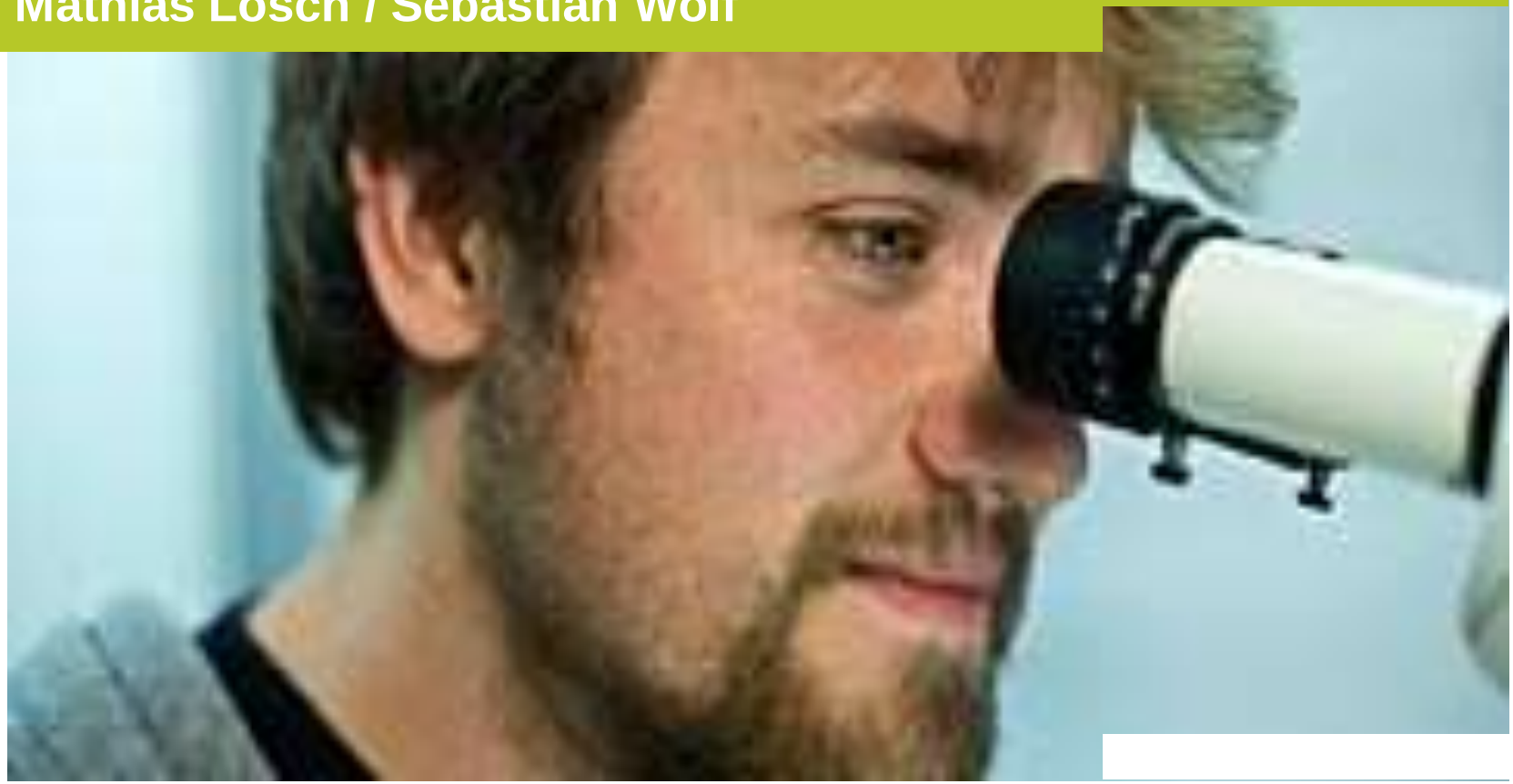
Universität Bielefeld Universitätsbibliothek

INFORMATION.plus!

Grundlagen der Suchtechnologie

Kolloquium Wissensinfrastruktur, 28.05.2010

Mathias Lösch / Sebastian Wolf



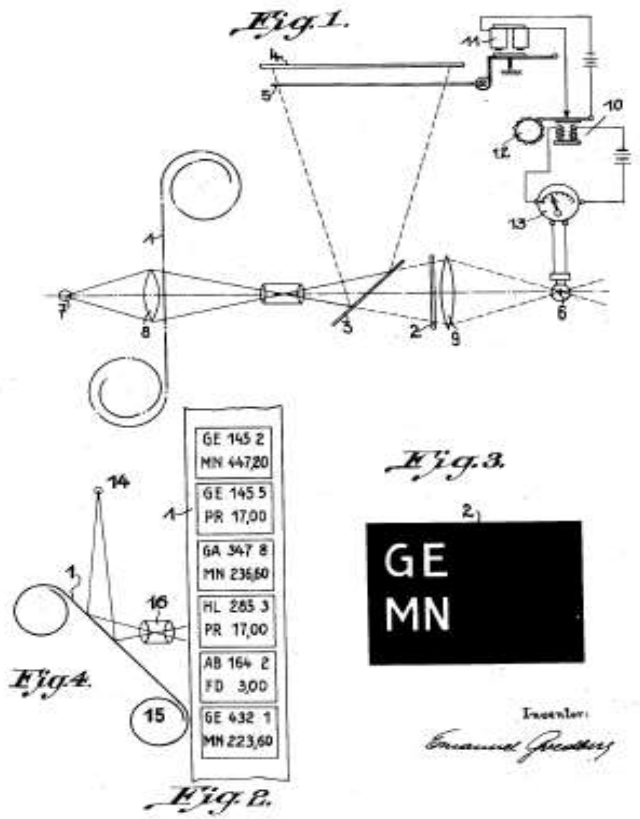
Inhalt

- Historie
- Indexierung
- Suchtechnologien
- Ranking-Verfahren
- Suchmaschinenoptimierung
- Ausblick

