Introduction to Information Retrieval

Ulf Leser

Verschiebung

- VL am 27.4. muss verschoben werden
- Alternativen
  - Donnerstag, 22.4., 16 Uhr
  - Freitag, 23.4., 15 Uhr
  - Montag, 3.5., 16 Uhr

Content of this Lecture

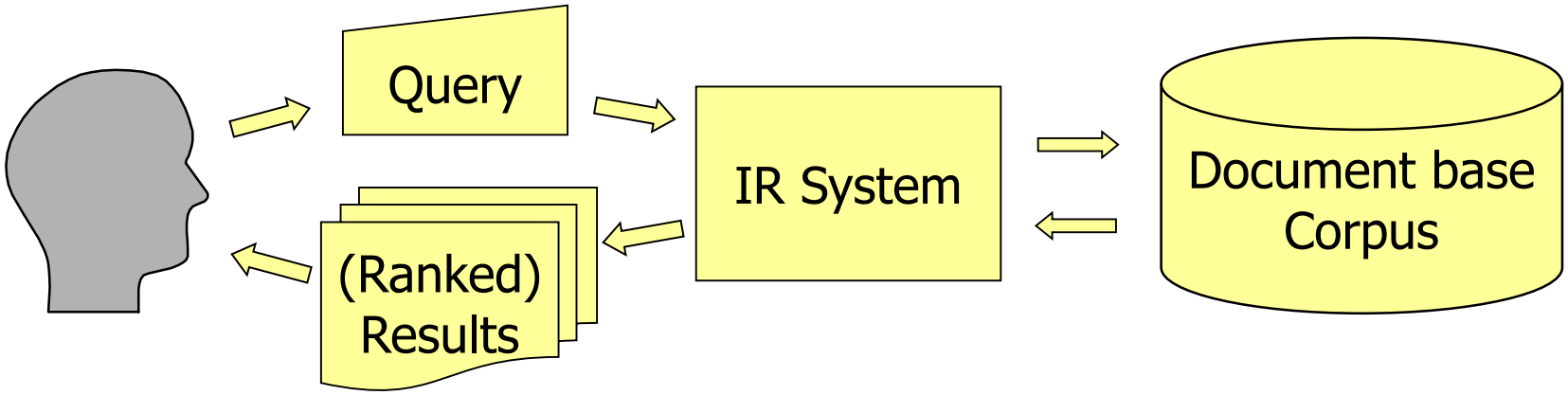
- What is Information Retrieval
- Documents
- Queries
- Related topics

Information Retrieval (aka "Search")

- Naïve: Find all documents containing the following words
- Advanced: „Leading the user to those documents that will best enable her to satisfy her need for information"
  - [Robertson 1981]
  - A user wants to know something
  - The user needs to tell the machine what he wants to know: query
  - Posing exact queries is difficult: room for interpretation
  - Machine interprets query to compute the (hopefully) best answer
  - Goodness of answer (relevance) depends on original intention of user, not on the query
  - Answer is always a set of docs (in classical IR)
  - "Leading": Sensible ranking of all potentially relevant docs

The Informal Problem

- Help user in quickly finding the requested information within a given set of documents
  - Set of documents: Corpus, library, collection, …
  - Quickly: Few queries, fast responses, simple interfaces, …
  - Requested: The "best-fitting" documents; the "most relevant" content



Difference to Database Queries

- Queries: Formal language versus natural language
- Result granularity: Set of documents versus relation as defined by query
- Exactly defined result versus loosely described relevance
- Result set versus ranked result list
- DB: Posing the right query is completely left to the user
- IR: Understanding the query is a problem of the software

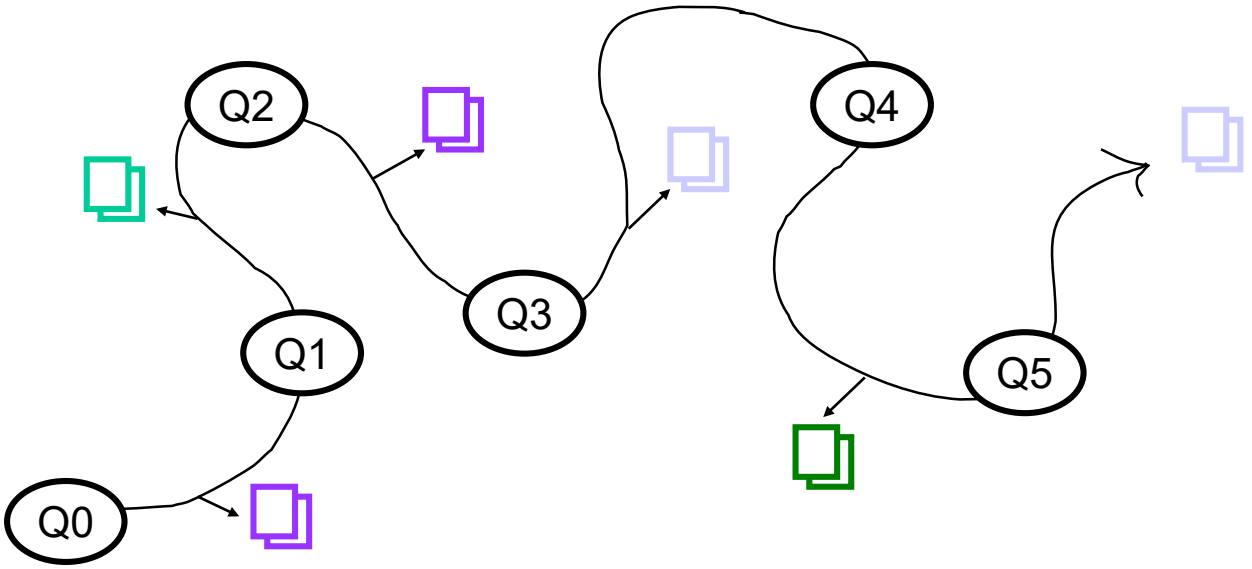
Why is it hard?

- Properties of human languages
  - Homonyms (context): Fenster (Glas, Computer, Brief, …), Berlin (BRD, USA, …), Boden (Dach, Fussboden, Ende von etwas, …
  - Synonyms: Computer, PC, Rechner, Desktop, Laptop, Tablet, …
- Properties of the corpus
  - Size: Corpora today may have billions of documents
  - Heterogeneity: Length, format, language, genre, grammatical correctness, special characters, …
- Heterogeneous users: Precise queries versus usability
  - Lay persons: Short queries with wide spectrum of interpretations
    - Average web queries have 1,6 terms
    - Additional knowledge: Location, current trends, popular answers, …
  - Professionals: Long queries trying to precisely define the intention
    - "Information broker" was/is a profession

Quickly

- Time to execute a query
  - Indexing, parallelization, compression, …
- Time to answer the request (may involve multiple queries)
  - Understand request, find best matches
  - Success of search engines: Better results (and fast!)
  - Process-orientation: User feedback, query history, …
- Information overload
  - "We are drowning in data, but starving for knowledge"
  - If the corpus is large, ranking is a must
  - Alternative: Result summarization (grouping on what?)
  - Different search modes: What's new? What's certain?

