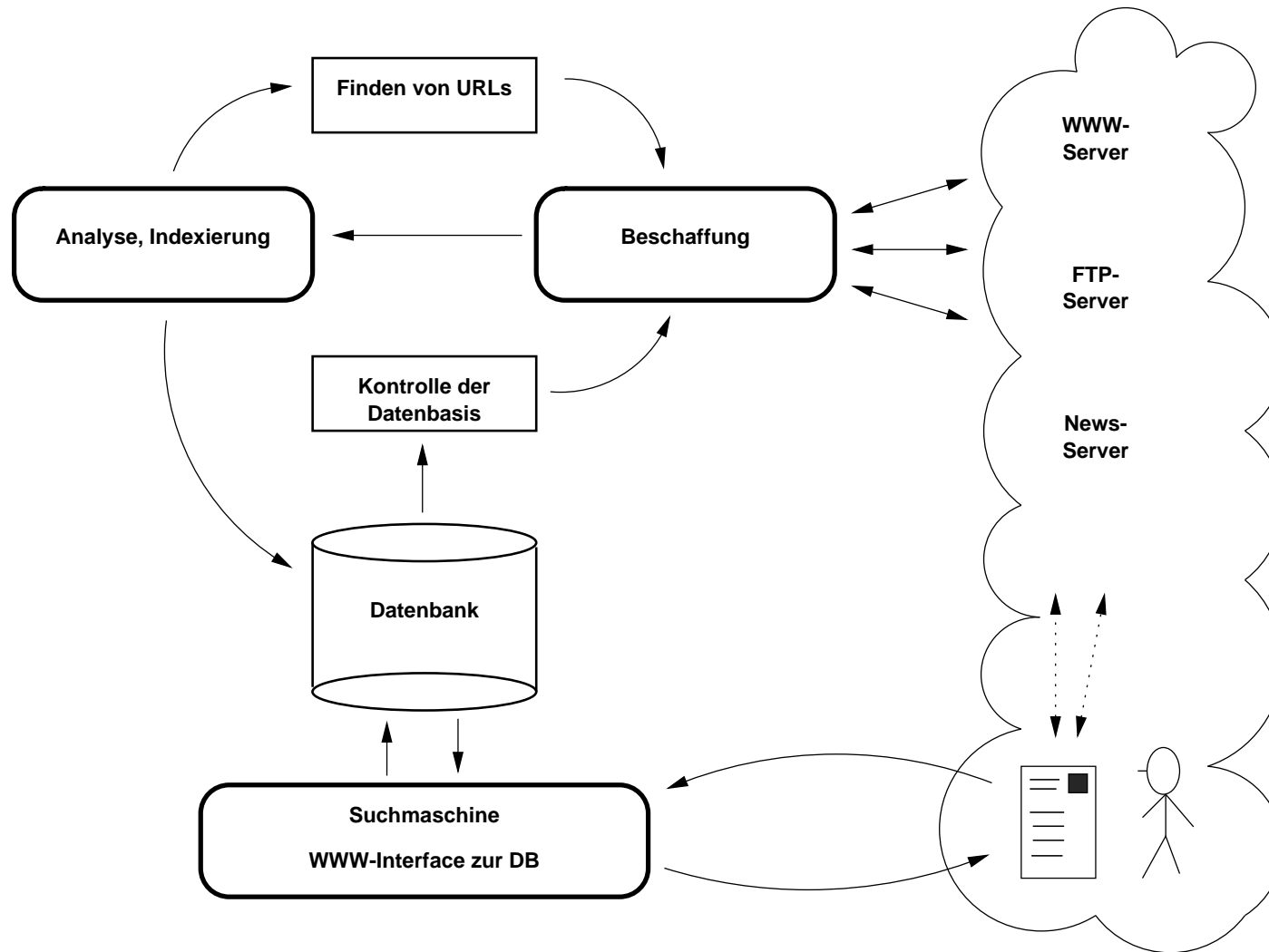


# Funktionsprinzip von Suchmaschinen



# Roboter, Spider, Crawler, Wanderer

Die Jäger und Sammler des WWW

- beschaffen WWW-Dokumente (auch FTP-Inhaltsverzeichnisse, Newsartikel)
- spüren neue Dokumente auf:
  - nutzen Verweis-Technik (Hypertext-Links)
  - Verzeichnisse, News-Gruppen, explizite Anmeldung
- arbeiten sich rekursiv durch's Web

