DRI1002 - IKT og informasjonssøking
Uke 5: Databaser og søking i fritekst

Hovedpunkter for forelesningen

- Databaser og fritekstsystemer – en innføring
- Introduksjon til søkemotorer
- Diskusjon

Et kort historisk overblikk

Datamaskinen – en regnemaskin (computer)

- *Digitaliseringen* : Alt representeres ved 0 og 1: **binær lagring** av tall, tekst, lyd, bilder, film,..)
- *Formalisering*: Både h**andlingsregler** og **informasjon** uttrykkes på presis form (matematiske/logiske utrykk)
- *Strukturering* : Organisering av data i bestemte, veldefinerte strukturer

⇓

Strukturerte Databaser

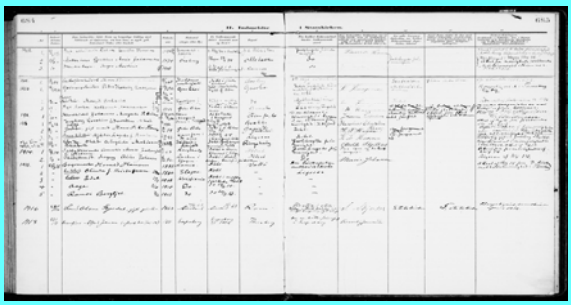
Manuelle databaser –eksempler

- Kirkebøker
- Leksikon, ordbøker
- Kataloger
- Kartoteker,
- Offentlige og private arkiver
- Medlemsregistre
- ....

Alle er karakterisert ved at de har en fast struktur for lagring og gjenfinning av informasjon (data)

Eksempel på manuell database:

Innmelde i statskyrkja i Slagen sokn i Sem 1905-1918



Automatisert behandling av strukturerte data – **Databasesystemer**

Det vokste raskt fram et behov for å beskrive og lagre data elektronisk *på en strukturert form*

De første eksempler på EDB-baserte databaser på 50-60tallet :

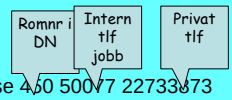
- Folkeregisteret (se //www.ssb.no/
- Skatt- og ligningsdata
- Bankenes og forsikringsselskapers kundekonti
- Medlemsregistre, adresselister,..
- ……

Hvorfor strukturering av data

Dette forstår de fleste:

Arild Johan Jansen, Hofstadgata, 1384 Asker
Dag Wiese Schartum, Harald Løvenskiolds v , 0760 Oslo

Men hva betyr dette :

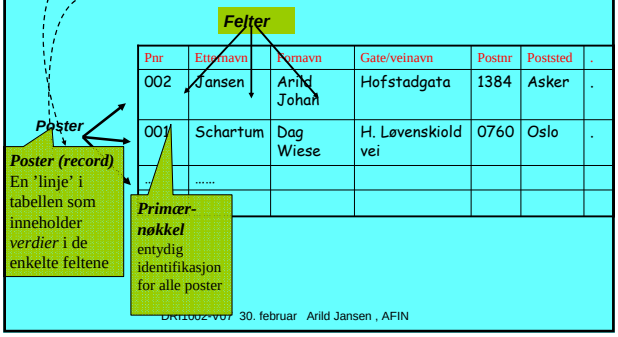


001 Schartum Dag Wiese 450 50077 22733373

002 Jansen Arild Johan 452 50075 66846814

Eksempel på enkel (filbasert) database

Arild Johan Jansen, Hofstadgate , 1384 Asker
Dag Wiese Schartum, Harald Løvenskiolds v , 0760 Oslo



Pnr	Etternavn	Fornavn	Gate/veinavn	Postnr	Poststed	.
002	Jansen	Arild Johan	Hofstadgata	1384	Asker	.
001	Schartum	Dag Wiese	H. Løvenskiold vei	0760	Oslo	.
.	.....					

Noen sentrale begreper knyttet til (filbaserte) databaser

- ***Data :*** et tegn (representert på digital, binær form:
- ***Felt :*** Inneholder et sett/samling av tegn som gir mening, f eks. en ord, tall, dato, klokkeslett,
- ***Post*** *(record) :* En 'linje' i tabellen som inneholder *verdier* i de enkelte feltene
- ***Primærnøkkel*** : et felt som gir entydig identifikasjon for alle poster (f eks. personnr, navn [dersom det gir entydighet)
- ***Fil:*** Poster som hører sammen, f eks. et medlemsregister, katalog, varelageroversikt,..

Men filbaserte databaser utgjør en 'gammeldags' tenkemåte, og vi har andre måter å organisere dataene på

Hva er en database?