Historie

Register / Index (1614)
Antonio Zara (1574-1621), Bischof von Petina, fügte seiner Enzyklopädie *Anatomia ingeniorum et scientiarum* erstmals einen umfangreichen Index an

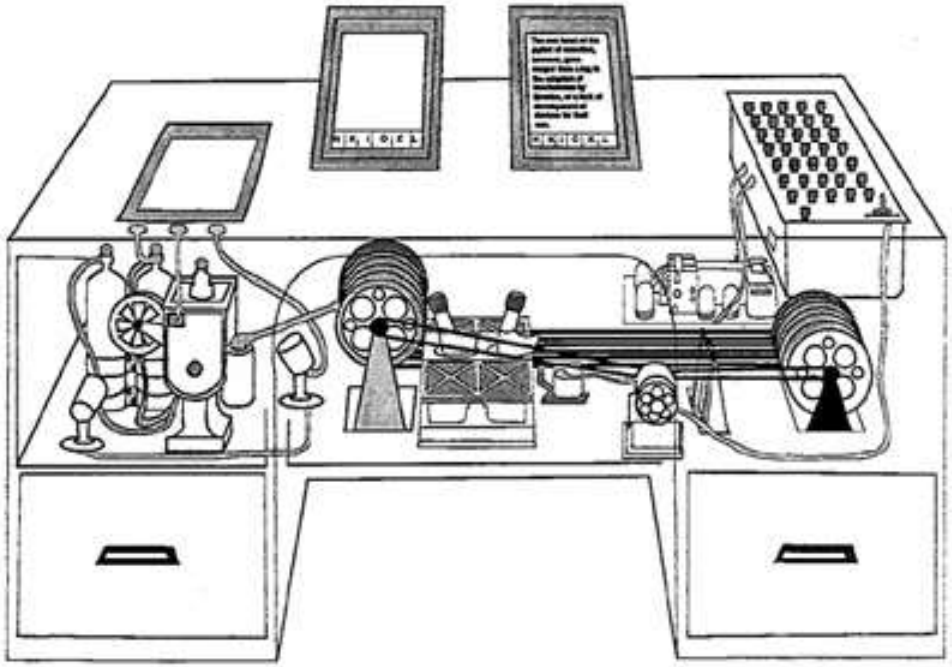
Historie



Statistische Maschine (1931)
Sollte Metadaten auf Rollen von Mikrofilm mit Hilfe von Fotozellen und Mustererkennung durchsuchen können. Vorgestellt von Emanuel Goldberg 1931 auf dem „VIII. Internationalen Kongreß für wissenschaftliche und angewandte Photographie“

Quelle: http://www.freepatentsonline.com/1838389.pdf

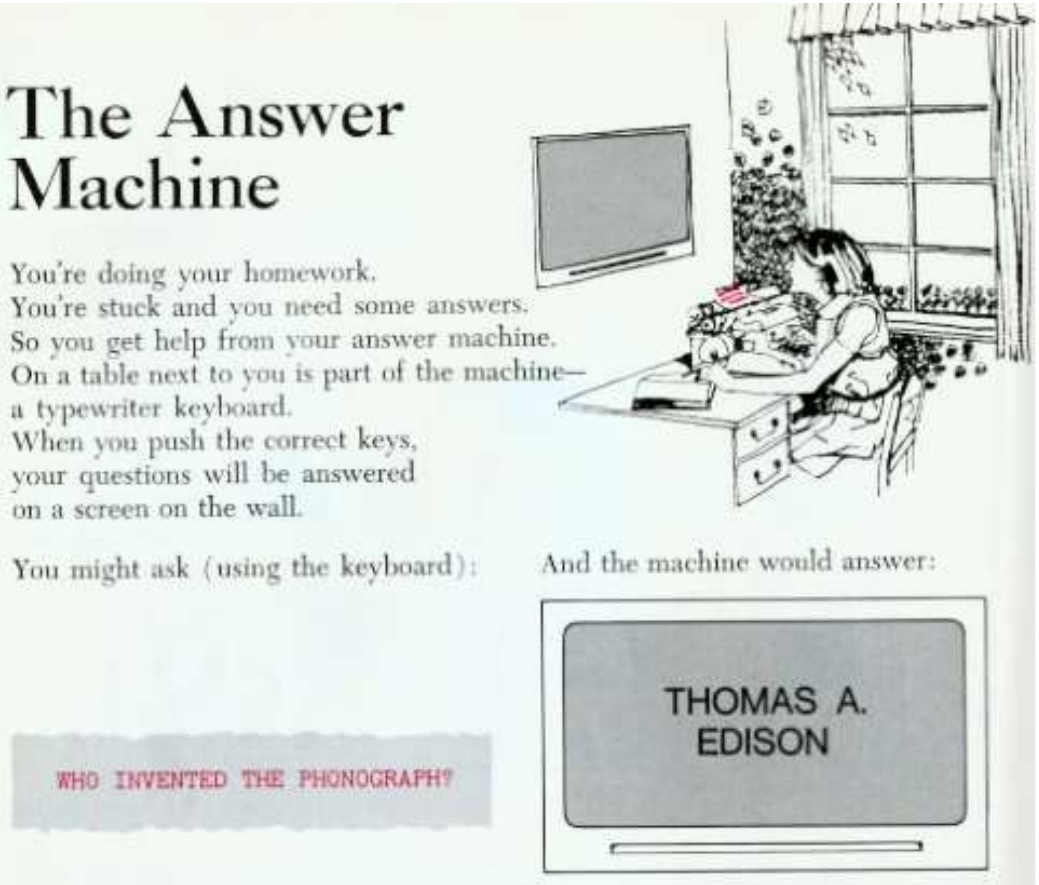
Historie



Quelle: https://atlas.colorado.edu/~hofmocke/digitalpoetry/memex.html

Memex (1945)
Als Wissensfindungs- und Verwertungssystem konzipierter Kompakt-Analog-Rechner, der 1945 von Vannevar Bush im Artikel „As We May Think“ fiktiv vorgestellt wurde

