

# Introduction to Information Retrieval

Ulf Leser

# 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 him/her to satisfy his/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
  - “Leading”: Sensible **ranking** of all potentially relevant docs

# Difference to Database Queries

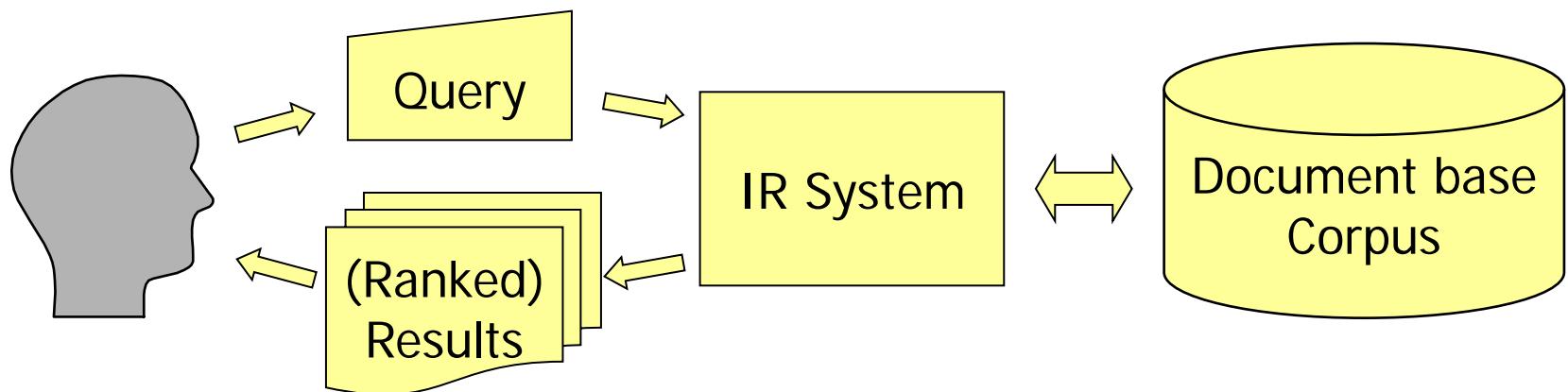
---

- 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**

# 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



# 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
  - Runtime: Corpora today may have **billions of documents**
  - Document heterogeneity: **Length**, format, language, genre, **grammatical correctness**, special characters, ...
- Heterogeneous users: **More precise** queries versus **usability**
  - Lay persons: Short queries with **wide spectrum** of interpretations
    - Average web queries have 1,6 terms
  - Professionals: Long queries often lead to **zero results**
  - “Information broker” was/is a profession

