DRI 2010
Databaser og fritekstsystemer

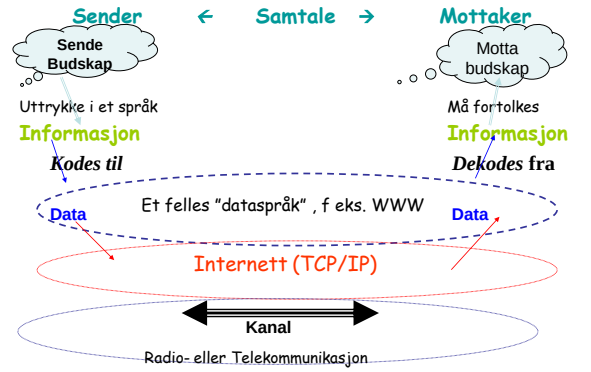
Hovedpunkter for forelesningen
Litt repetisjon fra 1. time
- Om støtteundervisning i INF1000
- Databaser, data og metadata
- Arkiver og offentlige journaler
- Fritekstsystemer,
- Litt om og WWW og HTML

Oppsummering 1. forelesning

Internett – hovedprinsipper
- Modularisering og lag-deling (3 "hovedlag")
  - De fysiske (tele)nettene, kjernen: IP/TCP++, og tjenestene
  - Minimums-løsninger og Ende- til-endeprinsippet
  - IP-adresse og URL
- Vise : Hva skjer når vi kobler oss opp til via et WLAN til Internett og klikker på en lenke

En kort video om IP/TCP mm

Hva er datakommunikasjon – en enkel modell



.

Databaser

Datamaskinen – Både regnemaskin (computer) og tekst-arkiv
Noen grunnleggende egenskaper

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

Strukturerte Databaser :
Systemer for lagring,behandling og gjenfinning av store datamengder

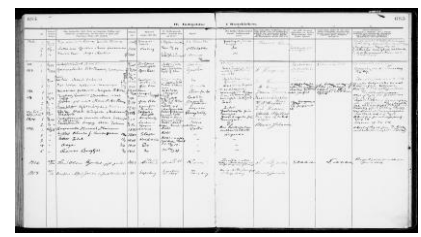
Manuelle databaser –eksempler

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

Hvordan er disse organisert ?

- 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 :

- Befolkningsdata (se f eks.) http://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 :

Romnr i DN | Intern tlf jobb | Privat tlf

001 Schartum Dag Wiese 450 50077 22733873

002 Jansen Arild Johan 452 50075 66846814

Hva er en strukturert 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)

Eksempel på enkel (tabellbasert) database

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

Felter

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

Poster

Poster (record) En 'linje' i tabellen som inneholder *verdier* i de enkelte feltene

Primær-nøkkel entydig identifikasjon for alle poster

Noen sentrale begreper knyttet til (tabell-baserte) 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,..

Tabellbaserte databaser utgjør en 'tradisjonell' tenkemåte, og vi har også andre måter å organisere dataene på (fritekstsystemer, lenkebaserte systemer, hypertekst,…)

Eksempler :

- Vitnemålsdatabasen, tabellene i søkerhåndboka (se http://info.samordnaopptak.no/soeking_opptak/soekerhandboka, oversikter over studier og studenter ved UiO, kontooversiktene hos bankene, medlemsregistre,…)

Data og metadata

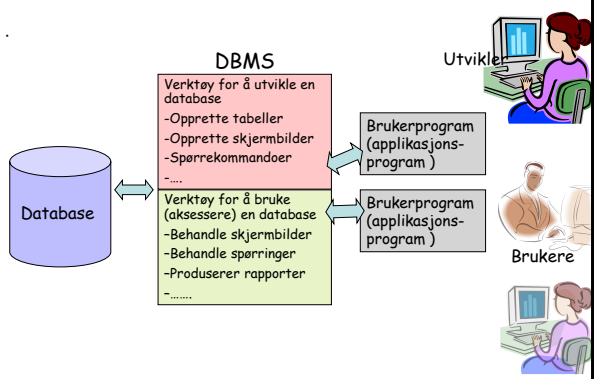
Dataelement: Enhet av data som er udelelig, f eks. f. navn, e.navn. p.nr, telefonnr. ,..

