die normalen Suchmaschinen. Solche Spezialsuchen sind mittlerweile auch als Option in die "großen" Suchmaschinen eingegliedert.

Suchmaschinen unterscheiden sich aber nicht nur in Bezug auf den Datenbestand, sondern auch in Bezug auf die Recherchemöglichkeiten, die sie bieten, sowie die Art der Ausgabe der Ergebnisse. Wenn eine Suchmaschine zwar einen guten Datenbestand hat, dann aber im Ranking versagt und Ihnen bei den ersten Auswahlmöglichkeiten nur Schrott und Werbung bietet, haben Sie nichts davon!

Suchmaschinen erschließen in aller Regel *Texte,*

- entweder Texte an sich (Volltexte) oder
- Metadaten von anderen Dateitypen.

Suchmaschinen bestehen aus

- Crawler/Spider/Robot
- Indexer
- Repräsentation und
- Searcher

Suchmaschinen erschließen

- den Text des Dokuments
- dem Dokument beigegebene Metadaten
- aus dem Dokument extrahierte Metadaten
- Metadaten aus der Webseite des Dokuments
- Metadaten aus dem Web (z.B. Page Rank)
- Nutzer-Suchverhalten (Personalisierung)!

In das Ranking fließt ein

- textspezifische Faktoren
- Popularität
- Aktualität
- Standort (Lokalisierung)
- Personalisierung
- technische Faktoren

Erkenntnisse über das Nutzerverhalten kommen

- via Toolbars
- via eigene Browser (Chrome!)
- via Personalisierungstools (Cookies, Google-Accounts etc.)
- via Analysedienste

Spezialsuchmaschinen (z.T. spezialisierte Crawler) erschließen

- besondere Arten von Inhalten und
- besondere Typen von Inhalten

***Beachten Sie bitte:*** *Eine Suche mittels Suchmaschinen bedeutet in aller Regel: Stichwortsuche!* Eine Suche nach Schlagwörtern können nur Kataloge und Datenbanken bieten.

Dies bedeutet, dass man bei einer Suche die verschiedenen grammatikalischen Formen (und, falls notwendig: auch in verschiedenen Sprachen) berücksichtigen muss. Eine Suche nach dem Thema "Frauen in Lateinamerika" beispielsweise müsste in verschiedenen Suchschritten und Verknüpfungen die Suche nach folgenden Begriffen beinhalten: "frau", "frauen" (ggf. durch Platzhalter oder stemming zusammenlegbar), "woman", "women", "mujer", "mujeres" (dito), "mulher", "mulheres", "Lateinamerika", "latin", "america", "latinoamérica", ...

Die meisten Suchmaschinen ermöglichen eine Suche mittels Verknüpfungen. Voreingestellt ist meist, wenn man zwei Begriffe eingibt, dass nach Ihnen mit einer ODER-Verknüpfung gesucht wird. Mit "+" sucht man meist nach einer UND-Verknüpfung (Ausnahme: Google, hier bitte das Wort mit „“ einschließen) und mit "-" schließt man das entsprechende Element aus (NICHT-Verknüpfung).

*Was* bei der einzelnen Suchmaschine möglich ist, erfahren Sie über die meist recht versteckt angebotene "erweiterte Suche", welche eine Hilfestellung bei der Eingrenzung der Suchfrage gibt (mittlerweile ist die erweiterte Suche meist erst nach einer Anfangsrecherche erreichbar) oder lesen Sie hierzu die Hilfetexte der entsprechenden Suchmaschine.

> **Tipp:** Wenn man die erweiterte Suche geladen hat, kann man in der Webadresse die aktuelle Suchanfrage bis einschließlich dem Fragezeichen herauslöschen, dann die Seite durch Drücken von „Return“ nochmal laden (es müsste eine leere erweiterte Suche angezeigt werden) – und dann können Sie die erweiterte Suche bookmarken!

Ambrosianum COLLEGE

Die Ergebnisse einer Suche in Suchmaschinen werden meist in einer "Ranking"-Liste gezeigt. Bitte haben Sie keine Scheu, wenn mehr als 120.000 Ergebnisse angezeigt werden, Sie sollten nur die ersten Seiten durchsehen, ob dort etwas Relevantes aufgeführt ist. Meist werden die Nennungen nämlich gewichtet, so dass relevantere Ergebnisse zuerst angezeigt werden (manchmal sogar mit einem Prozentzeichen, inwieweit das Ergebnis der Anfrage "entspricht", nun na).