## Was Suchmaschinen nicht finden

- Dokumente (oder Server) ohne Referenz
- geschützte Dokumente
- Dokumente hinter Gateways, in Datenbanken
- neue/geänderte Dokumente
- für Roboter verbotene Dokumente
- dynamische Dokumente

## Hinweise für Serverbetreiber

Minimale Steuerung der Robots: Standard for Robot Exclusion

`http://server.domain/robots.txt`

```
User-agent: ugly-spider      # der darf nichts holen
```

```
Disallow: /
```

```
# alle anderen: keine temporaeren Doks und Scripts
```

```
User-Agent: *
```

```
Disallow: /tmp/
```

```
Disallow: /counter/
```

```
Disallow: /cgi-bin/
```

Aber nicht alle halten sich dran ... :-)

## Analyse und Indexierung

Analyse der Dokumente (Text, HTML, ...) → Datenbank

- ... das große Geheimnis der Betreiber
- große Unterschiede - ständige Weiterentwicklung
- Texterfassung: Volltext oder teilweise (Anfang, wichtige Passagen)
- Häufigkeit, Stellung der Wörter
- Bewertung nach HTML-Tags: Überschriften, Titel, Keywords, Hervorhebungen, Anker, Bilder, Applets, ...
- mathematische Methoden zur Bewertung, um später "ähnliche" Dokumente finden zu können → Relevanz Feedback

## Bearbeitung von Suchanfragen

Entgegennahme der Suchwörter → Analyse → DB-Abfrage →  
Sortierung

- Sortierung der Treffer nach Relevanz → Ranking
  - Anzahl der gefundenen Wörter – ↑ viele Suchwörter
  - ...im Vergleich zur Gesamtlänge – ↑ kurze Dokumente
  - Gesamthäufigkeit einzelner Wörter in DB – ↓ häufig
  - Position der Wörter – ↑ Titel, Keyword, Überschrift, Anfang
  - Abstand der Begriffe – ↑ nah beieinander
- Geschwindigkeit geht vor Exaktheit → nicht reproduzierbare Ergebnisse

## ht://Dig - Intranet searching engine

- htdig: Startseite, max. Linktiefe, Beschränkung auf Domain
- htmerge: Wörter + Informationen in DB (GDBM)
- htsearch: Web-Interface zur Suche in DB
  - Relevanz: Titel, Keywords, H1, ...
  - Boole'sche Verknüpfung, Fuzzy-Algorithmen (Soundex, Metaphone, ...)

Einsatz an der TU:

- Dokumente bis 5 Klicks von TU-Homepage
- ca. 15 000 Dokumente auf 20 Servern, 200 000 Wörter

## Hinweise für HTML-Autoren

- Stabile, langlebige URLs, Hinweise bei Verlegung
- regelmäßige Aktualisierung
- Anmelden bei Suchmaschinen / thematischen Verzeichnissen
- Dokumentenaufbau HTML-konform
- Nutzung von HTML-Tags:

```
<HEAD><TITLE>Aussagekräftiger Titel</TITLE>
```

```
<META NAME="keywords" CONTENT="schlagwort1 ...">
```

```
<META NAME="description" CONTENT="Kurze Beschreibung">
```

```
<META NAME="ROBOTS" CONTENT="NOINDEX | NOFOLLOW">
```

```
</HEAD><BODY>
```

```
<H1>Das ist eine Überschrift</H1>
```

```
<FONT SIZE=7>Das ist große Schrift</FONT>
```