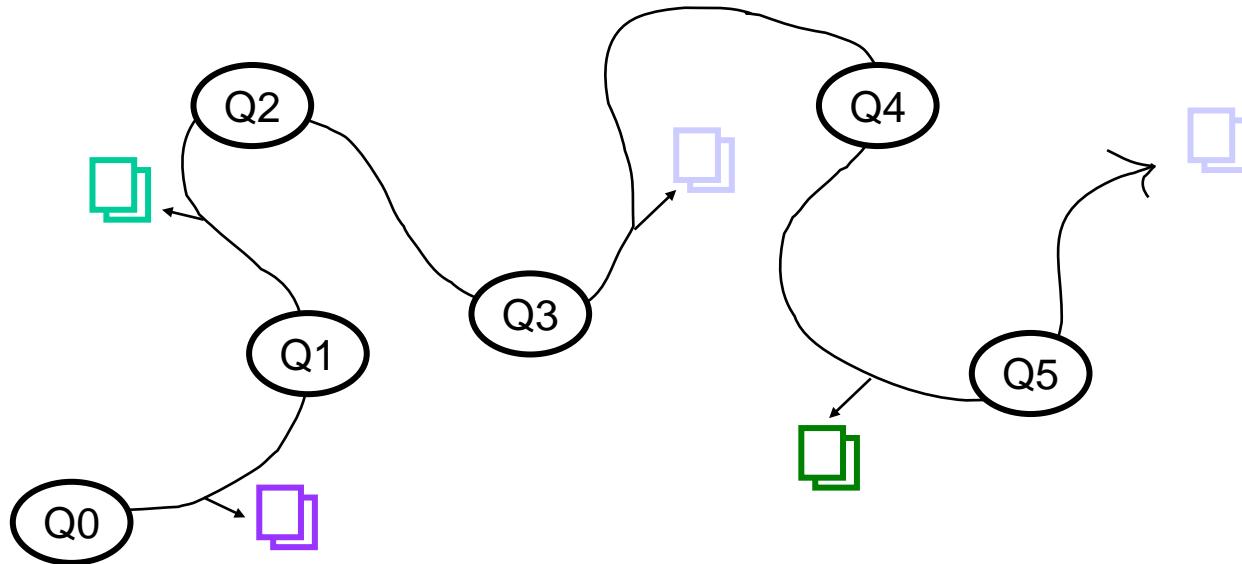
# Quickly

---

- 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

---

- 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

Screenshot of the Universitätsbibliothek der Humboldt-Universität Digitale Bibliothek OPAC interface.

The search results page shows 1-10 von 31 geholten Einträge. A blue box highlights the first result, and a large blue arrow points from the circled second result to the same highlighted entry.

Search results:

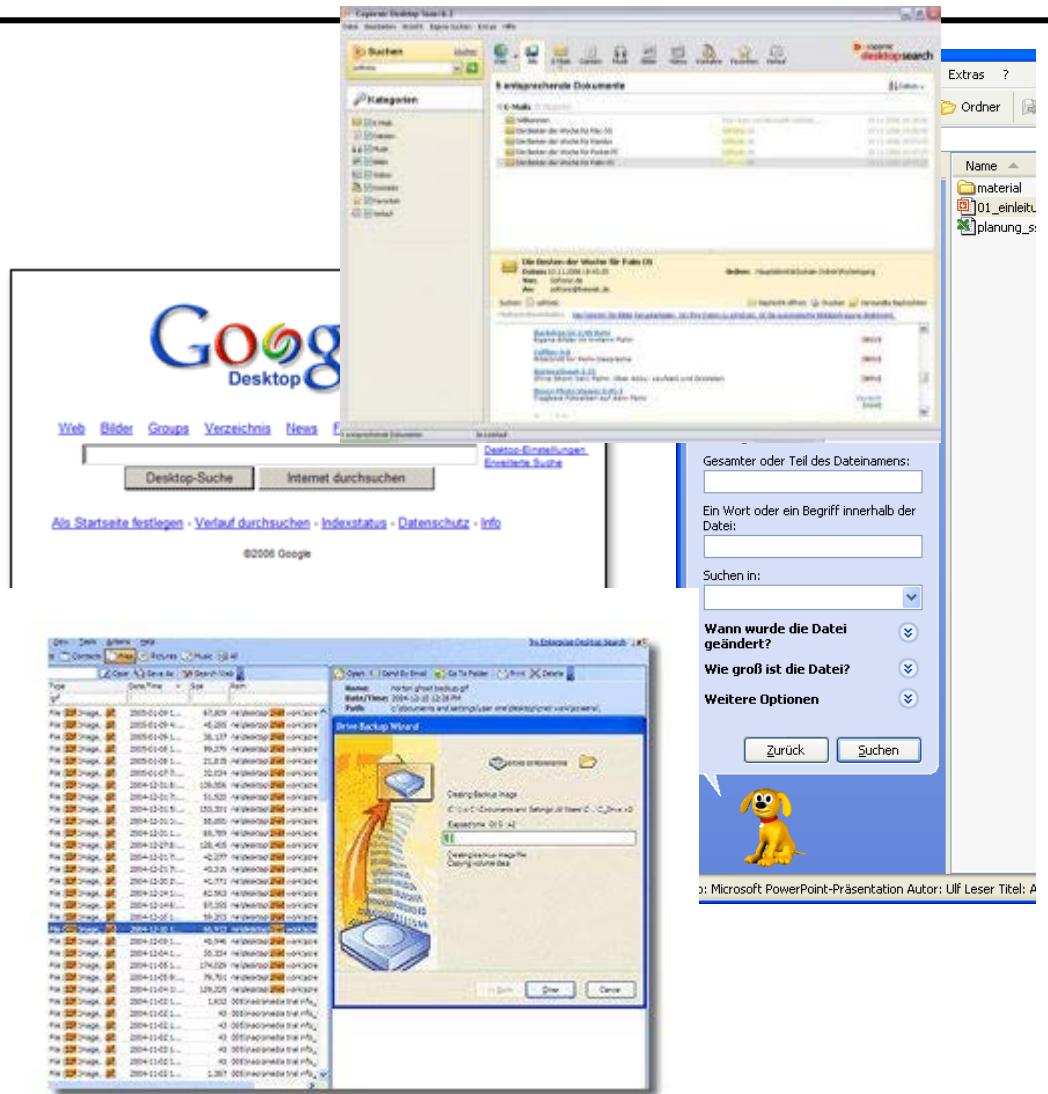
No.	Autor	Titel	Jahr	Verlag
1	Leser, Ulf	Informationsintegration Integration verteilter und heterogener Datenquellen		
2	Leser, Ulf	A query language for biological networks		
3	Leser, Ulf	Informationsintegration Integration verteilter und heterogener Datenquellen		
4	Leser, Ulf [Hrsg.]	Data integration in the life sciences : third International Workshop, DILS 2006, Hinxton, UK, July 20 - 27, 2006, Proceedings	2006	KOBV Berlin-Brandenburg
5	Leser, Ulf	Informationsintegration : Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen	2007	KOBV Berlin-Brandenburg
6	Leser, Ulf	A query language for biological networks	2005	KOBV Berlin-Brandenburg
7	Leser, Ulf	Query planning in mediator based information systems	2000	KOBV Berlin-Brandenburg
8	Leser, Ulf	Query planning in mediator based information systems	2000	KOBV Berlin-Brandenburg
9	Heyden, Ulf	Zielgruppen des Romans	1986	Staatsbibliothek Berlin
10	Heyden, Ulf	Zielgruppen des Romans : Analysen, Fazit, Romanvorworte d. 19. Jh.	1986	KOBV Berlin-Brandenburg