Samling med data som er organisert for å tjene et bruksområde. Organiseringen av data er gjort i henhold til en tenkt struktur som beskriver dataenes karakteristikker og sammenhengen mellom dem.

Et *databasehåndteringssystem* (DBMS – data base management system) er et programsystem som laget for *opprette* og *vedlikeholde* databaser

- Eks: Access, Oracle,

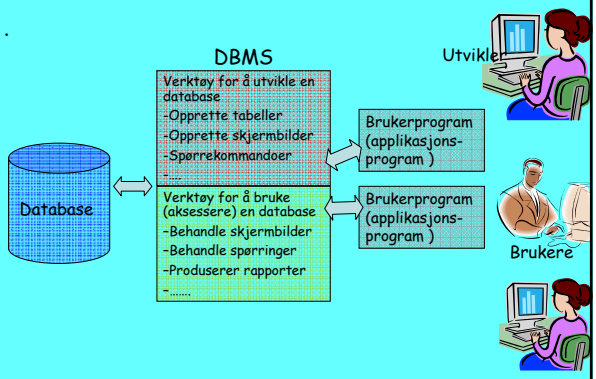
Når vi snakker om tradisjonelle, *strukturerte databaser* mener som regel databaser på tabellform (i motsetning til fritekst-systemer)

Eksempler

: http://www.uio.no/studier/program/

- http://www.uio.no/studier/emnegrupper/
- http://www.uio.no/studier/emner/

"Moderne" databaser



Operasjoner på en database

- *Opprette* en database : definere strukturen
  - **DDL**: data definition language
- *Vedlikeholde* og *bruke* en database : legge inn/rette data
  - **DDM**: Data manipulation language
  - ***Spørring og rapporter***: Hente ut data fra en database, enkeltvis ellers oversikter

Det er utviklet ulike standardspråk for å opprette og bruke databaser (f eks. **SQL** (Structured Query language)

- Opprette: DDL: CREATE TABLE <..>, CREATE INDEX..
- Spørre: SELECT student FROM <tabell> WHERe <betingelse>

Eksempler på bruk av databasesystem

- Her blir det gjennomgått et eksempel på forelesningen – og dette utdypes i gruppetimene
- Se øvelsesoppgaver på siste lysark i denne presentasjonen

Pause – og nytt tema

Datamaskinen ble også en tekstbehandler

- Fritekstsystemer :
  - Med *fritekst* mener vi en vanlig prosatekst inndelt i kapitler, avsnitt og setninger - i utgangspunktet uten spesielle skilletegn og markører. Fritekstsystemer har i Norge blitt brukt til databaser over arkeologisk gjenstandsmateriale, utdrag fra middelalderdiplomer og tingbøker innenfor historiefaget.

  Rettslig materiale er kanskje det felt hvor tekstsøking har blitt mest anvendt i Norge, jf de juridiske databasene hos stiftelsen *Lovdata*.

  (hentet fra Heimen 2/1996 Av: *Gunnar Thorvaldsen, se* *http://www.rhd.uit.no/art/heim96-2.html*

Informasjonssøking

- Computer-aided information search and retrieval
  - historie om lag like gammel som datamaskinene
  - første skikkelige gjennombrudd på 50-talet i samband med søk og erstatt av uttrykk i lovtekst
  - IR = Information Retrieval
- Før WWW har informasjonssøk særlig vært knyttet til databaser og databasesøk, men også enkle fritekstsøkesystemer
- Internett/WWW har endra dette ved søk i store, ustrukturerte datamengder

Søking på Internett i et historisk lynglimt

- I begynnelsen var.... Archie
  - utvikla i 1990 av Alan Emtage, pre-web søkemotor (ftp)
- The World Wide Web Wanderer (Wandex) – første søkerobot på web'en
- Galaxy (1994), den første internett-katalogen
- **WebCrawler** (1994) – første fulltekstindeksering av web
- **Yahoo**! (1994) →
  - 10 år med internett-søk har vist at det skjer raske endringer og mange søketjenester har relativt kort levetid. Yahoo! er en av få tjenester som har vært med heile tida
- **AltaVista** var ei viktig tjeneste fram til slutten av 90-talet. På berre ca. et halvt år forsvant den nesten helt då **Google** tok over

Noen aktuelle søketjenester i dag

- **Google** er i dag den mest brukte og langt mer enn en søkemotor (http://www.google.no/, http://www.google.com
- **Kvasir** (http://www.kvasir.no/
- SESAM : (http://www.sesam.no/
- Yelo: http://www.yelo.no/yelo/
- For oversikter mm , se
  - http://searchenginewatch.com/links/
  - http://www.seoforum.no/index.php
  - http://www.startsiden.no/

Ulike typer søketjenester

- Katalog
  - menneskeskapt hierarkisk database over nettressurser (Yahoo, Open Directory, LookSmart, Kvasir)
- Søkemotor
  - robot, database, brukargrensesnitt mot database (Google, AltaVista, Teoma, Kvasir…)
  - samme søkemotor kan være motor i ulike tjenester (Google blir brukt i Yahoo, AOL, Kvasir…) – outsourcing av søk!
- Metasøkemotor
  - søkemotor som bruker andre søkemotorer som kilde, parallellsøk i mange underliggende baser
    HotBot, *Queryster*, *DogPile, Excite, MetaCrawler, mamma*
- I praksis er i dag de fleste søketjenester en kombinasjon av kataloger og søkemotorer

Hva er en søkemotor ?