Worin unterscheiden sich Suchmaschinen? Da ist zunächst einmal die Qualität des Rankings, die beispielsweise bei Google oft besonders hoch ist. Dann ist es die Transparenz der Darstellung der Ergebnisse: Manche Suchmaschinen stellen die Ergebnisse viel strukturierter und transparenter dar als andere. Weiter sind Suchmaschinen zu bevorzugen, welche Ihnen Möglichkeiten an die Hand geben, die Suchmenge weiter einzuschränken mit Hilfe von "refine your search", indem Stichworte angeboten werden, die dies erlauben. Für die wissenschaftliche Suche ist Bing (bzw. eine der anderen Suchmaschinen, die den Bing/Yahoo-Index nutzen) noch von besonderer Bedeutung, weil es eine Suche mit Hilfe Boole'scher Operatoren (auch Klammerungen) erlauben. Exalead bietet auch einen NEAR-Operator. Gerne wird auch DuckDuckGo http://www.duckduckgo.com/ ausgewählt, eine Suchmaschine mit Bing/Yahoo!-Index, aber netten AddOns, wie z.B. Datenschutz und Icons mit Wiedererkennungswert.

Derzeit sinnvolle Suchmaschinen sind:

- **Google:** Derzeit die leistungsfähigste Suchmaschine, die personalisiert nutzbar, aber auch anpassbar ist. Wenn man englischsprachige Webergebnisse recherchieren will, zuerst suchen, dann oben rechts das Zahnrad aufrufen, bei der erweiterten Suche umstellen (klappt nicht immer). Suchwörter werden auf den Wortstamm zurückgeführt. Mit Hilfe von Suchaspekten kann man die Ergebnisse sehr gut spezifischer erzielen. Übersetzte Suche ist bei Google möglich mit Hilfe von 2lingual http://www.2lingual.com/
- Google anonymisiert suchen: **Startpage** https://www.startpage.com/
  Yahoo!-Index:
- **DuckDuckGo** (Yahoo/Bing Index) https://duckduckgo.com/, evtl. zuerst auf Deutsch umstellen. Die Suchmaschine trackt nicht (= Dateschutz) und bietet Icons in der Ergebnismenge zur Wiedererkennung an.
- **Ecosia** ((Yahoo/Bing Index) http://www.ecosia.org/?c=de
- **Bing** http://www.bing.com/ (Operatoren, auch Boole'sche Operatoren. Near-Operator zurzeit inaktiv)
- **Brave Search** https://search.brave.com/

Tipps und Tricks:
*Alle Suchmaschinen*
**“ “:** Phrasensuche
**+:** Erzwingen eines Begriffes in der Ergebnismenge (außer Google, dort Begriff in „“ setzen)
**intitle:** sucht im Titel von Webseiten.
**inurl:** sucht in der URL einer Webseite.

**link:** findet Webseiten, die einen Link auf die angegebene Webadresse beinhalten (sehr unzuverlässig). Alternative wäre https://openlinkprofiler.org/
**domain:** sucht nach gewünschter Top-Level-Domain (z.B. .de, .eu oder .int).
**site:** sucht nur auf der angegebenen Webseite (z.B. site:http://europa.eu +asyl)
Nach speziellen Dokumenttypen suchen:
**filetype:** sucht bestimmte Dokumenttypen (pdf, ppt, doc etc.)

*Zusätzliche Suchkommandos für Google*
**“ “**: Erzwingen eines Begriffes in der Ergebnismenge
**“ * “** ersetzt ein oder mehrere Wörter zwischen zwei Begriffen (unzuverlässig).
**500..600** sucht nach von..bis, z.B. Jahreszahlen
**allintext:** alle Wörter müssen im Text der Seite enthalten sein
**intext:** mindestens eines der Wörter muss im Text der Seite enthalten sein
**allintitle:** alle Wörter müssen im Titel der Seite enthalten sein
**allinurl**: alle Wörter müssen in der URL enthalten sein
**cache:** sucht die bei Google gespeicherte Version einer Seite (cache:www.uni-tuebingen.de/pol)
**info:** sucht nach Informationen über eine Seite (or id:) (info:europa.eu)
**related:** sucht nach Seiten, die ähnlich sind

Die Strategie der Suche mit Hilfe von Suchmaschinen sollte sein, möglichst vollständige und möglichst qualitätvolle Ergebnisse zu bekommen. Deshalb sollten *mehr als eine* Suchmaschine verwendet und die Ergebnisse verglichen werden. Nur so bekommen Sie ein Gefühl dafür, ob und wie sich die Ergebnismengen in der Qualität unterscheiden und verändern.

Die Suche mittels **Metasuchmaschinen** ist zu empfehlen, wenn Sie *mehrere* Suchmaschinen simultan recherchieren möchten. So durchsuchen Sie größere Datenbestände und sparen Zeit. Sie sollten aber darauf achten, dass die „großen“ Suchmaschinen alle berücksichtigt sind, also Google und Yahoo!. Der Nachteil ist manchmal, dass man seine Suchanfrage nicht so detailliert stellen kann wie bei einzelnen Suchmaschinen.

Empfehlenswerte Metasuchmaschinen sind

- **eTools.ch** (http://etools.ch/),
- Carrot Search https://search.carrot2.org/#/web und
- bedingt - MetaGer (http://www.metager.de/).