Annotations:

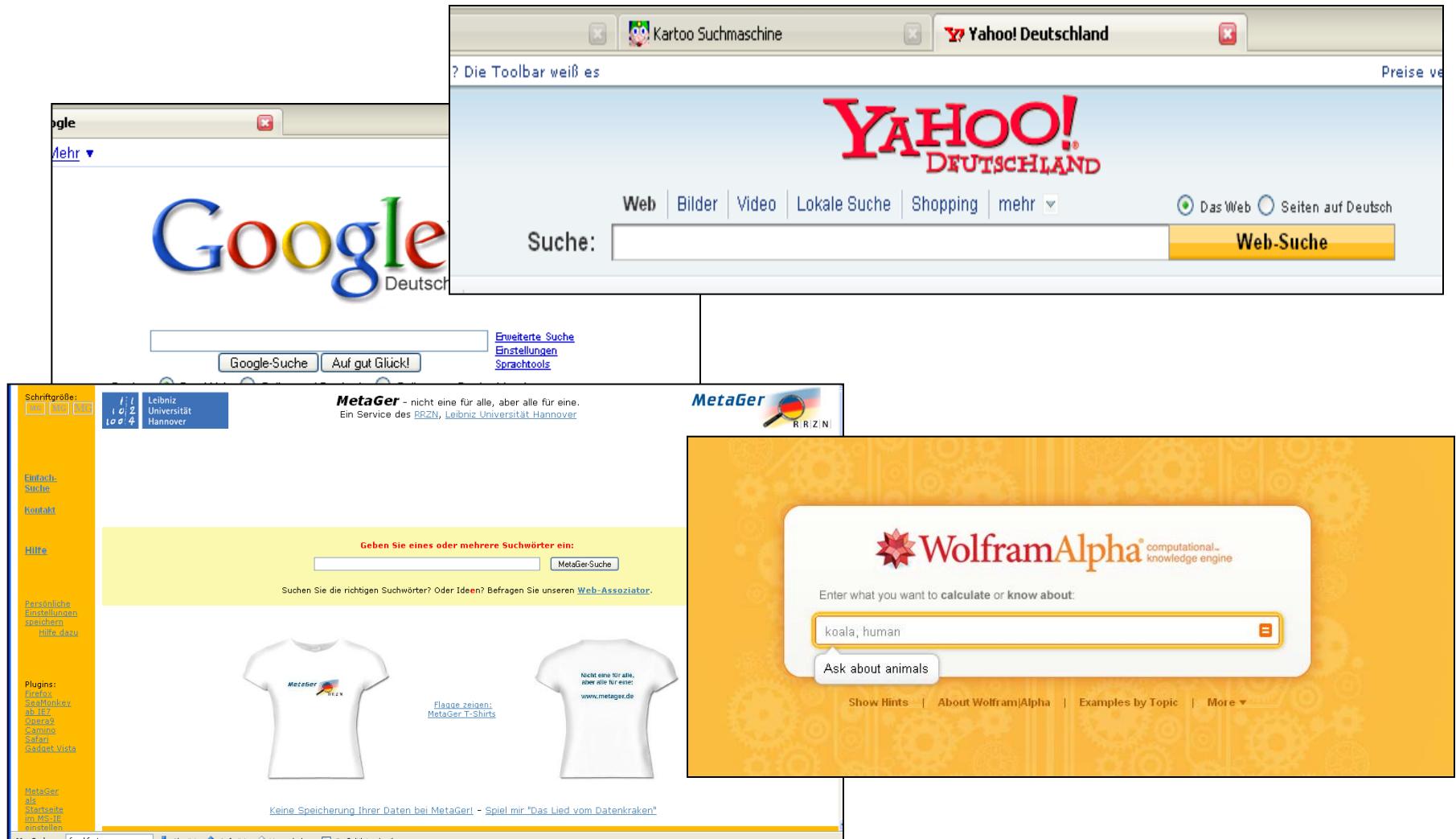
- A blue box surrounds the first result (No. 1).
- A large blue box surrounds the entire search results area, containing the text "Obviously not perfect".
- A blue arrow points from the circled second result (No. 2) to the highlighted first result (No. 1).