- I Søkerobot (*crawler, bot, spider, vevkjerring*)
  - program som følger lenker på veven og kopierer informasjon (tekst) inn i den sentrale databasen
- II Database
  - informasjonen samla av roboten blir lagra i en data-base med en del tilleggsinfo
  - indekseringa i etterkant av informasjonsinnhenting inneber m.a. statistikk over ord, plassering av ord i teksten, analyse av lenker m.m.
- III Søkegrensesnitt
  - brukeren sin interaksjon med søkemotoren
  - enkelt søkefelt eller grensesnitt for avansert søk

Søkemotor og søkeroboter

- Søkerobot
  - ikke en, men mange roboter (program) som traverserer nettet og henter inn informasjon
  - en tjeneste som Google vil vanligvis indeksere en vevtjeneste en gang i måneden
  - søkeroboten leser vevsider som en "primitiv" tekstbasert nettleser

Søkemotor: Søkegrensesnitt

- Søkegrensesnitt
  - Søkeboks for enkle søk
  - Avansert søk med hjelp til avgrensing
  - Problem:
    - Ingen standard for søk i søkemotorar
      - korleis fungerer søket "epler pærer" i Google?
        (finsk undersøking viste at > 60% av brukerene tok feil)
    - Variabel støtte for Boolsk logikk (AND, OR, NOT)
  - For mer informasjon om oppbygging av en søkemotor, les "The Anatomy of a Large-Scale Hypertextual Web Search Engine" av Larry Page og Sergey Brink (grunnleggjarane av Google)

Finst det andre søkemotorar enn Google?

- Fort å se seg blind på en dominerende aktør som Google
  - bør stadig prøve ut andre søkemotorer
  - nisjesøk
    - dersom du søker etter offentlig informasjon i Norge, bør norge.no vera en betre søkemotor enn Google (men ikke heilt sikker på at den er det…)
    - Kvasir er avgrensa til .no-domenet og bør slik sett kunna gi betre resultat enn ei meir omfattande tjeneste (i realiteten er det Google som leverer søkeresultat til Kvasir; skilnaden blir emnekatalogen Kvasir har bygd opp)
    - Startsiden.no tilbyr søk i Google, Kvasir, AltaVista, Yahoo! og Alltheweb (som nå er en del av Yahoo!)
- Google har i dag en for dominerende rolle siden den også blir brukt som motor for mange andre konkurrenter
- Meir informasjon:
  - Search Engines and controversy: http://www.firstmonday.dk/issues/issue9_1/gerhart/

Synlege vevtenester

- For sluttbruker er søkegrensesnittet den synlige delen av søkemotoren
- For tjenestetilbydar er søkeroboten den viktigste delen
  - søkerobotar les vevsider som "primitive" nettlesarar
  - http://www.delorie.com/web/lynxview.html for å sjå korleis søkemotoren les sidene
- Den usynlige (skjulte) vev
  - http://www.hib.no/biblioteket/Kildekritikk.asp#aapen

  eksempel på usynleg side: www.kjornes.no/start.htm

Noen Øvelser i søking

1. Gå inn på http://www.samordnaopptak.no/
   - Finn eksempler på *strukturerte databaser*. Hvordan søker du i disse?
   - Finn eksempler på informasjon som ikke er lagret i form av en (søkbar) database
1. Gå inn på http://www.nb.no/ og søk i databaser
2. Går inn i Lovdata. Søk etter lover mm som omtaler hvilke rettigheter og plikter dere har som studenter
   - Hva fant dere?
   - Hva slag type søking tror dere ble utført (databasesøk, fritekst,…)
3. Bruk minst to åpne søkemotorer og gjør tilsvarende søk, både enkle og mer avanserte
   - Hva fant dere ? Sammenlign ulike resultater fra ulike søkemotorene
   - Forklar hvorfor ble dette forskjellig for ulike søkemotorene. Drøft resultatet.
4. Gå inn på UiO's søketjeneste
   - Hvordan vurderer du denne søketjenesten