IR: An Iterative, Multi-Stage Process



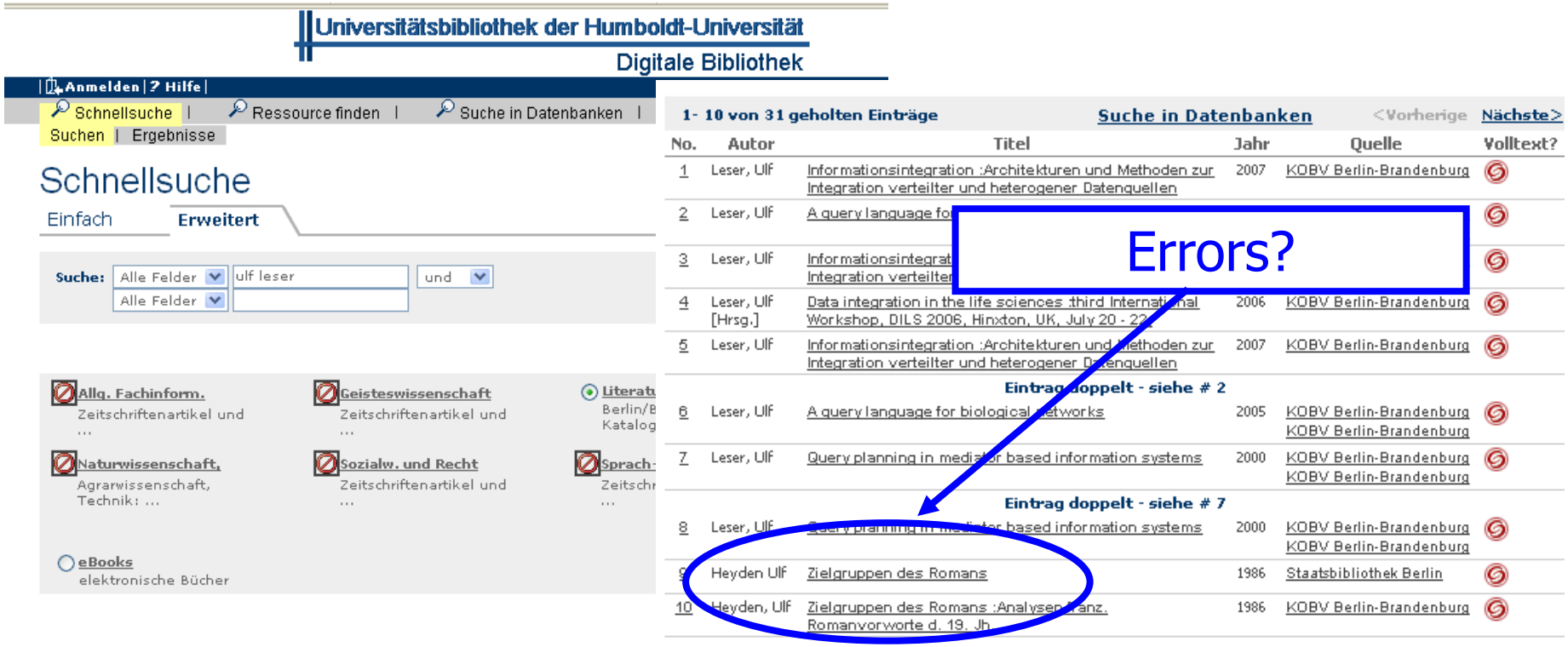
- IR process: "Moving through many actions towards a general goal of satisfactory completion of research related to an information need."
  - "Berry-picking" [Bates 89]

Gesellschaft für Informatik (2014)

- Im Information Retrieval (IR) werden Informationssysteme in Bezug auf ihre Rolle im Prozess des Wissenstransfers vom menschlichen Wissensproduzenten zum Informations-Nachfragenden betrachtet. … Fragestellungen, die im Zusammenhang mit vagen Anfragen und unsicherem Wissen entstehen .... auch solche, die nur im Dialog iterativ durch Reformulierung (in Abhängigkeit von den bisherigen Systemantworten) beantwortet werden können ... Die Unsicherheit resultiert meist aus der begrenzten Repräsentation von dessen Semantik (z.B. bei Texten oder multimedialen Dokumenten);... Aus dieser Problematik ergibt sich die Notwendigkeit zur Bewertung der Qualität der Antworten eines Informationssystems, wobei in einem weiteren Sinne die Effektivität des Systems in Bezug auf die Unterstützung des Benutzers bei der Lösung seines Anwendungsproblems beurteilt werden sollte.