# 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



# 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) 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, metadata standards
- Mid 90s: The web, web search engines, XML
- End 90s: Personalized search engines, recommendations
- 2010: Mobile and context-based search, social networks

# Content of this Lecture

---

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

# Document or Passage

The image displays three search results side-by-side, illustrating different approaches to information retrieval:

- Universitätsbibliothek HU Berlin Catalogue:** Shows a search for "shakespear death" in their library catalogue, returning 1-10 of 37 results. The results are mostly book titles by William Shakespeare.
- Google Search:** Shows a search for "shakespear death" on Google. The top result is a link to a page about Shakespeare's death, stating he died in 1616. Other results include links to his will, death quotes, and biographies.
- WolframAlpha Knowledge Engine:** Shows a search for "when did shakespeare die?" The input interpretation is "William Shakespeare date of death". The result is Saturday, April 23, 1616. It also provides date formats in Julian, Jewish, and Islamic calendars, and calculates time differences from today (Thursday, October 21, 2010).

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
- 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
  - 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 (e.g. MS-Word help)
  - (Semi-)Structured queries (author=... AND title~ ...)
  - Voice (Siri)
- 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)

The screenshot shows the PubMed search results for the query "Myers-g[au] sequence[ti]".

**Left Sidebar (Circled):** Contains links for Entrez PubMed, PubMed Services, and Related Resources.

**Search Bar (Circled):** Shows the search term "for Myers-g[au] sequence[ti]" and various search controls like "Go", "Clear", and "Save Search".