- **Datadefinisjon:** *Type og formatbeskrivelse* av et dataelement
- **Metadata :** Data om dataelementer, inkl. datadefinisjon, dataeierskap, tilgangsrettigheter,.....
  - Metadata brukes både i tradisjonelle (relasjons) databaser og andre typer databaser, f eks. XML-baserte databaser.

Metadata omfatter mer enn [rene]datadefinisjoner :

- Bidrar til å opprette logiske sammenhenger, der de ikke finnes fra før
- Bidrar til å gi opplysninger entydige egenskaper
- Bidrar til å knytte informasjon til informasjonens tilhørende sammenheng

"Moderne" databaser



Offentlige arkiver og journaler

- Hva er et arkiv
- Arkivnøkler - avgrensning
- Arkivnøkler som klassifikasjonssystem
- Arkivnøkler og offentlig informasjon
- Offentlig journal

Hva er et arkiv

- Dokumenter mottatt eller skapt av en virksomhet som en del av virksomhetens virkeområde (også kalt *enke/tarkiv*).
  - Eks: dokumentsamling, brevsamling, osv som er blitt til som ledd i organisasjonens virksomhet
  - Et arkiv er organisert i henhold til virksomhetens formål, definert gjennom eit klassifikasjonssystem - **en arkivnøkkel**

Om offentlig arkiver

- Offentlege organ pliktar å ha arkiv, og desse skal vera ordna og innretta slik at dokumenta er tryggja som informasjonskjelder for samtid og ettertid.
- Eit offentleg organ skal ha ein eller fleire journalar for registrering av dokument i dei sakene organet opprettar
- Offentlige arkiver er regulert av arkivloven og arkivforskriften, se f eks.
  - http://www.arkivverket.no/arkivverket/lover/arkivloven.html

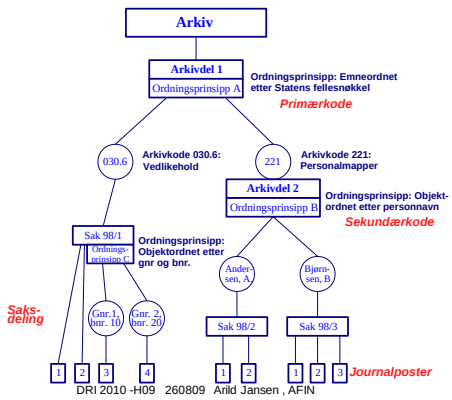
NB: Et Biblioteker er ikke et arkiv, men omfatter bøker og kataloger

Arkivnøkkelen

- Opprinnelig en måte å klassifisere dokumenter på for å kunne fysisk organisere dem i henhold til en rekkeorden slik at man kan finne frem dokumentet
  - "På hvilken reol og hylle befinner dette dokumentet seg"
  - Utgjør en del av et klassifikasjons "scheme" (Egentlig regime, men også system går bra)
- Det finnes en rekke forskjellige typer klassifikasjonsmåter (kronologisk, alfabetisk, temabasert, saksbasert,...)

- *Om Elektronisk journalføring* (Forskriften, §**2-9**.
- For elektronisk journalføring skal offentlege organ normalt nytte eit arkivsystem som følgjer krava i Noark-standarden. Nye system skal vere godkjende av Riksarkivaren før dei blir tekne i bruk.

Arkivstruktur / ordningsprinsipp



Innenfor et journalarkivsystem

- Vil det være flere ordningsprinsipper
- Arkiv
- Arkivdel
- Arkivnøkkel
- Arkivnøkkel som er emneordnet i utgangspunktet
- Suppleres med objektordnede underserier
  - For eksempel gårds- og bruksnummer
  - Navn eller fødselsnummer
- Saknummer
- Sak/Journalpost
- De enkelte dokumentene

Hva inneholder offentlig journal

- Eksempel
- http://www.asker.kommune.no/
- http://www.fredrikstad.kommune.no/
- http://www.regjeringen.no/nb/dep/fad.html?id=339
- http://www.digitalarkivet.no/

- Nytt tema -

Fra Strukturerte databaser til fritekstsystemer

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*.