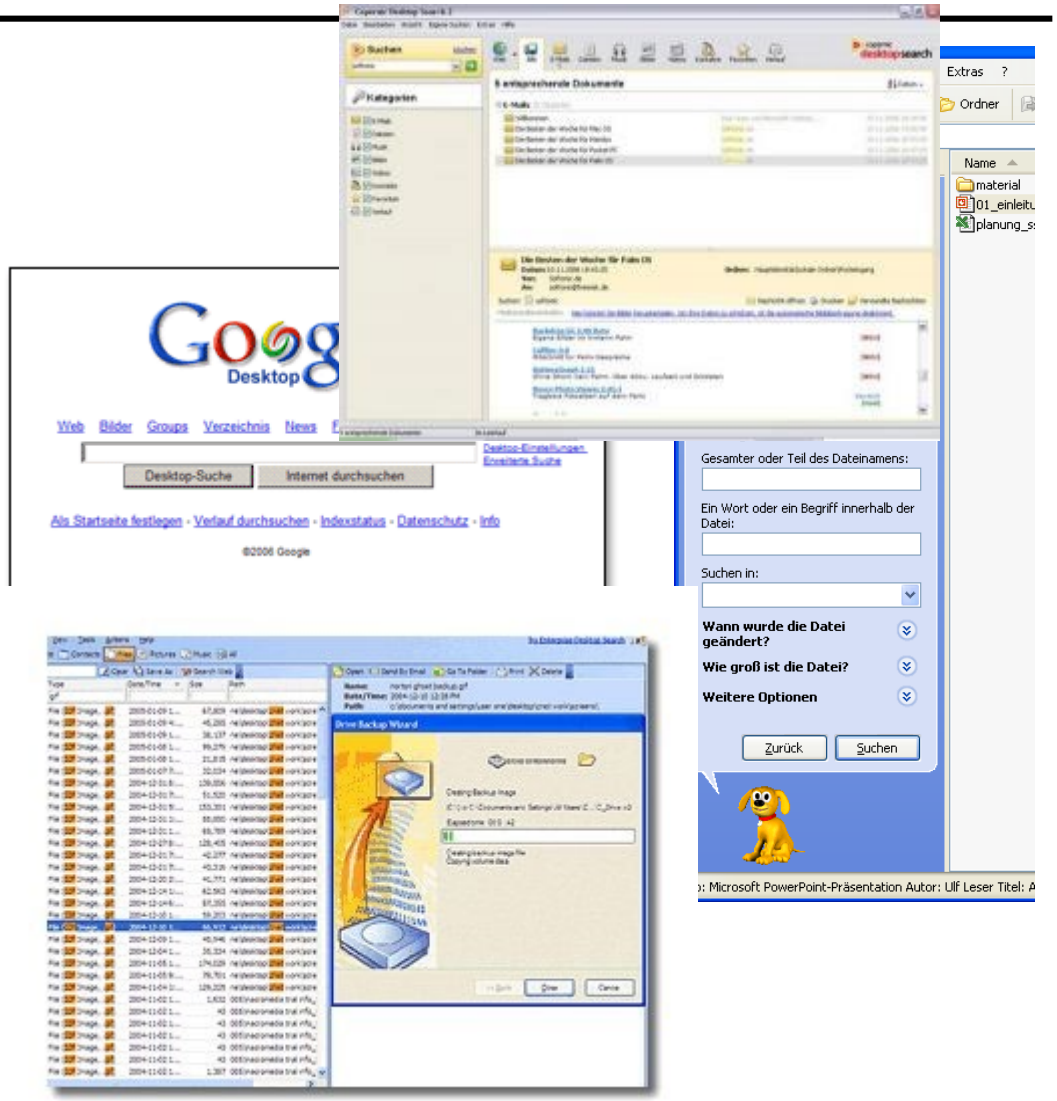
Prominent Systems I: Digital Libraries

- E.g. OPAC
  - Combination of structured attributes and IR-style queries

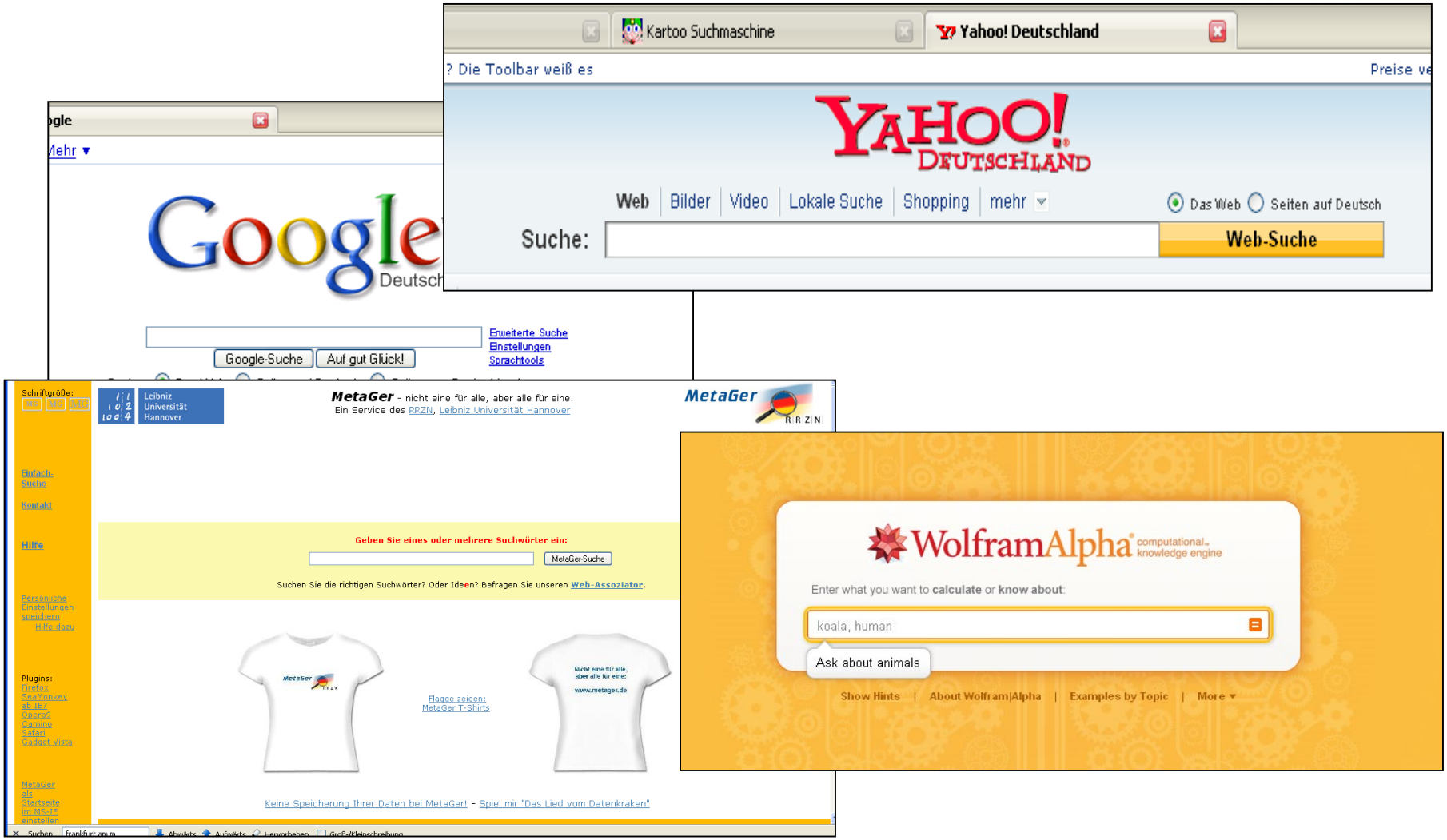


Prominent Systems II: Desktop Search

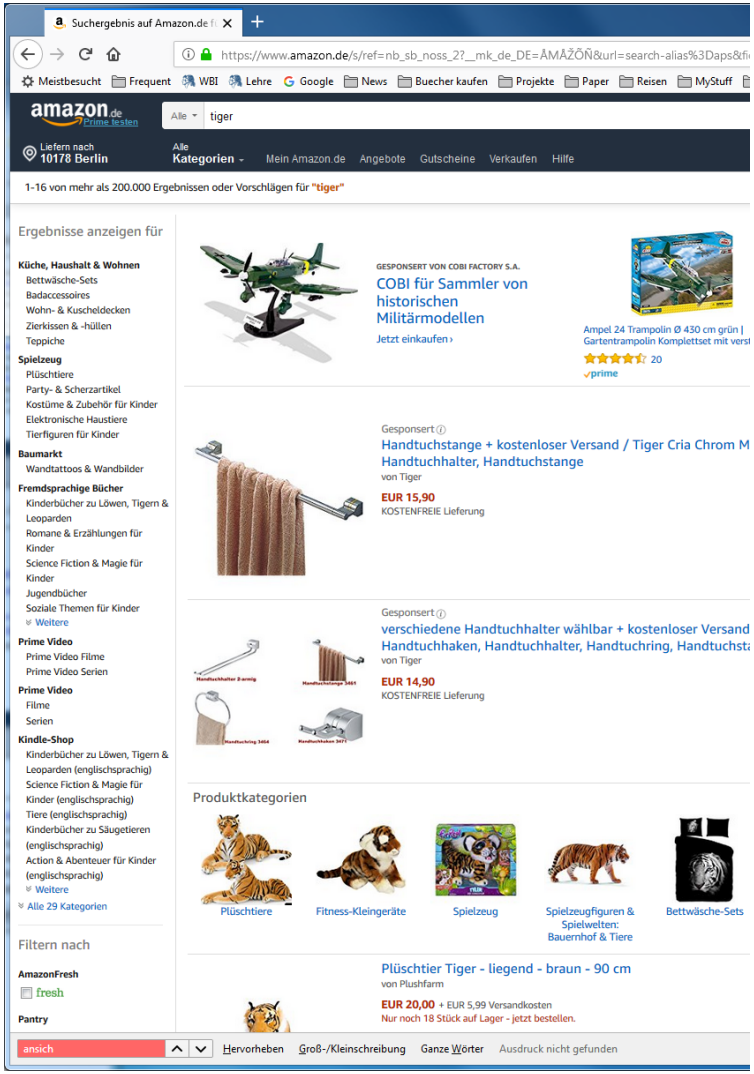
- Much activity in 2000-2010
- Various search engines and indexing mechanisms
- Important: Search different types of files (txt, doc, mail, ppt, pdf, tex, odp, xls, ...)