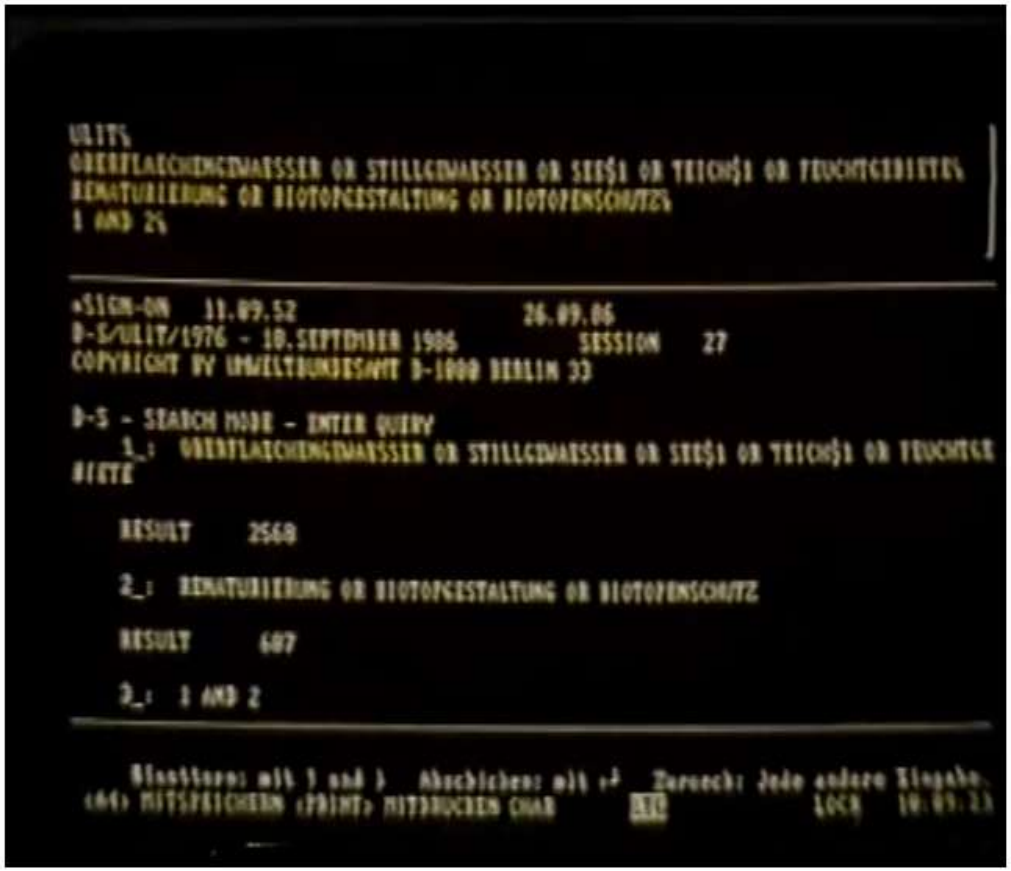
Historie



Quelle: http://eduspaces.net/csessums/weblog/179458.html

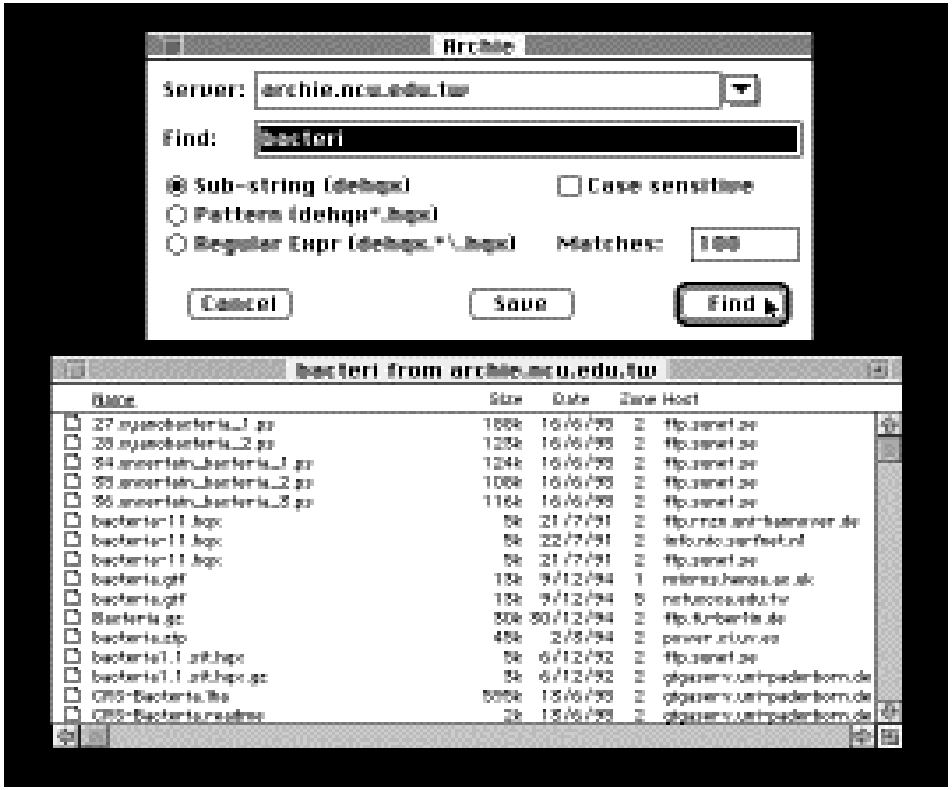
The Answer Machine (1964)
Theoretischer Entwurf einer Suchmaschine (in: Childcraft Vol. 6 How Things Change)