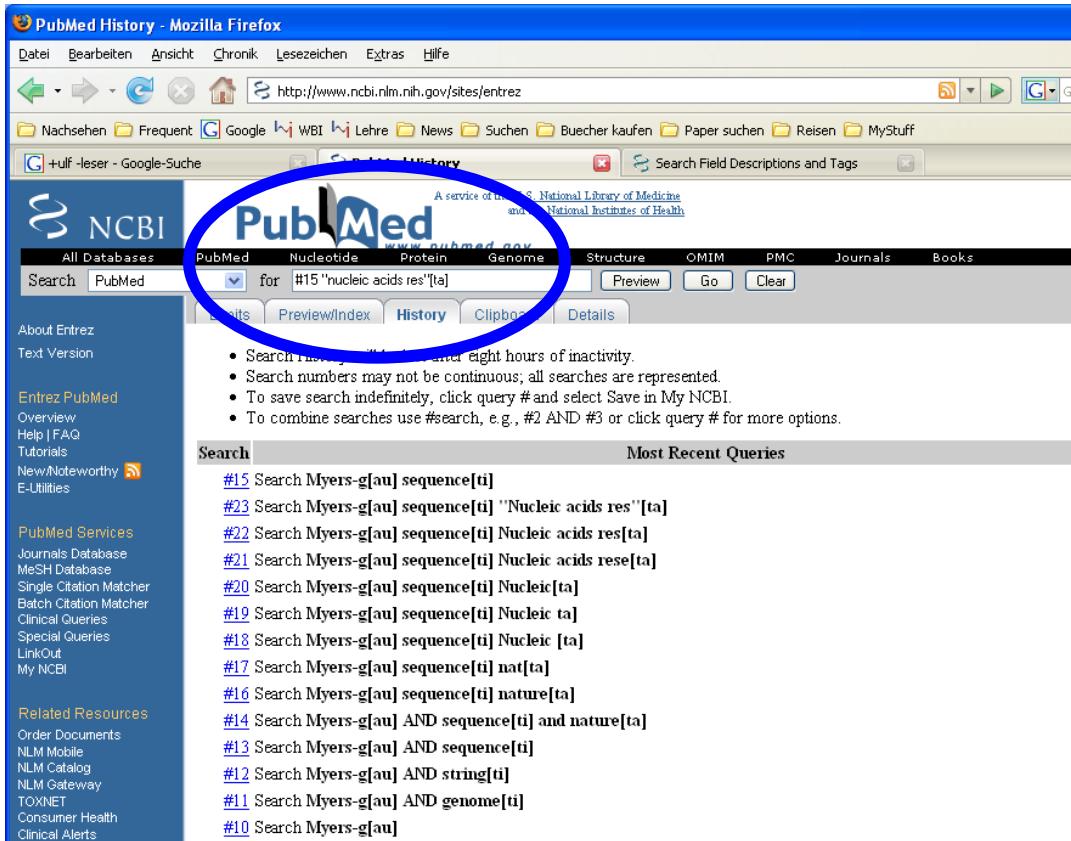
**Right Sidebar (Circled):** Includes "One page.", "Related Articles, Links", and "Links".

**Search Results:** Displays 9 results, with the first one being a genome sequence article by Myers et al. from 2007.

**Search Field Descriptions and Tags:** A table showing various search fields and their abbreviations.

Affiliation [AD]	Issue [IP]	Place of Publication [PL]
Article Identifier [AID]	Journal Title [TA]	Publication Date [DP]
All Fields [ALL]	Language [LA]	Publication Type [PT]
Author [AU]	Last Author [LASTAU]	Secondary Source ID [SI]
Comment Corrections	Location ID [LID]	Subset [SB]
Corporate Author [CN]	MeSH Date [MHDAT]	Substance Name [NM]
EC/RN Number [RN]	MeSH Major Topic [MAJR]	Text Words [TW]
Entrez Date [EDAT]	MeSH Subheadings [SH]	Title [TI]
Filter [FILTER]	MeSH Terms [MH]	Title/Abstract [TIAB]
First Author Name [1AU]	NLM Unique ID [JID]	Transliterated Title [TT]
Full Author Name [FAU]	Other Term [OT]	UID [PMID]
Full Investigator Name [FIR]	Owner	Volume [VIL]
Grant Number [GR]	Pagination [PG]	
Investigator [IR]	Personal Name as Subject [PS]	
	Pharmacological Action MeSH Terms [PA]	