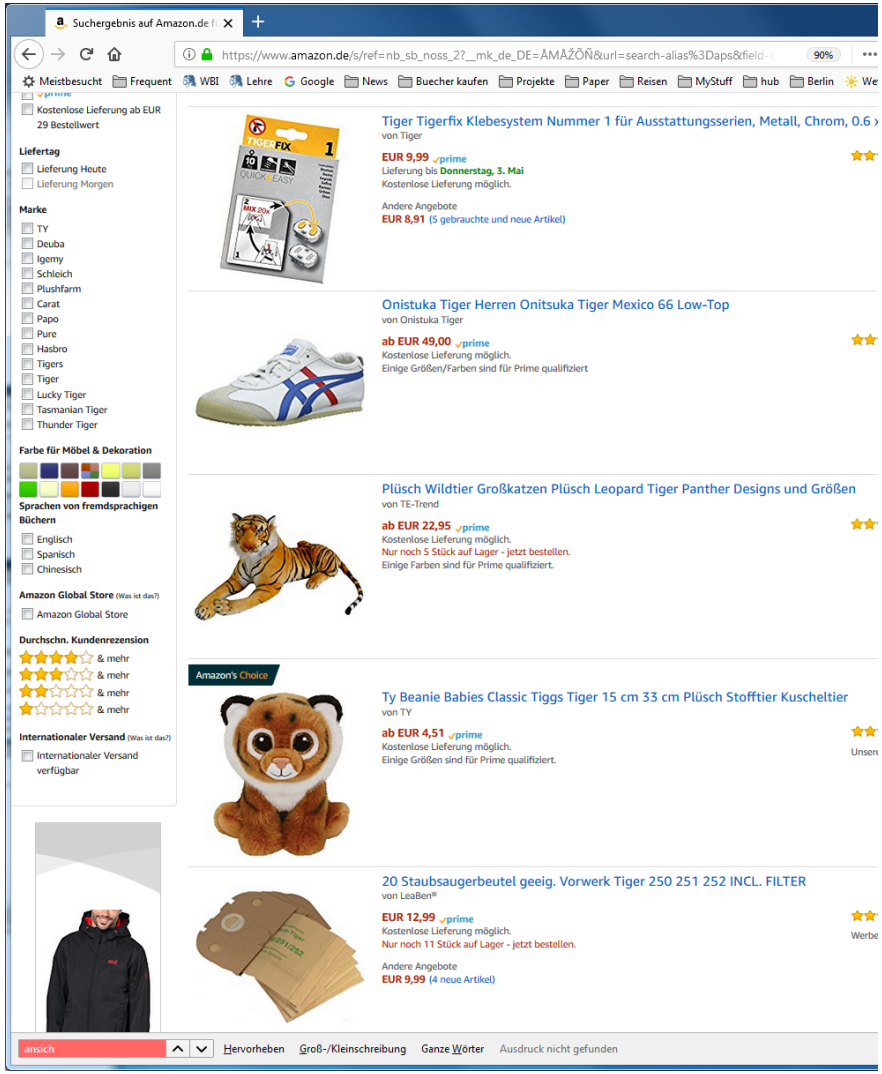


Prominent Systems III: Web Search Engines



Almost Any Web Site





Properties of Information Retrieval (IR)

- IR is about helping a user
- IR is about finding information, not about finding data
- IR builds systems for end users, not for programmers
  - No SQL
  - IR (web) today is used by almost everybody, databases are not
- IR searches unstructured data (e.g. text)
- 90% of all information is presented in unstructured form
  - Claim some analysts