Historie



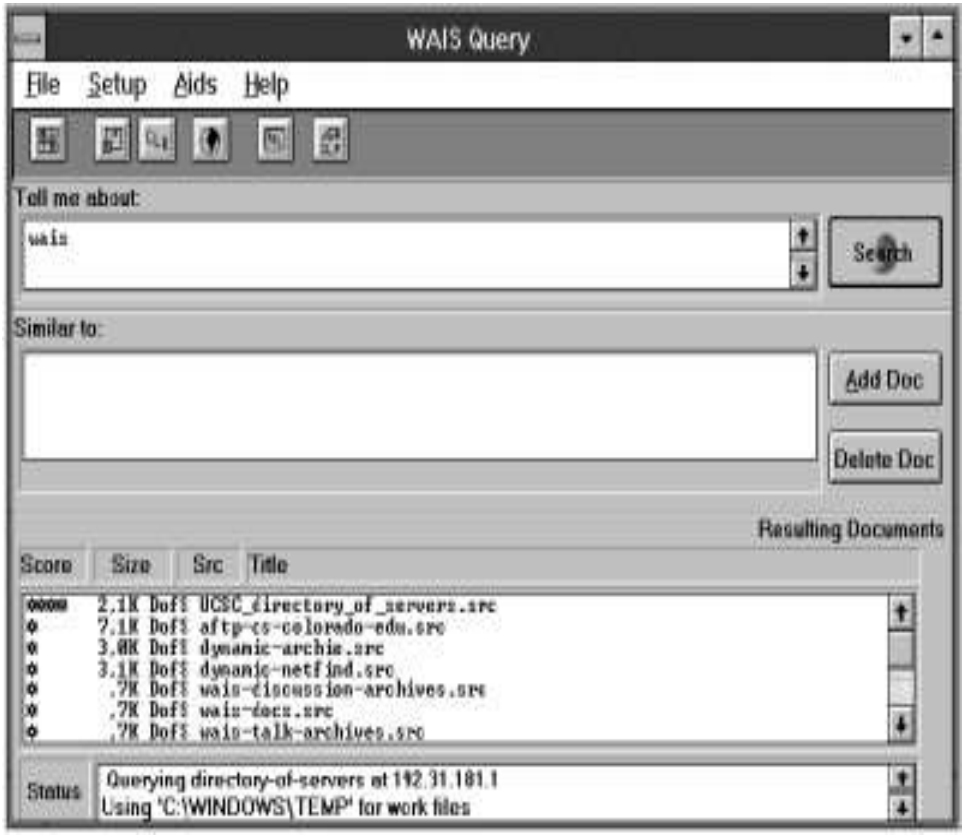
DIALOG (1972)
Der Datenbankanbieter Dialog stellte 1972 mit dem DIALOG Information Retrieval Service das weltweit erste kommerzielle Angebot mit Zugriff auf eine Online-Datenbank vor.

Historie



Archie (1990)
Archie war eine Suchmaschine, die speziell für das Indizieren von FTP-Archiven entwickelt wurde. Gesucht werden konnte nach (Teilen von) Dateinamen.

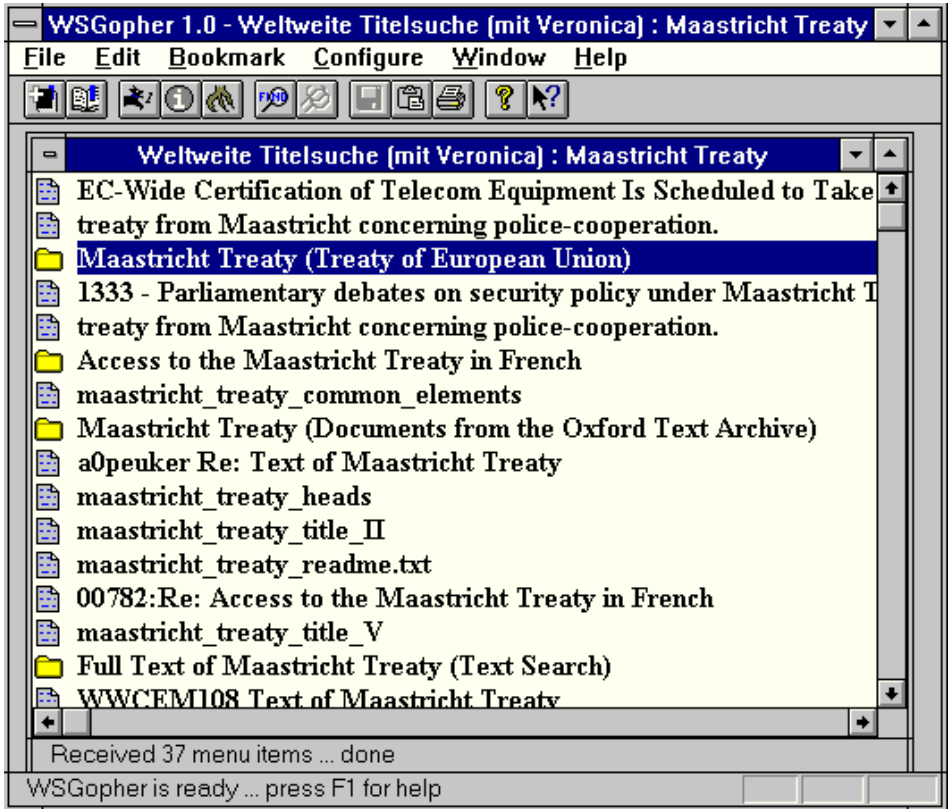
Historie



WAIS (1991)

„Wide Area Information Server System“. Innerhalb des WAIS-Systems wurden Dateien gelistet, die im Internet erreichbar sind. WAIS ermöglicht erstmals eine Volltextsuche innerhalb des Datenbestandes.

Historie



Veronica (1992)
„Very Easy Rodent-Oriented Netwide Index to Computerized Archives“. Suchdienst für „Gopher“-Webseiten (Titel bzw. Dateinamen).