# 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

---

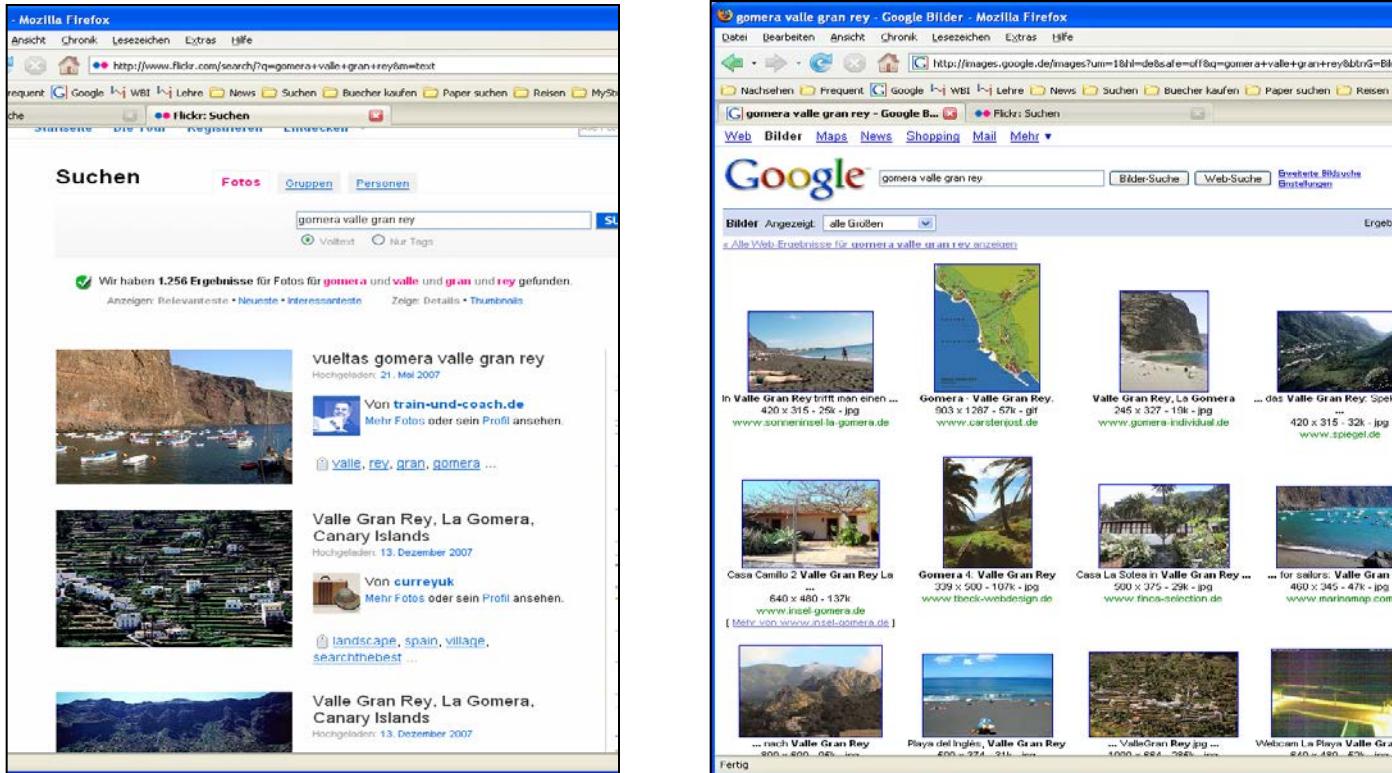
```
<head profile="http://dublincore.org/documents/dcq-html/">
<title>Dublin Core</title>
<link rel="schema.DC" href="http://purl.org/dc/..." />
<link rel="schema.DCTERMS" href="http://purl.org/..." />
<meta name="DC.format" scheme="..." content="text/htm" />
<meta name="DC.type" scheme="..." content="Text" />
<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>
```

