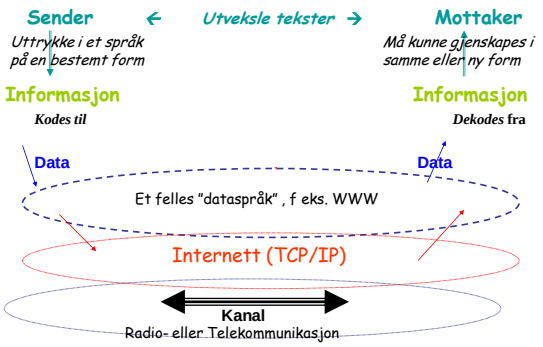
Utviklingen av fritekstsystemer

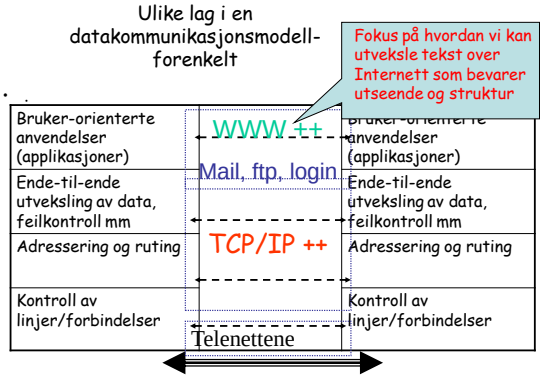
Hovedpunkter for forelesningen

- Litt repetisjon fra 2. time
  - Om støtteundervisning i INF1000
- Ulike typer Fritekstsystemer
- Litt om HTML og XML

Hva er datakommunikasjon – en enkel modell

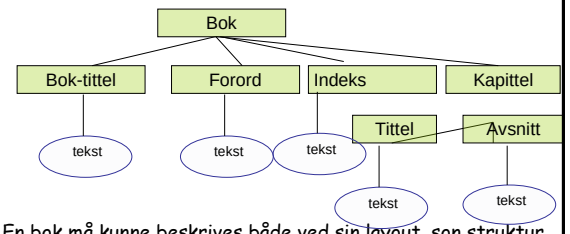


Ulike lag i en datakommunikasjonsmodell- forenkelt



Hva består en bok av – og hvordan representere denne

Eksempel på en typisk struktur på en bok



En bok må kunne beskrives både ved sin layout, son struktur og sitt innhold !

Hvordan kan beskrive denne strukturen ?

- Bruke "Word-format" (.doc) ?
- Bruke .ODT (open tekst) format?
- Bruke .pdf-format?
- Legge det inn i en (strukturert) database?
- Bruke HTML?
- Andre løsninger?

Et alternativ er ***XML*** (Extensible Markup Language)
Et annet alternativ er emnekart (Topic Map)

Data og metadata

- **Data** : [Her forstått som] formalisert representasjon av informasjon i en eller annen form (tekst, lyd, bilde) Data kan være fri tekststrenger eller strukturerte data med bestemt formell betydning

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. i HTML, XML , osv.)

Eksempel :

- Ola Nordmann er født: 05/07/09 : *Data* (vi tror vi forstår)
- Formater på datoer er med mer/dd/yy (amerikansk notasjon) : metadata

- Metadata bidrar til at data blir forstått riktig (skal gi mening)
- Metadata er nødvendig for at datamaskiner skal kunne tolke data

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

Litt om WWW og HTML

World Wide Web – hva er det ?

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

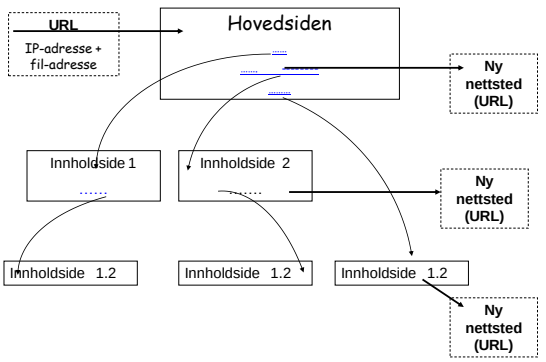
Svaret er kanskje både ja og nei:
Det kan framstå som alle disse formene –
men er egentlig ingen av delene

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



Hvordan representeres dokumenter

- Dokumenter har blant annet
  - En **identifikasjon** (vanligvis 'navn', kap i en bok, forfatter/dato), På Internett ved *URL !*
  - ***Utseende*** – formatet eller layout, dvs. slik det framstår (presenteres) på skjerm eller papir. Her bruker vi **HTML** for å bestemme utseende på WWW.
  - ***Strukturen*** – Hvordan et dokument er organisert i ulike deler , f eks. bok: Tittel, forfatter, innholdsliste, kapitler, sider,..
  - ***Innholdet***, dvs. teksten vi er interessert i
    - Innholdet kan skrives ved tekstredigeringsprogram (Word, OpenOffice Writer, men det gir lite/ingen informasjon om hva innholdet betyr

Et dokument kan ha ulike typer fysisk *representasjon*, f eks. skriftlig, nedkopier på mikrofilm, digital på disk, disketter, Cd-rom,

Det kan formateres på ulike måter

Det kan organiseres på ulike måter .

Men innholdet (betydningen, meningen) skal ikke endres

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-sekvensen:

  .....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 ? JA
- Struktur? Begrenset
- Innhold ? Svært lite

- HTML er primært en markeringsnotasjon, som beskriver hvor et dokument skal se ut

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

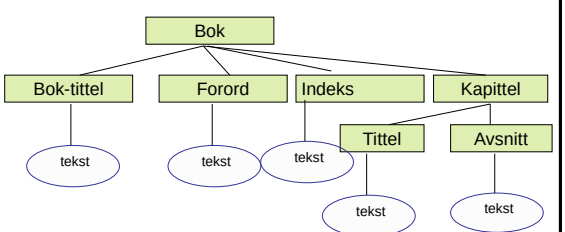
Se mer: http://www.w3.org/XML/

- Se eksempler på

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

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>
```

(fra Skagestein, kap. 17 forenklet. Fargene er for å synliggjøre teksten)

Eksempler på metadata på emnesidene

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/

Beskrivelse av innholdet i tekst bruk av ontologier

- Ontologi : an **ontology** is a formal representation of a set of concepts within a domain and the relationships between those concepts. (http://en.wikipedia.org/wiki/Ontology_%28computer_science%29)
  - Eks på ontologier et Linnee's plantelære, begreper innen ulike deler av matematikken, osv.
- Det blir nå definert ontologier for å kunne beskrive innholdet i tekster på ulike fagområder (domener), f eks. for å beskriver lover, eller dokumenter i forvaltningen ved stardard begreper
  - Eks. på XML-baserte ontologier er LegalXML, Government XML
- XML-baserte ontologier inngår som en sentral del av den semantiske veven

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

DRI 2010 -H09 90909 Arild Jansen , AFIN

Meget kort om emnekart

- **Emnekart** (eng. *Topic Maps*) er en ISO-standard for representasjon og utveksling av strukturert og semistrukturert informasjon.
- Et emnekart består av et sett med *emner* av forskjellige typer. Disse emnene er knyttet sammen i en grafstruktur gjennom *assosiasjoner*. Et emne representerer et *tema*
- Eksempler på bruk av emnekart i Norge:
  - GREP, Se http://www.utdanningsdirektoratet.no/grep
  - Regjeringen.no, se http://www.regjeringen.no/nb.html?id=4

Bruk av emnekart vil bli gjennomgått på forelesningen 23.9