Historie

Webcrawler (1994)
Erste Volltext-Internet-Suchmaschine, die für die für die Ergebnisanzeige eine Relevanzbewertung (Ranking) vornahm

Historie

Google (1998)
Führte eine neue Form des Rankings (Link-Popularität) ein

Indexierung

Dieser Teil ist als separate Präsentation abrufbar:

Starten

Suchtechnologien

- Abgleich von Zeichenketten
- Reguläre Ausdrücke
- „Fuzzy“-Suche
- Phonetische Suche
- Semantische Suche

Suchtechnologien: Abgleich von Zeichenketten

- Das gesuchte Wort wird mit dem im Index verglichen
- Über die Phrasensuche können Wortgruppen abgeglichen werden („Universität Bielefeld“)
- Bei der Eingabe mehrerer Wörter, wird i.d.R. mit „UND“ verknüpft (teilweise mit „ODER“)
- Weitere Verknüpfungsmöglichkeit z.B. durch Verwendung Boolescher Operatoren (AND, OR, NOT)

Suchtechnologien:
Reguläre Ausdrücke

- Ersetzen eines oder beliebig vieler Zeichen bzw. eines Zeichenbereichs (a-z, 0-9, f-q, B oder D, Leerzeichen, Tabulatoren oder anderer Steuerzeichen)
- Für große Datenmengen ungeeignet (Suchzeiten)
- Suchmaschinen beherrschen (wenn überhaupt) nur das „Wildcard“-Zeichen am Wortende (Trunkierung; ersetzt beliebig viele Zeichen)

Suchtechnologien: Fuzzy-Suche

- Normierung von Umlauten / Sonderzeichen (ä = ae, á = a)
- Rechtschreibkontrolle anhand eines Wörterbuchs (Eifelturm = Eiffelturm)
- Ermittlung von Synonymen anhand eines Wörterbuchs (Fernsehgerät = Fernseher)
- Unscharfe-UND-Verknüpfung („Fuzzy-AND-Search“)

Suchtechnologien: Fuzzy-Suche

- Stemming / Lemmatisierung = verschiedene morphologische Varianten eines Wortes auf ihren gemeinsamen Wortstamm zurückgeführt
- Numerus (Bibliotheken = Bibliothek)
- Flexion (Hauses = Haus / schneller = schnell / geworden = werden)
- Komposita (Fußballstadion = Fußball + Stadion)

Suchtechnologien: Fuzzy-Suche

- „Unscharfer“ Abgleich des Suchbegriffs (Levenshtein-Distanz / N-Gramm-Analyse)
- Levenshtein-Distanz = minimale Anzahl von Einfüge-, Lösch- und Ersetz-Operationen um die erste Zeichenkette in die zweite umzuwandeln
- Beispiel: Die Begriffe Tier und Tor haben eine Levenshtein-Distanz von 2:
  Tier → Toer (Ersetze i durch o) → Tor (Lösche e)

Suchtechnologien: Fuzzy-Suche

- N-Gramm-Analyse = Berechnung der Wahrscheinlichkeit, dass auf eine bestimmte Buchstaben- oder Wortreihenfolge ein bestimmter Buchstabe oder ein bestimmtes Wort folgen wird
- Durch Abgleich mit einem Referenzdokument lassen sich Cluster bilden
- Je näher ein Dokument am Referenzdokument liegt, umso wahrscheinlicher ist, dass sich der Inhalt um dessen Thema dreht

Suchtechnologien: Phonetische Suche

- Gleichklingende Wörter werden mit einer identischen Zeichenfolge kodiert
- Bei einer Suche nach einem Wort, wird auch das ähnlich klingende Wort gefunden (oder als Suchbegriff vorgeschlagen)

Suchtechnologien:
Phonetische Suche (Beispiel „Soundex“)

- Ähnliche Laute besitzen den gleichen Code (im englischen werden z.B. B, F, P und V mit der Ziffer „1“ codiert)
- Vokale, Umlaute und bestimmte Konsonanten werden ignoriert
- Der Anfangsbuchstabe des Wortes bleibt erhalten
- Neben dem Anfangsbuchstaben werden maximal 3 Konsonanten kodiert, ggf. wird mit 0 aufgefüllt

Suchtechnologien:
Phonetische Suche (Beispiel „Soundex“)

- Ergebnisse sind oft zielführend, aber manchmal auch irreführend oder unsinnig
- **G**oo**g**l**e** = **G240**
- **G**oo**g**l**e**bombe = **G241**
- **G**u**g**elhu**p**f = **G241**
- Für jede Sprache braucht man einen eigenen Code (Deutsch = „Kölner Phonetik“)

Suchtechnologien: Semantische Suche

- Klassifizierung von Dokumenten mit Hilfe texttechnologischer Verfahren
- Ermittlung des Kontextes, in dem ein Suchbegriff steht (z.B. Laster = LKW / Schlechte Angewohnheit)
- Ermittlung der Intention des Nutzer (wonach sucht der Nutzer)
- Problem: Bei kurzen Suchanfragen nicht geeignet

Ranking-Verfahren

- Formale Sortierung
- On-Page-Faktoren
- On-Site-Faktoren
- Link-Faktoren
- Eigenschaften und Verhalten der Nutzer

Ranking-Verfahren: Formale Sortierung

- Alphabetische Sortierung nach Titel oder Autor
- Chronologische Sortierung nach Erscheinungsdatum
- Nur in Bereichen sinnvoll und möglich, die über entsprechende Metadaten verfügen

Ranking-Verfahren: On-Page-Faktoren

- Häufigkeit und Position (Dichte, Abstand) der Terme
- Funktion (URL, HTML-Auszeichnung: Titel, Überschrift, Linktext)
- Format (Schriftgröße und –farbe)

→ Hohes Missbrauchspotential

Ranking-Verfahren: On-Site-Faktoren

- Analyse globaler Faktoren der jeweiligen Domain
- Alter der Domain
  Art der Domain (Kommerziell / Wissenschaftlich)
  Thematische Ausrichtung
  Gesamtzahl der indexierten Seiten
  Linkpopularität der gesamten Domain