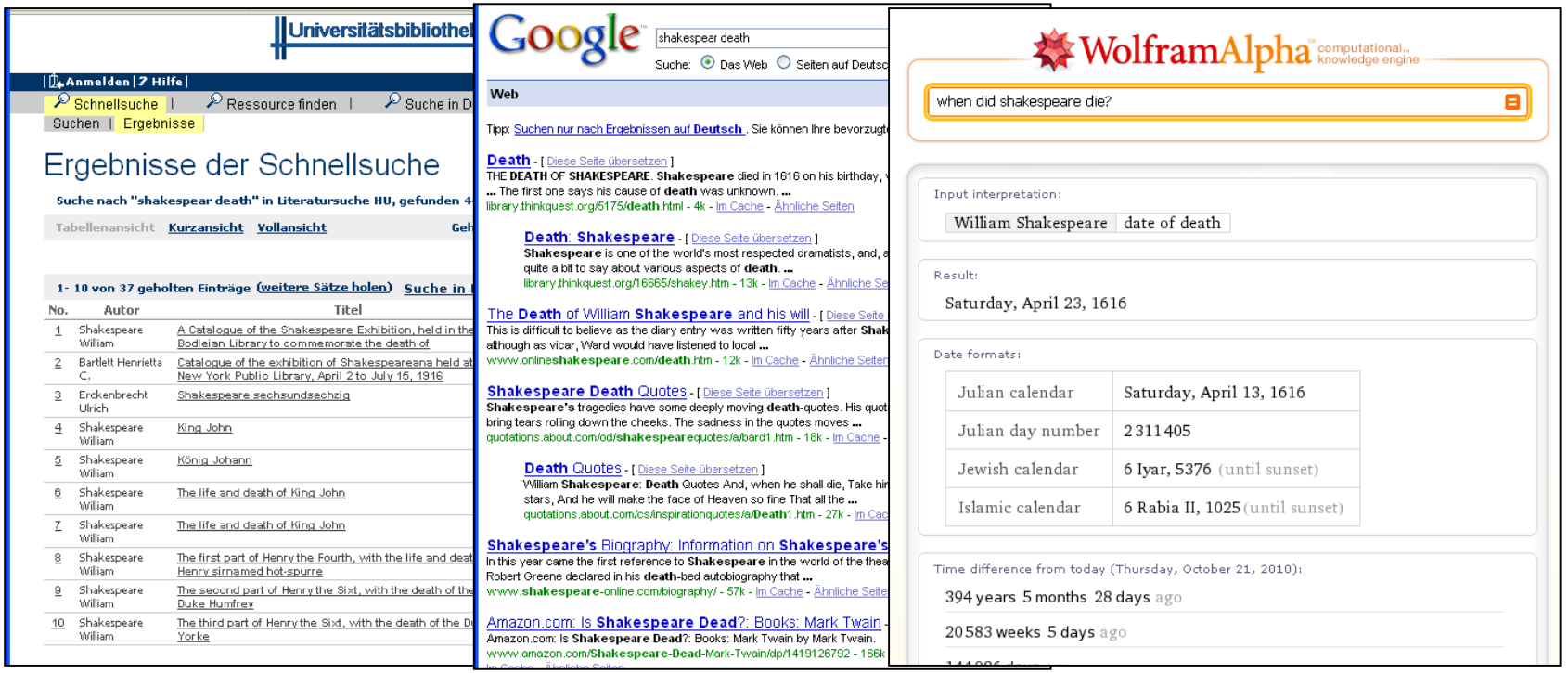
History

- ~300 ad. Library of Alexandria , ~700.000 „documents“
- 1450: Bookprint
- 19th century: Indices / concordance
- Probabilistic models: Maron & Kuhns (1960)
- Boolean queries: Lockheed (~1960)
- Vector Space Model: Salton, Cornell (1965)
  - Faster, simpler to implement, better search results
- 80s-90s: Digital libraries, SGML, hypertext, metadata standards
- Mid 90s: The web, web search engines, XML, federations
- End 90s: Personalized search engines, recommendations
- 2010 - : Mobile/localized search, user-generated content, social networks
- 2015 - : Knowledge graphs, entity search, personalization
- 2018 - : Question answering, language models, machine-learning based

Content of this Lecture

- What is Information Retrieval
- Documents
- Queries
- Related topics

Document or Passage



Searching only metadata

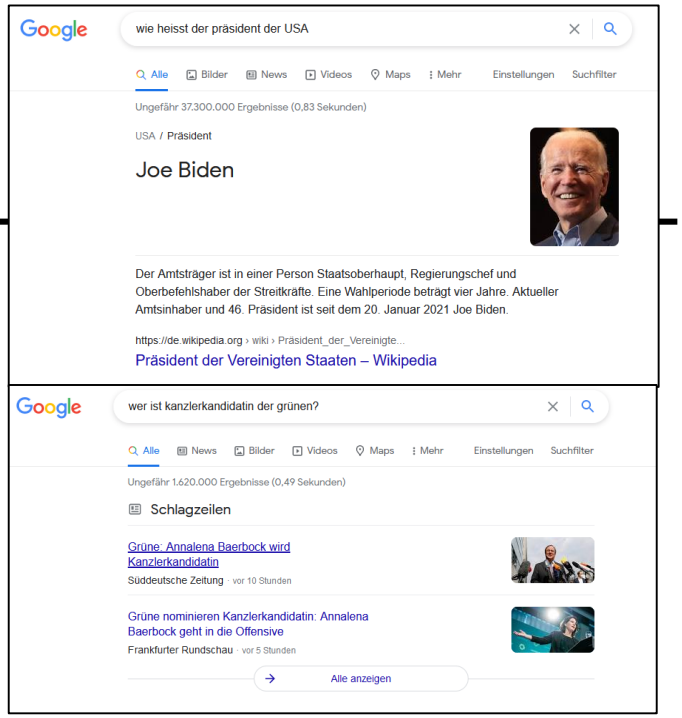
Searching tokens within documents

Interpreting natural text

Documents

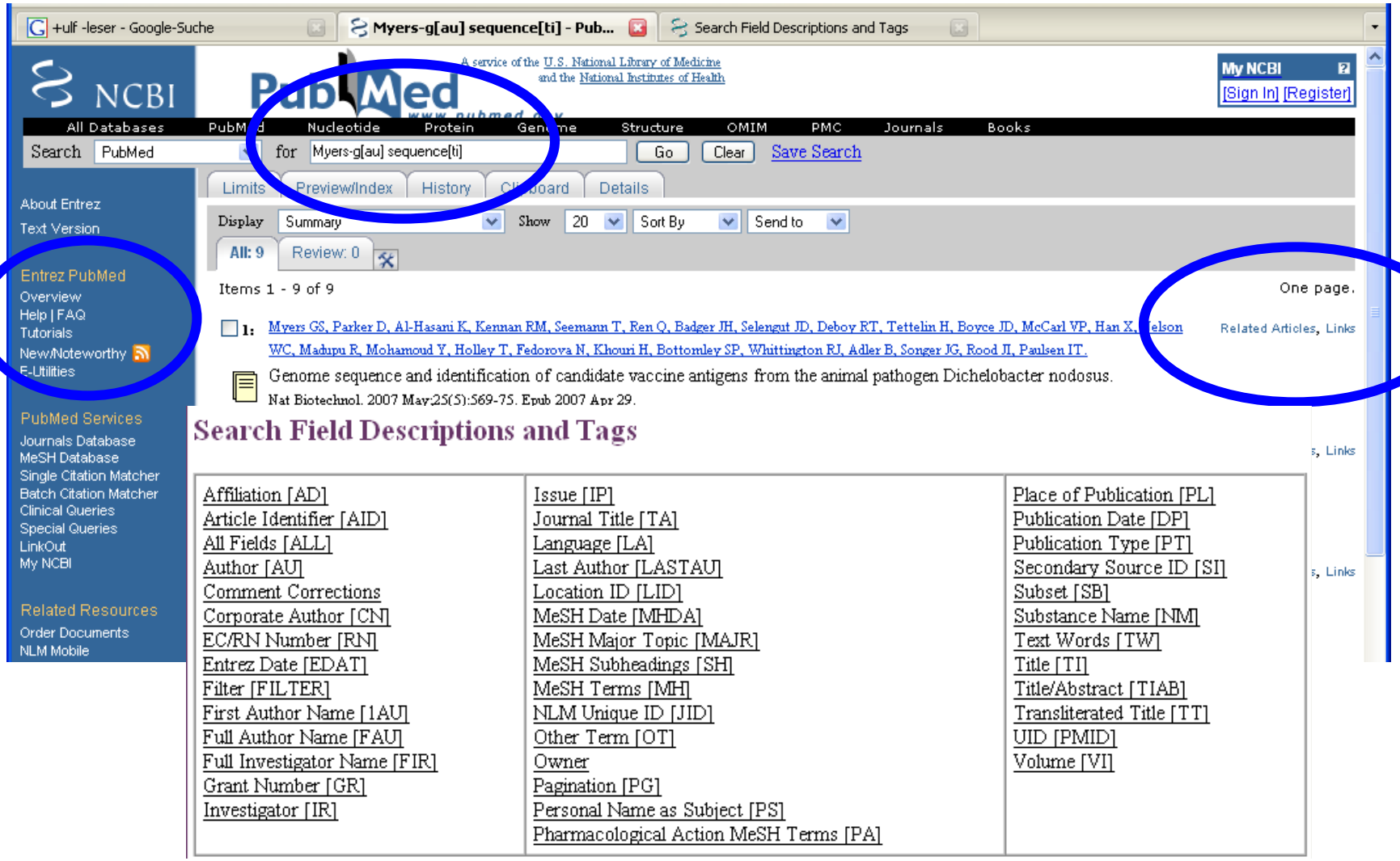
- This lecture: Natural language text
- Might be grammatically correct (books, newspapers) or not (blogs, Twitter, spoken language)
- May have structure (title, abstract, chapters, …) or not
- May have associated (explicit or in-text) metadata or not
  - Author, title, year, publisher, …
- May be in different languages or even have mixed content
  - Foreign characters
- May have various formats (ASCII, PDF, DOC, XML, …)
- May refer to other documents (hyperlinks)
- Not covered here
  - Semi-structured data (XML)
  - Structured data (But: Keyword search in relational databases)