Litt om organisering av tekstlig informasjon

Et tekstlig dokument kan (blant annet) karakteriseres ved

- *Innhold*: Hva teksten utrykker/formidler,
  - Eks: Roman, dikt, fagstoff, lovtekst, offentlig rundskriv, brosjyre,
- *Struktur* :Måten innholdet er organisert,
  - Eks. Bind, kapitler, avsnitt, nummerering, referanser,…
- *Form/utseende* (Layout, "design")
  - Skrifttyper/størrelser, farger/grafikk, sidestørrelse, spalter, bokser,
- Disse er ikke uavhengige av hverandre

Fra tekst til Hypertekst

Tradisjonelt er en tekst en sekvensielt organisert samling av setninger, avsnitt eller kapitler.
- Referanser skjer gjennom fotnote, sidehenvisning eller

- **Hypertekst** : er et brukergrensesnitt-mønster for å presentere dokumenter som inneholder automatiske kryssreferanser til andre dokumenter (noder), kalt hyperlenker (linker, hyperlinker, lenker, pekere). Når man aktiverer en hyperlenke vil datamaskinen raskt presentere dokumentet det er lenket til. (Wikipedia)
- Hypertekst innebærer at tekstelementer knyttes til hverandre, f eks. gjennom *pekere* eller *lenker* i WWW-terminologi
- HTML tillater en å angi slike lenker, som en kan "klikke på" for få tilgang til et nytt element.

Merking av fritekst : HTML

HTML: Hyper Text Markup Language –

Et standard "språk" for å beskrive layout (format) av et dokument i fritekstformat for presentasjon

HTML-kodene angi hvordan dokumentet skal presenteres:

Et HTML-dokument består av 2 deler (nivåer)

1.: Det vi ser på skjermen

2: Kodene i dokumentet (normalt vises de ikke)

HTML: "Markup –språk "

- Beskriver utseende (layout,format), ikke innhold
  - I HTML markeres ("tagges") tekst for å angi format

Eks: HTML-sekvensens:

…..Vanlig tekst **<B>** uthevet **<B/>** **<I>** kursiv **</I>** **
** ny tekst

blir således :

Vanlig tekst **uthevet** *kursiv* **…**
ny tekst

- HTML består av et bestem sett av markeringer (Tag-typer)
- HTML –setninger kan leses av alle nettlesere (forutsatt at de bruker standard)

Eksempel på HTML - side

Noen grunnleggende HTML-koder:

```
<HTML>
  <HEAD>
   <TITLE>Avdeling for forvaltningsinformatikk</TITLE>
  </HEAD>
 <BODY>
…….. <A href="http://www.jus.uio.no/">JURIDISK FAKULTET</A>
 </body>
</HTML>
```

Hentet fra http://www.afin.uio.no/

Hvordan representeres dokumenter

- Dokumenter har blant annet
  - En **identifikasjon** (vanligvis 'navn', kap i en bok, forfatter/dato),

  Men på Internett ved *URL !*
  - ***Innholdet,*** dvs. teksten vi er interessert i
    - F eks skrevet ved tekstredigeringsprogram : Word, Framemaker, OpenOffice Writer,.....
  - ***Utseende*** – formatet eller layout, dvs. slik det framstår (presenteres) på skjerm eller papir. Her bruker vi **HTML** for å bestemme utseende på WWW.

Et dokument kan ha ulike typer fysisk *representasjon*, f eks. skriftlig, nedkopier på mikrofilm, digital på disk, disketter, Cd-rom,
Et dokument kan *presenteres* (visualiseres) på ulike måter: på papir, på skjermen, på film osv..

World Wide Web – hva er det ?

- Et informasjonssystem ?
- En bok?
- Et (velorganisert) bibliotek ?
- Et leksikon ?

Svaret er kanskje både ja og nei – det kan framstå som alle disse formene – men er egentlig ingen av delene

Organisering av informasjon (data) på WWW

Noen hovedbegreper:

- *Hjemmeside* (home page): Førstesiden (ofte kalt startsiden for et *nettsted*, (web-site)
  - Eks: http://www.uio.no/, http://www.afin.uio.no/
  - En hjemmeside identifiseres ved en **URL** (Unified resource locator)
- *Lenke* : peker til et annet dokument
- *Hypertekst :* tekst som inneholder lenker til andre dokumenter (*URL'er)
- Nettleser (Browser) som henter og presenter filer på WWW ved hjelp av http og HTML.
  - Eks. Internet Explorer, Opera, Firefox,..

Strukturen på ett nettsted