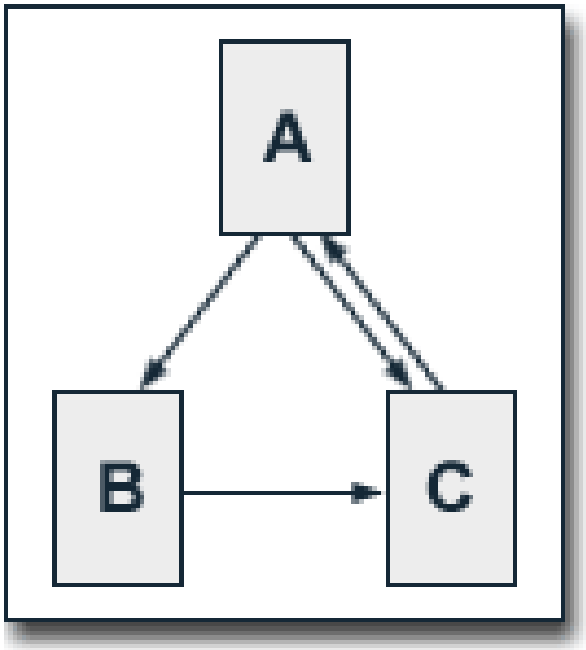
Ranking-Verfahren: Link-Faktoren

- Ranking basierend auf der Analyse der Referenzstruktur im Web
- Annahme: Link stellt ein Qualitätsurteil dar (ähnlich einem Zitat / Zitationsanalyse)
- PageRank: Ermittlung der Wichtigkeit einzelner Dokumente durch Analyse der Verweisstruktur aller indexierten Webseiten

Ranking-Verfahren:
PageRank (einfache Fassung)

- **PR(A)** = (1-**d**) + **d** (**PR(T1)**/**C(T1)** + ... + **PR(Tn)**/**C(Tn)**)
- **PR(A)** = PageRank der Seite A
- **d** = Dämpfungsfaktor, wobei 0 <= d <= 1 ist (i.d.R. 0,85)
- **PR(Tn)** = PageRank der Seite Tn, von denen ein Link auf die Seite A zeigt
- **C(Tn)** = Gesamtanzahl der Links von Seite Tn

Ranking-Verfahren:
PageRank (einfache Fassung, d = 0,5)



Quelle: http://pr.efactory.de/d-pagerank-algorithmus.shtml

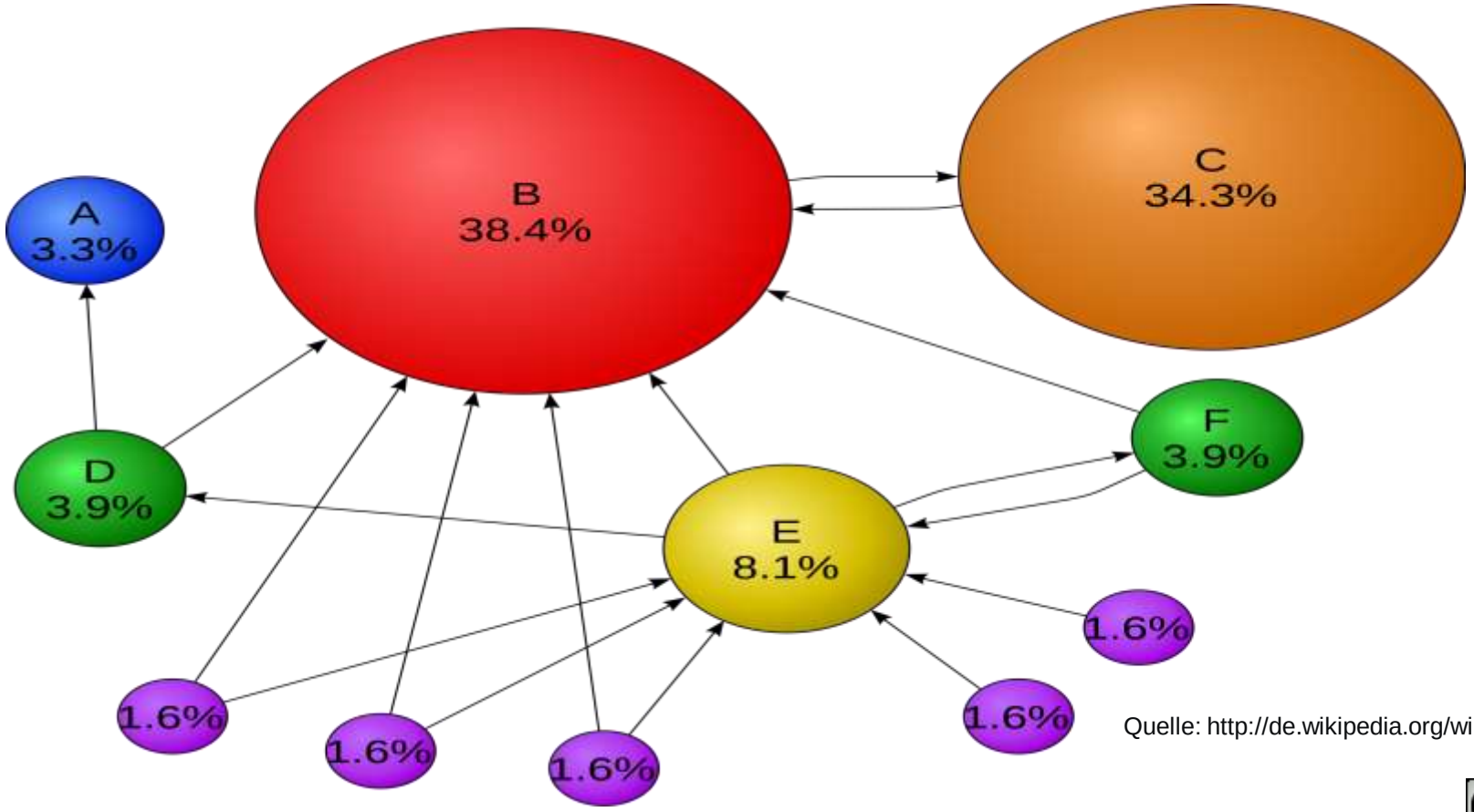
- PR(A) = 0.5 + 0.5 PR(C)
  PR(B) = 0.5 + 0.5 (PR(A) / 2)
  PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))
- Bei Ausgangswert 1 für alle Seiten ergibt sich nach 13 Iterationen
- PR(A) = 14/13 = **1.07**692308
  PR(B) = 10/13 = **0.76**923077
  PR(C) = 15/13 = **1.15**384615