IR Queries

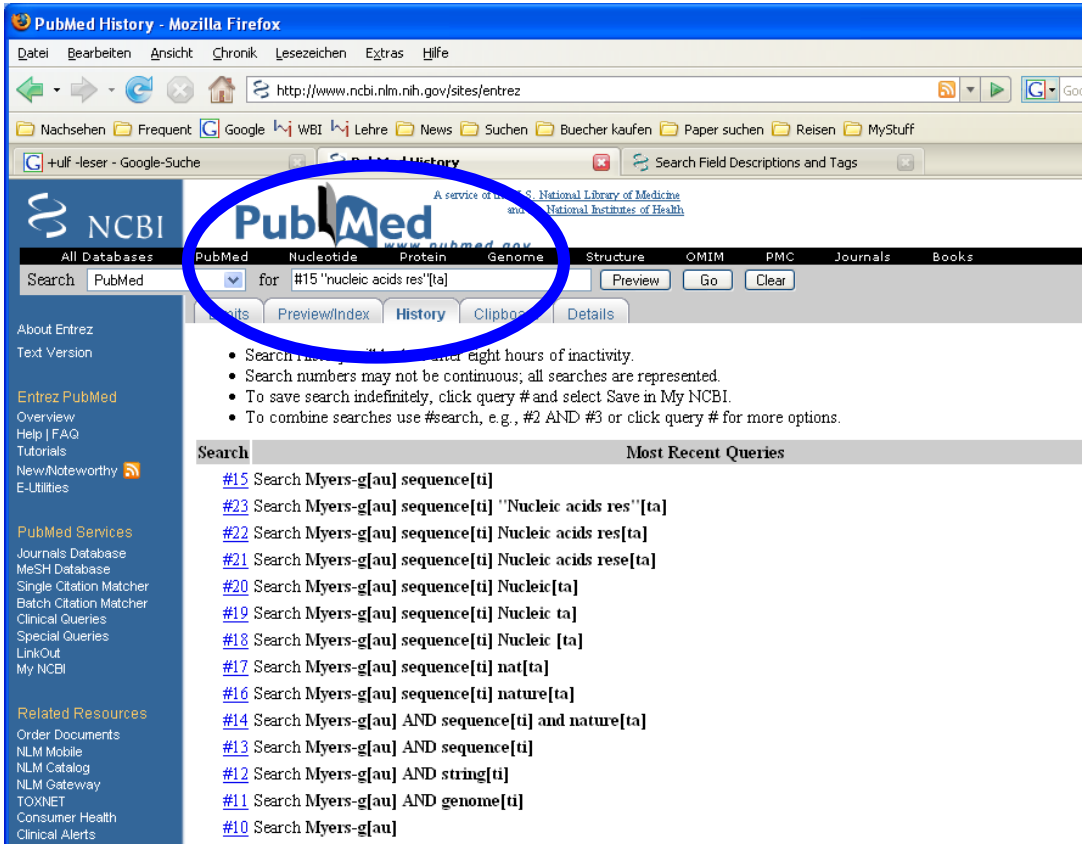


- Users formulate queries
  - Keywords or phrases
  - Logical operations (AND, OR, NOT, ...)
    - Also other operators: "-ulf +leser"
  - Natural language questions
    - Question answering, e.g. wolfram alpha
  - (Semi-)Structured queries (author=… AND title~ …)
  - Voice (Siri, Alexa, …)
- Documents as queries: Find documents similar to this one
- Query refinement based on previous results
  - Find documents matching the new query within the result set of the previous search
  - Use relevant answers from previous queries to create next query

Searching with Metadata (PubMed/Medline)



Query Refinement



Dublin Core Metadata Initiative (W3C), 1995

- identifier: ISBN/ISSN, URL/PURL, DOI, …
- format: MIME-Typ, media type,
- type: Collection, image, text, …
- language
- title
- subject: Keywords
- coverage: Scope of doc in space and/or time
- description: Free text
- creator: Last person manipulating the doc
- publisher:
- contributor:
- rights: Copyright, licenses, …
- source: Other doc
- relation: To other docs
- date: Date or period

Usage in HTML

```
<head profile="http://dublincore.org/documents/dcq-html/">
<title>Dublin Core</title>
<link rel="schema.DC" href="http://purl.org/dc/..." />
<link rel="schema.DCTERMS" href="http://purl.org/..." />
<meta name="DC.format" scheme="…" content="text/htm" />
<meta name="DC.type" scheme="…" content="Text" />
<meta name="DC.publisher" content="Jimmy Whales" />
<meta name="DC.subject" content="Dublin Core Metadata" />
<meta name="DC.creator" content="Björn G. Kulms" />
<meta name="DCTERMS.license" scheme="DCTERMS.URI"
       content="http://www.gnu.org/copyleft/fdl.html" />
</head>
```

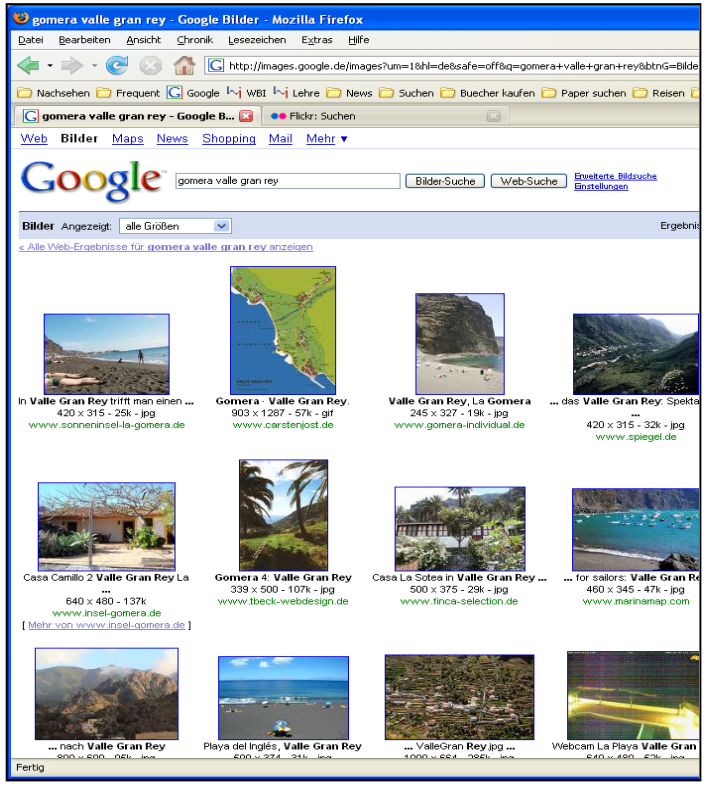
Knowledge Quiz

- Search engines are used only in the web
- Most search engines return documents, not direct answers to questions
- Search primarily is about faster answers to a query
- “Keyword queries" and “Boolean queries" are synonyms
- Search engines rarely offer sub-token search apart from morphology