# Content of this Lecture

---

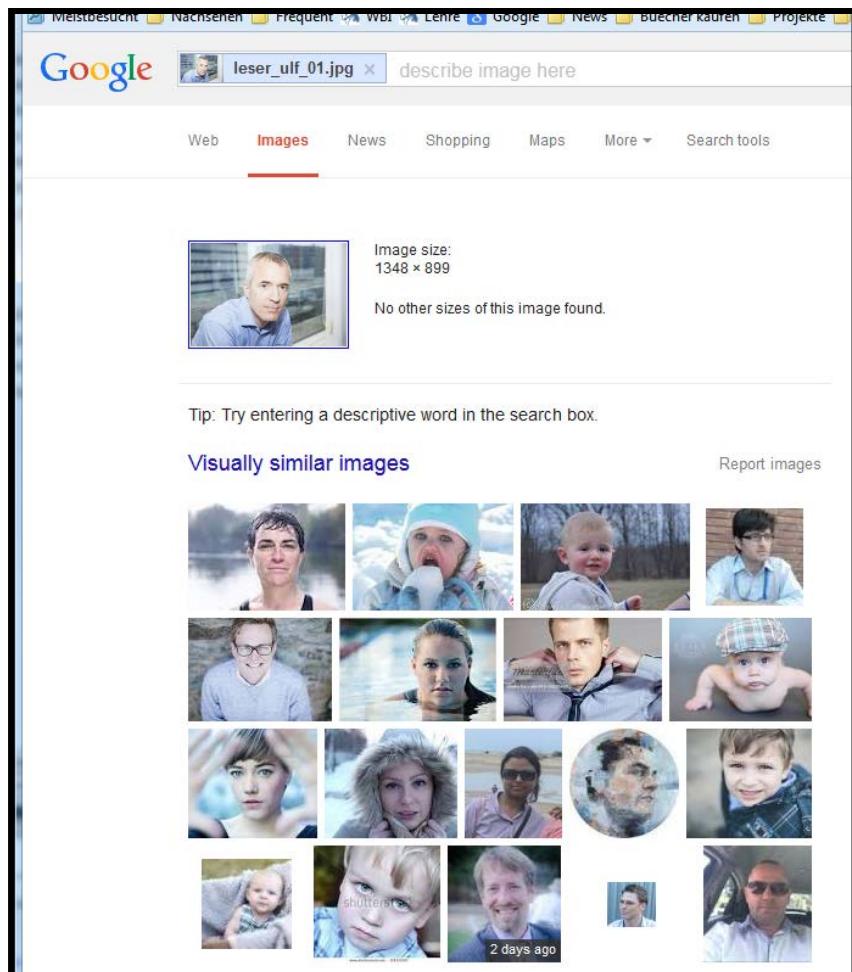
- What is Information Retrieval
- Documents
- Queries
- Related topics