Ranking-Verfahren:
PageRank-Wert und echter PageRank

Toolbar-PR	Tatsächlicher PR		
0/10	0.15	-	0.9
1/10	0.9	-	5.4
2/10	5.4	-	32.4
3/10	32.4	-	194.4
4/10	194.4	-	1,166.4
5/10	1,166.4	-	6,998.4
6/10	6,998.4	-	41,990.4
7/10	41,990.4	-	251,942.4
8/10	251,942.4	-	1,511,654.4
9/10	1,511,654.4	-	9,069,926.4
10/10	9,069,926.4	-	0.85 × N + 0.15

Quelle: http://pr.efactory.de/d-pagerank-algorithmus.shtml

Ranking-Verfahren:
Besuchs-Wahrscheinlichkeit



Quelle: http://de.wikipedia.org/wiki/PageRank

Ranking-Verfahren: Link-Faktoren

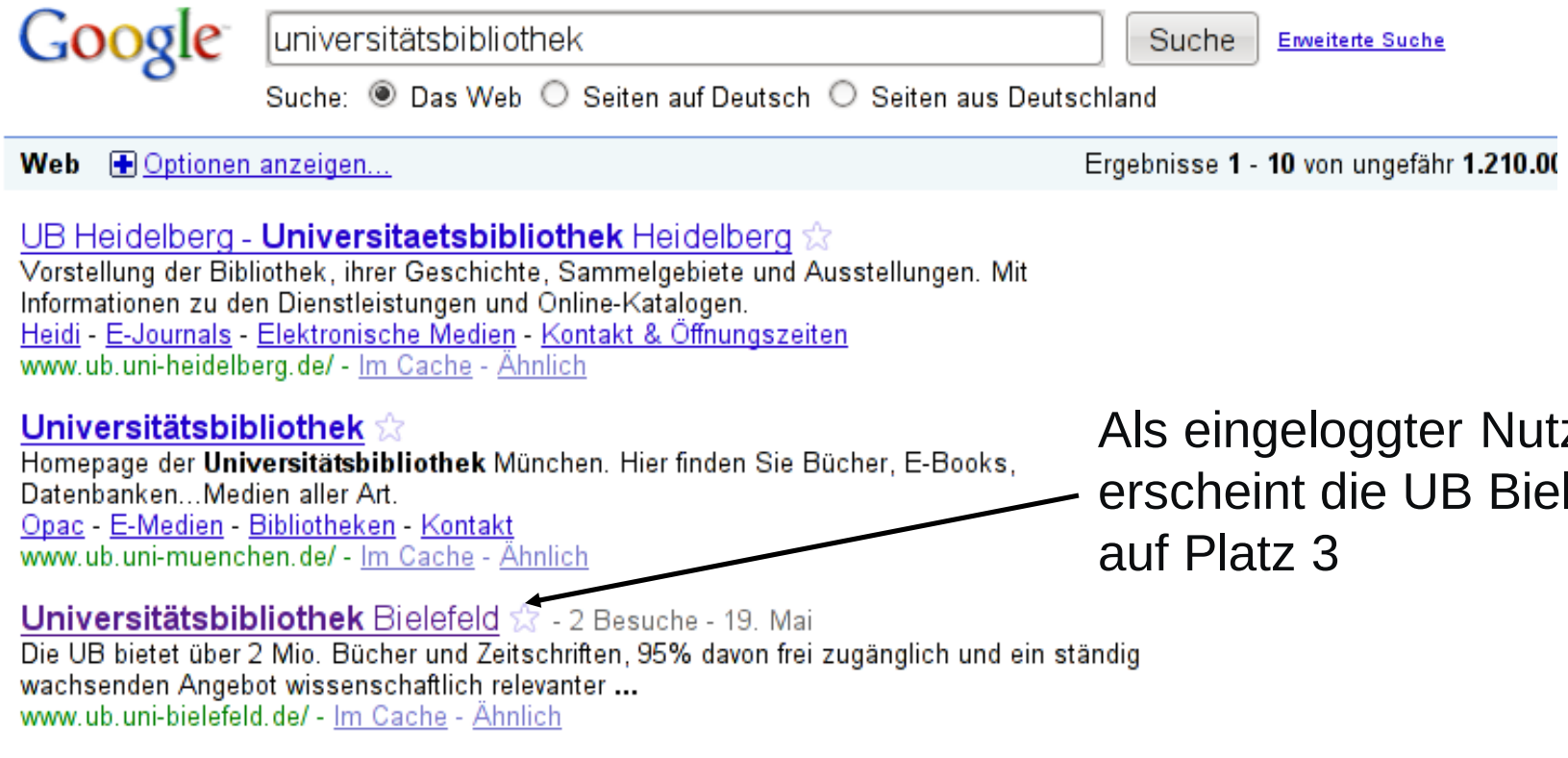
- Linktexte von eingehenden Links werden ebenfalls indexiert und in das Ranking einer Webseite einbezogen
- Eine Webseite ist auch dann auffindbar bzw. wird als Treffer angezeigt, wenn das gesuchte Wort lediglich im Linktext eines externen Verweises, nicht aber auf der Seite selbst vorkommt
- Missbrauchspotential: „Google Bombing“

Ranking-Verfahren: „Google Bombing“

Ranking-Verfahren:
Eigenschaften und Verhalten der Nutzer

- Geografische Zuordnung des Users anhand der IP-Adresse (Land, Region, Stadt)
- Darstellung und Reihenfolge in Abhängigkeit der geografischen Zuordnung
- Suchergebnis wird mit Karten angereichert (Standorte von Firmen, Kneipen, Bibliotheken)
- Bei persönlichem Login zusätzlich Analyse der Suchhistorie des Anwenders

Ranking-Verfahren:
Eigenschaften und Verhalten der Nutzer



Als eingeloggter Nutzer erscheint die UB Bielefeld auf Platz 3

Ranking-Verfahren:
Eigenschaften und Verhalten der Nutzer

Als anonymer Nutzer erscheint ein anderer Treffer auf Platz 3

Ranking-Verfahren:
Eigenschaften und Verhalten der Nutzer

Auch als anonymer Nutzer
Erkennt Google meinen Standort anhan der IP
Und gibt passende Ergebnisse aus