Content of this Lecture

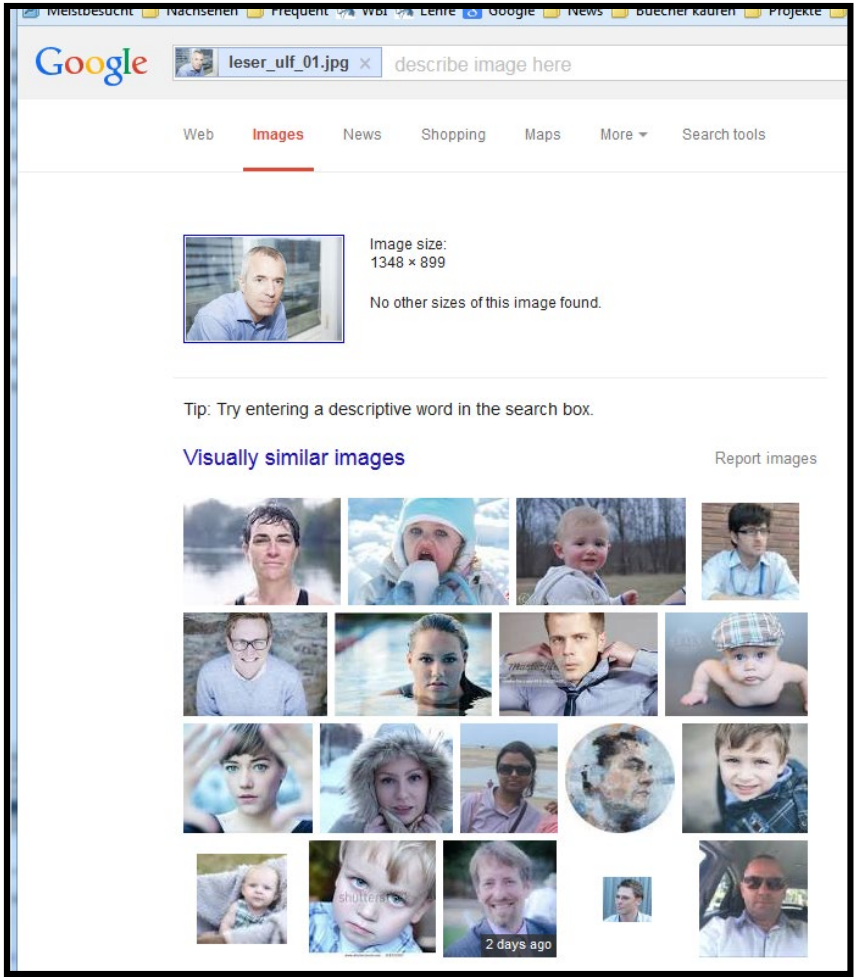
- What is Information Retrieval
- Documents
- Queries
- Related topics

