DRI 2010

Databaser, fritekstsystemer, hypertekst og semantisk web.

Hovedpunkter for forelesningen

- Databaser
- Data og metadata
- Arkiver og offentlige journaler
- Fritekstsystemer
- Litt om HTML og XML

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 ved matematiske/logiske utrykk)
- **Strukturering** : Organisering av data i bestemte, veldefinerte strukturer

↓

Strukturerte Databaser

Manuelle databaser –eksempler

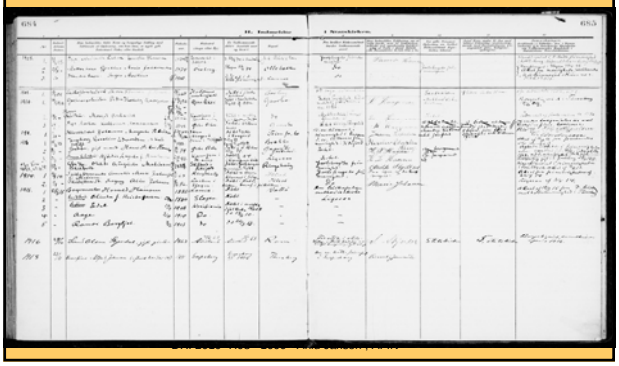
- Kirkebøker
- Leksikon, ordbøker
- Kataloger
- Kartoteker,
- Offentlige og private arkiver
- Medlemsregistre
- ....

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 22733373

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

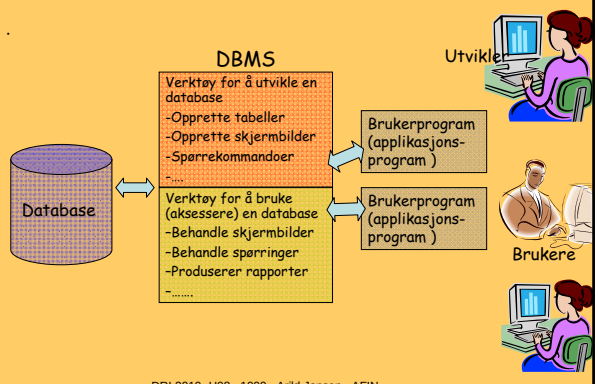
Dataelement: Enhet av data som er udelelig, f eks. f. navn, e.navn. p.nr, telefonnr. ,..

- **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 *enkeltarkiv*).
  - 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

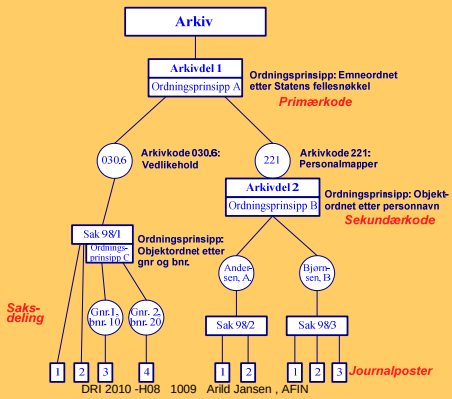
NB: Et Biblioteker er ikke et arkiv, men omfatter bøker og kataloger

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

Hva er viktigst av disse for bøker ??

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 merkes "tagges" tekst for å angi format
    - (Stammer fra boktrykkeriene, eks å markere "ingresss", avsnitt" i margen på en side)

  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)
  - Word kan oversette fra *.doc* format til *.html* (men lager dårlig .html-kode !!!

Hva er HTML- fortsatt

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/

Er HTML tilstrekkelig for å beskrive fritekst

- Layout ?
- Struktur?
- Innhold ?

Kort om XML

- Extensible Markup Language (XML) er enkelt språk for å beskrive dataformater (struktur og innhold, og ikke layout-utseende).
- XML kan brukes til å utveksle data mellom systemer
- XML kan brukes til å lagring av semistrukturerte data, f eks. boktekster, web-sider, …
- XML har en strengere syntaks (grammatikk) enn HTML

Se mer: http://www.w3.org/XML/

- Se eksempler på

http://www.brreg.no/samordning/grunndata/gr1b_basisdatamini.html

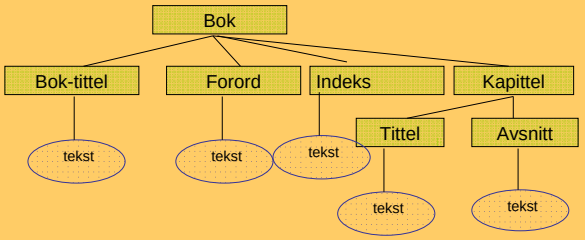
Metadata og hypertekst

Eksempler

- http://www.uio.no/studier/program/
- http://www.uio.no/studier/emner/jus/afin/DRI1001/h08/
- http://www.uio.no/studier/emnegrupper/
- http://www.uio.no/studier/emner/

XML – Extensible markup language

- XML kan beskrive struktur og innhold
- Eks. en beskrive struktur i en bok



Eksempel på XML-kode, inkludert HTML-kode

```
<?XML versjon ="1.0" Encoding = "ISO-8859-1"?>
<bok>
   <desciption>
        <title> Fra kjernen og ut, fra skallet og inn </title>
        <author>
        <first-name> Gerhard </first name>, <Last-name Skagestein</last-name>
        </author>
   </description>
   <body>
        <Forord > I denne boka vil jeg....</forord>
        <chapter title ="Innledning" >
        <p> I dette kapitlet ser vi på ……..
        ……
        </chapter >
        Chapter title ="systemutviklingprosessen"
   </body>
<book>
```

Noen forskjeller mellom HTML og XML

- HTML beskriver bare utseende – ikke hva dataene betyr
- HTML har en løs syntaks (*feil oppdages ikke lett)*
- HTML har et begrenset sett av fast definerte *markeringer* og tilhørende attributter (egenskaper)
- XML kan beskrive både struktur og utseende
- XML har en strengere syntaks
  - Dette gjør at feil kan oppdages før et program brukes
- XML tillater egendefinerte markeringer og attributt-navn

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

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!
  - 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
- Om Google, se f eks. http://en.wikipedia.org/wiki/Google Metasøkemotor

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