Suchmaschinen-Optimierung (SEO)

- Optimierung einer Webseite hinsichtlich der Indexierung und der Relevanzbewertung einer Suchmaschine
- Aussagekräftige Titel, Überschriften und Linktexte
- Optimierung der internen und externen Verlinkungsstruktur
- ASEO = Academic Search Engine Optimization (Optimierung von PDFs für Google Scholar)

Suchmaschinen-Optimierung (SEO)

URL: www.ub.uni-bielefeld.de

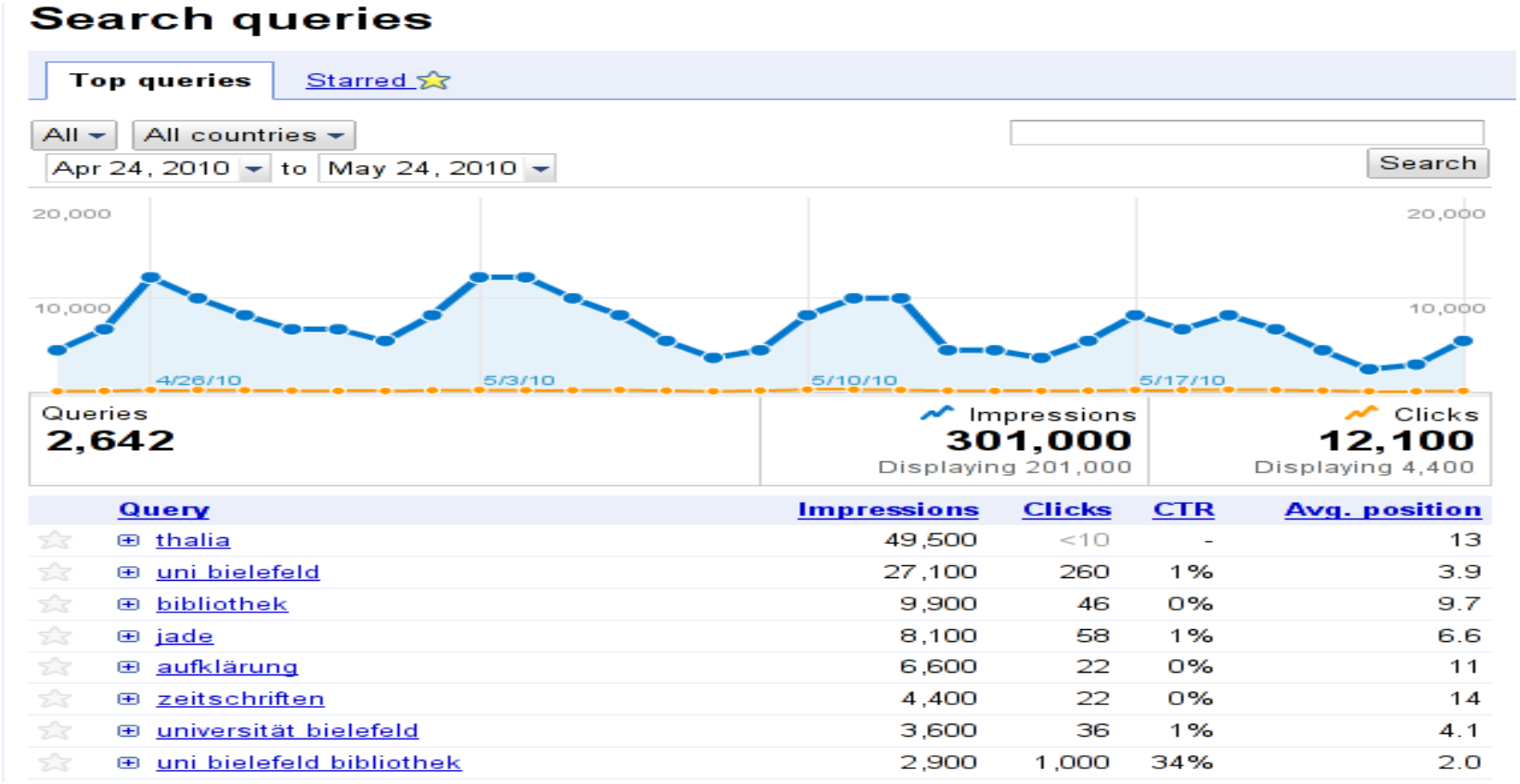
Seitwert	Ranking (Top 100)
54,88	393
von 100 Punkten	von 541135 Seiten

Seitwert-Berechnung

Gewichtung bei Google:	19,05 / 29	66%	Details
Gewichtung bei Yahoo:	8,14 / 17	48%	Details
Externe Wertungen:	6,34 / 10	63%	Details
Technische Details:	12 / 13	92%	Details
Social Bookmarks:	0,35 / 22	2%	Details
Sonstiges:	9 / 9	100%	Details

Quelle: http://www.seitwert.de

Suchmaschinen-Optimierung (SEO)



Search queries

Top queries | Starred

All | All countries | Search

Apr 24, 2010 to May 24, 2010

Queries	Impressions	Clicks
2,642	301,000	12,100
	Displaying 201,000	Displaying 4,400

Query	Impressions	Clicks	CTR	Avg. position
thalia	49,500	<10	-	13
uni bielefeld	27,100	260	1%	3.9
bibliothek	9,900	46	0%	9.7
jade	8,100	58	1%	6.6
aufklärung	6,600	22	0%	11
zeitschriften	4,400	22	0%	14
universität bielefeld	3,600	36	1%	4.1
uni bielefeld bibliothek	2,900	1,000	34%	2.0

Quelle: Google Webmaster Tools

Ausblick

- Relevanzbewertung verstärkt nach Eigenschaften und Verhalten des Nutzers („Local Search“)
- „Universal Search“ = Neben Ergebnissen aus Webseiten auch Anzeige von Bildern, News, Blogeinträgen etc.
- Weitere Fortschritte bei der semantischen Suche

Vielen Dank für Ihre

Aufmerksamkeit!

Präsentation beenden