# Multimedia Retrieval

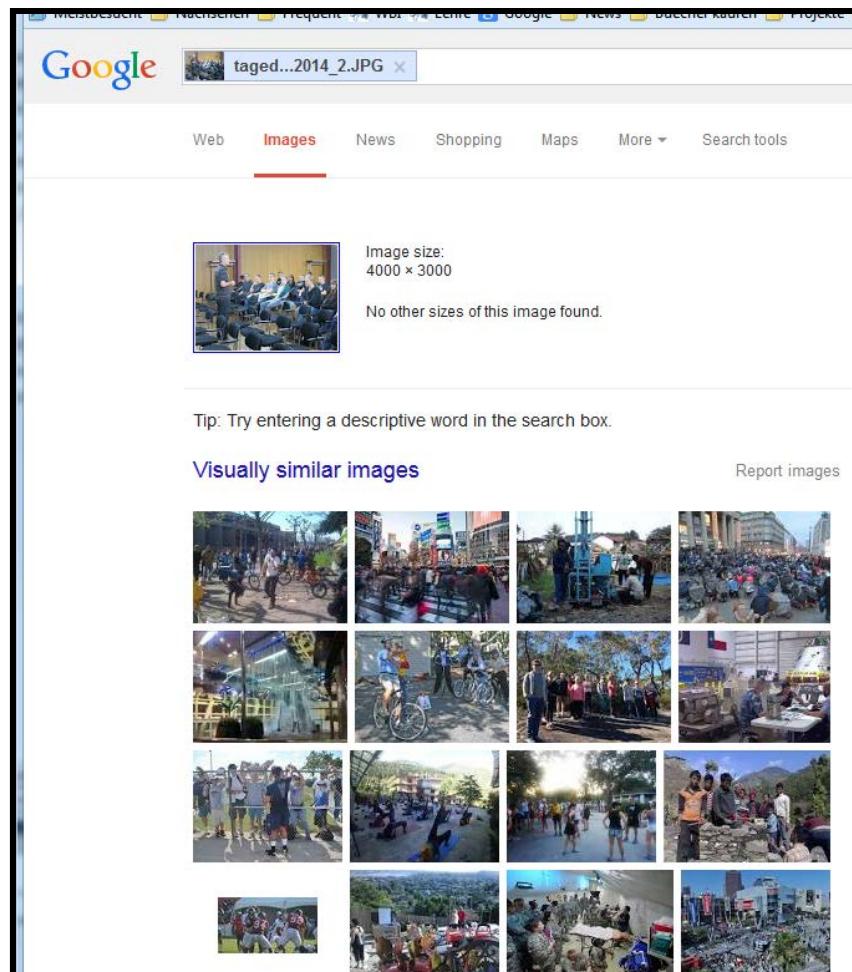


- Note: Neither searches within images
  - Flickr: tags (“folksonomy”)
  - Google: text in neighborhood

# „Search by Image“ (10/2014)



Google Images search results for the query "leser\_ulf\_01.jpg". The search bar shows the image file path. The results page includes a navigation bar with Web, Images (selected), News, Shopping, Maps, More, and Search tools. A thumbnail of the input image is displayed with its dimensions: 1348 x 899. Below it, a message states "No other sizes of this image found." A tip at the bottom suggests entering a descriptive word in the search box. A "Visually similar images" section shows a grid of 20 images, including various people and objects, with a "Report images" link above it.



Google Images search results for the query "taged...2014\_2.JPG". The search bar shows the image file path. The results page includes a navigation bar with Web, Images (selected), News, Shopping, Maps, More, and Search tools. A thumbnail of the input image is displayed with its dimensions: 4000 x 3000. Below it, a message states "No other sizes of this image found." A tip at the bottom suggests entering a descriptive word in the search box. A "Visually similar images" section shows a grid of 20 images, mostly depicting outdoor scenes like people in parks, markets, and urban areas, with a "Report images" link above it.

# 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?
- Most prominent: IBM Watson (2011)
  - “IBM Watson is a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data”
- Hot topic for personal assistants
  - E.g. Amazon Echo, Apple Siri, Google Assistant, ...
- QA: Mixture of statistical NLP, Machine learning and IR



# Historic Texts

---



- SachsenSpeigel, ~1250
  - "Swerlenrecht können 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

---

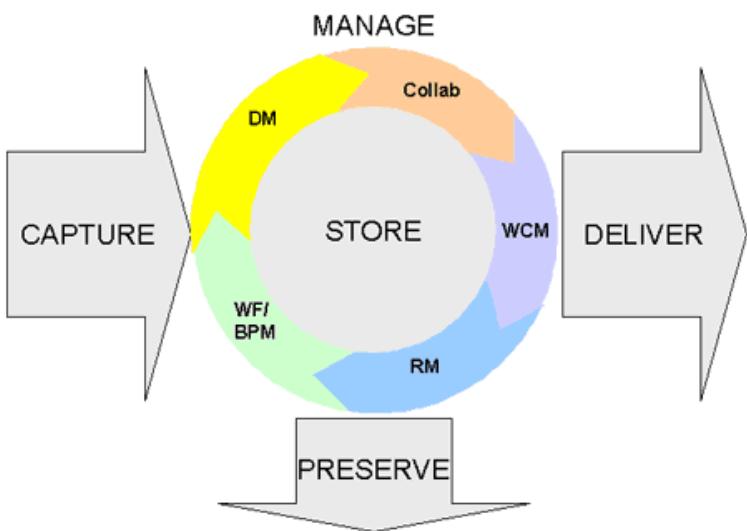
- Document management systems (DMS)
  - Large **commercial market**, links to OCR, workflow systems, etc.
  - Many legal issues (compliance, reporting, archival, ...)
  - Essentially all companies run some form of a DMS
  - Every DMS includes an IR system
- Knowledge management
  - “More sophisticated” 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**“

- 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, ...
- ...



Quelle: AIIM International

# 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, <20M citations, >500K added per year
  - Institute for Scientific Information (**ISI**)
    - 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?