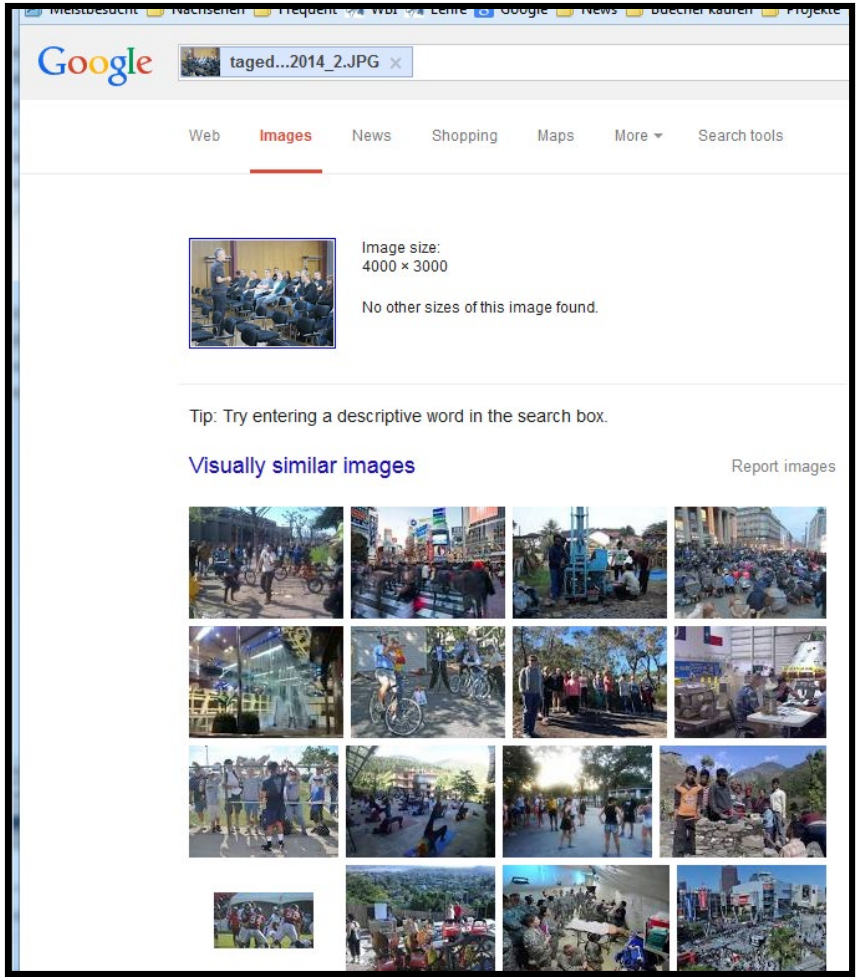
Multimedia Retrieval



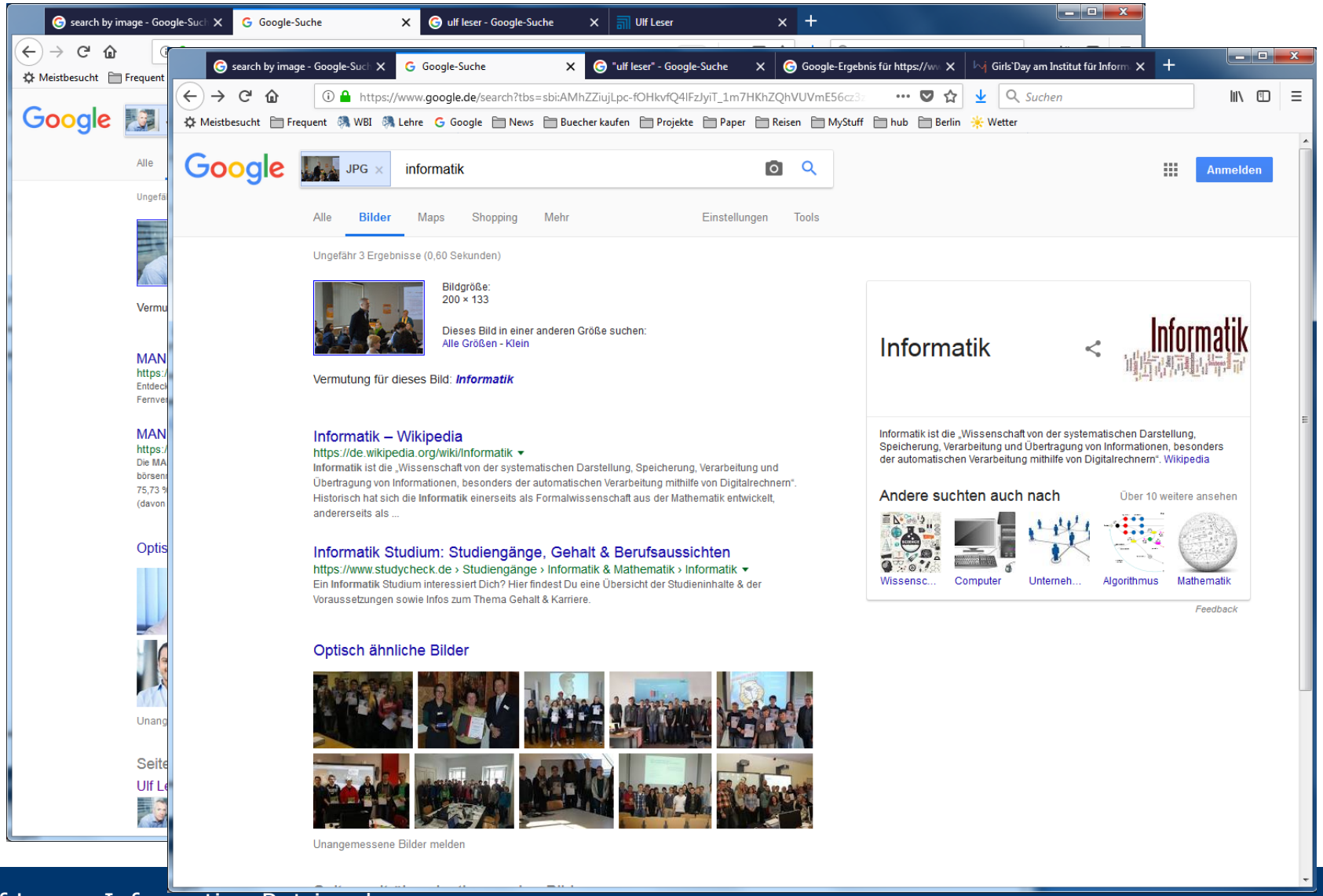
- Note: Neither searches within images
  - Flickr: tags ("folksonomy")
  - Google: text in neighborhood

„Search by Image“ (10/2014)





Search by Image 4/2018 – it's difficult ...



Question Answering

- Asking for a specific bit of information
  - What was the score of Bayern München versus Stuttgart in the DFB Pokal finals in 1998?
  - How many hours of sunshine has a day in Crete in May?
  - When does the next S9 leave this station?
- Prominent until recently: IBM Watson
  - "IBM Watson is a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data" [2011]

- Hot topic for personal assistants
  - E.g. Amazon Echo, Apple Siri, Google Assistant, …
- QA: Mixture of statistical NLP, Machine learning and IR

Historic Texts

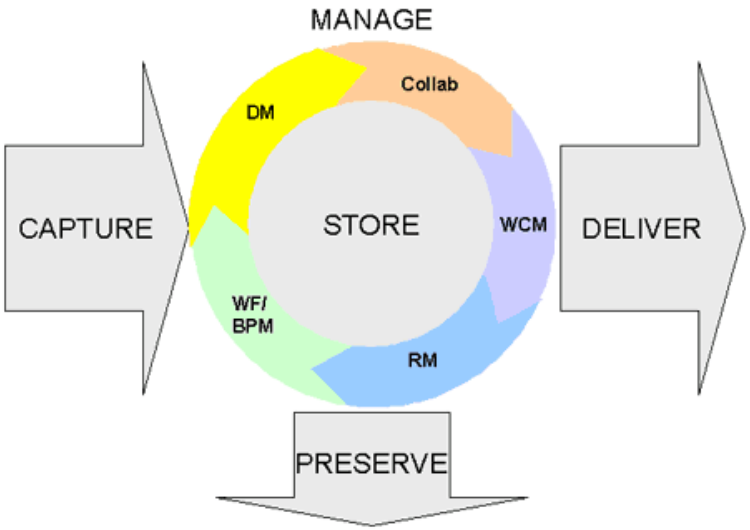
- Sachsenspiegel, ~1250
  - *"Swerlenrecht kůnnen wil•d~ volge dis buches lere.alrest sul wi mer ken, daz ..."*
- Multiple representations
  - Facsimile
  - Digitalization / diplomacy
    - How well can the facsimile be reproduced from the dig. form?
  - Differences in individual writers (proliferating errors)
  - Different translations
  - Different editions

Other Buzzwords

- Document management systems (DMS)
  - Large commercial market, links to OCR, workflow systems, etc.
  - Many legal issues (compliance, reporting, archival, …)
  - Essentially all companies run some form of a DMS
  - Every DMS includes an IR system
- Knowledge management
  - "More sophisticated" DMS with semantic searching
    - Ontologies, thesauri, topic maps, …
  - Social aspects: Incentives, communities, enterprise standards, …
- Digital libraries
  - Somewhat broader and less technical
  - Includes social aspects, archiving, multimedia, …

Enterprise Content Management

- „The technologies used to capture, manage, store, deliver, and preserve information to support business processes"



Quelle: AIIM International

- Authorization and authentication
- Business process management and document flow
- Compliance: legal requirements
  - Record management
  - Pharma, Finance, …
- Collaboration and sharing
  - Inter and intra organizations
  - Transactions, locks, …
- Publishing: What, when, where
  - Web, catalogues, mail push, …
- …

Technique versus Content

- IR is about techniques for searching a given doc collection
- Creating doc collections is a business: Content provider
  - Selection/filtering: classified business news, new patents, …
  - Augmentation: Annotation with metadata, summarization, linking of additional data, …
- Examples
  - Medline: >5000 Journals, >28M citations, >700K added per year
  - Thompson Reuter
    - Impact factors: which journals count how much?
  - Web catalogues ala Yahoo
  - "Pressespiegel", web monitoring

Self Assessment

- Give a definition of „Information retrieval“
- How is information retrieval different from database query evaluation?
- What are means to shorten the number of queries necessary to fulfil an information request?
- What is the difference between classical IR and Question Answering?
- What are possible types of answers to a IR query?