

## Databasestøttet Webpublicering, forår 2002

### Forelæsning 11

#### Publication, User Tracking, and How to add images to a site

- Publication og Annoncering af web-site
  - Exponering af web-site med søgemaskiner
- User Tracking: Hvor kommer brugerne fra?
  - Statistik fra access-log
  - Banner-add click-throughs
- Publicering af billeder på web-sites
- Klikbare billeder — server-side image maps og client-side image maps
- Introduktion til Øvelse 10

#### Søgemaskiner

En søgemaskine består typisk af tre dele:

- web-crawler
- database of URLs
- query processor

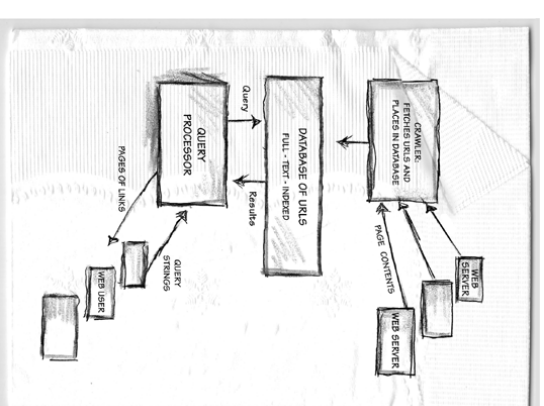
Databasen indeholder bl.a. frekvenstabeller over hvor tit ord forekommer i URL'er:

F.eks. vil teksten "**Home page for the course**"

generere følgende frekvenstabell:

Ord	Frekvens
Home	1
page	1
course	1

**Mange brugere vil komme fra søgemaskiner!**



#### Publicering og annoncering af webside

Mange muligheder:

- Bannerreklamer — specielt i søgemaskiner (køb af ord)
  - <http://www.google.com>
- Prime time TV
- Aviser/blade
- Web-biblioteker
  - <http://www.submit-it.com>
- Søgemaskiner

#### Hvordan kommer man øverst i søgeresultater?

- Det gælder om at stå øverst i søgemaskinens resultat
- Søgemaskiner sælger ud af ord til dem der vil betale penge for at stå øverst i søgeresultater

#### Du kan forbedre dine chancer for at stå øverst

Hvis du ikke har råd til at købe ord kan du sørge for at der er indhold på de sider du vil have indekseret  
Søgemaskiner forstår ikke billeder — heller ikke hvis du har betalt et reklamebureau kr. 50.000 for dem!

#### Uærlig forbedring af dine chancer:

```
<META name="keywords" content="sex sex money fast money  
money money money money money money fast fast sex">
```

Det kan nok ikke betale sig at tynde sin side med tonsvis af *keywords*

#### Ærlig forbedring af dine chancer:

```
<META name="description" content="Journal for sophisticated  
web publishers, speciallizing in RDBMS-backed sites.">
```

## Indhold kan skjules for søgemaskiner!

Nogle gode grunde til at skjule indhold:

- Mirror-sites
- Privat dokument delt blandt få personer
- Du er ved at redigere et dokument der allerede ligger on-line
- Intranettet

*Hvordan man skjuler indhold med wile:*

Det er muligt at instruere web-crawlere om ikke at søge i bestemte filer.

The Robots Exclusion Protocol: (robots.txt, som skal ligge i roden af dit site):

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /tmp/
Disallow: /_joe/
```

The Robots META tag:

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

Mere info: <http://info.webcrawler.com/mak/projects/robots/robots.html>

Bemærk: Der er ingen garanti for at søgemaskiner overholder dine retningslinjer!

## Indhold kan skjules for søgemaskiner — fortsat

*Til tider skjules indhold på trods af at det ikke var hensigten:*

- Hvis du kræver registrering
- Hvad med Tcl-filer og data i databasen?

Løsning:

- Eksporter data til dummy HTML-filer som web-crawlere kan se
- Konstruer indexes fra disse sider til relevante Tcl-filer, således at brugerne ser den rette information.

## Total Exposure

Kom på forsiden af

- [www.cnn.com](http://www.cnn.com)
- [www.newyorktimes.com](http://www.newyorktimes.com)
- [www.dr.dk](http://www.dr.dk)
- ...

## User Tracking: Hvor kommer brugerne fra?

— og hvor mange hits er der på mit web-site?

Se webserverens access-logs:

```
$ wc -l /web/login/log/access.log
1143 access.log

$ less /web/login/log/access.log
194.237.174.86 - - [12/Apr/2000:01:47:35 +0200] \
"GET /faerdig.html HTTP/1.0" 200 1131 "" "AltaVista V2.3A crawler@evreka.com"
194.237.174.86 - - [12/Apr/2000:03:37:35 +0200] \
"GET /robots.txt HTTP/1.0" 404 212 "" "AltaVista V2.3A crawler@evreka.com"
...
```

## Statistik fra access-log

Følgende oplysninger kan hentes fra en web-server's access-log:

- Typen af browser en bestemt bruger benytter
- Antal brugere som har efterspurgt ikke-eksisterende filer — og hvor de har URL'erne fra
- Antal brugere der efterspørger en bestemt fil
- Tiden en bruger i gennemsnit bruger på en fil før brugeren fortsætter med en anden fil
- Antal brugere der klikker på bestemte banner-ads
- Kommer en bruger tilbage?

Se filen /web/login/log/access.log på [hug.it.edu](http://hug.it.edu) — udsøgt i Login med dit brugernavn

### Ikke-eksisterende filer

Eksempel: bruger indtaster en forkert URL direkte i browserens "location-bar"

Søg efter 404 (File Not Found) i access-log:

```
130.226.141.250 - - [17/Feb/2000:15:51:29 +0100] "GET /temperatur.html
HTTP/1.1" 404 212 "-" "Mozilla/4.0 (compatible; MSIE 5.0; Windows 98; DigExt) "
```

- Filnavnet temperatur.html skulle have været temperature.html

- Brugeren benytter Internet Explorer (MSIE 5.0) på en Windows98 maskine

- Med nslookup kan det ses at 130.226.141.250 svarer til stud250.itu.dk:

```
22:49-S111]# nslookup 130.226.141.250
Server: ns000.worldonline.dk
Address: 212.54.64.170
```

```
Name: stud250.itu.dk
Address: 130.226.141.250
```

- Brugeren har højst sandsynligt tastet URL'en direkte i browseren — ingen "røfterer" side.

### Antal brugere som har efterspurgt en bestemt fil

Bestem antallet af linjer i access-log indeholdende f.eks.

```
GET /slideExtractor/slide_extractor.tcl
```

Det gøres således:

```
[21:41-Log]# grep 'GET /slideExtractor/slide_extractor.tcl' access.log* | wc -L
1609
```

Her er en af linjerne (med ekstra linjeskift):

```
130.226.133.160 - - [05/Feb/2001:11:05:55 +0100]
```

```
"GET /slideExtractor/slide_extractor.tcl?presentation_id=48903
HTTP/1.0" 200 3428
```

```
"http://www.itu.dk/courses/W2/F2001/"
```

```
"Mozilla/4.73 [en] (X11; U; Linux 2.2.14-12 1686) "
```

- Med nslookup kan det ses at 130.226.133.160 svarer til vip160.itu.dk — det er en lokal bruger

- Brugeren er sendt fra siden <http://www.itu.dk/courses/W2/F2001/> den 5. februar 2001 kl. 05.55

- Tiden for et hit, tilsammen med de øvrige linjer, kan bruges til at bestemme hvor lang tid en person bruger på en side

### Ikke-eksisterende fil p.g.a. forkert link

Vi søger igen efter 404 (File Not Found) i access-log:

```
213.237.71.166 - - [20/Mar/2001:02:10:46 +0100]
```

```
"GET /F2001/lec8/list2v.tcl HTTP/1.0" 404 456
```

```
"http://hug.itu.dk:8077/slideExtractor/slide_extractor.tcl?presentation_id=49403"
"Mozilla/4.73 [en] (X11; U; Linux 2.2.14-12 1686) "
```

- Filen /list2v.tcl skulle have været /listv2.tcl

- Brugeren benytter Netscape v. 4.73 på en Linux maskine (kerne v. 2.2.14-12).

- Med nslookup kan det ses at 213.237.71.166 svarer til

```
213.237.71.166.adsl.suoe.worldonline.dk.
```

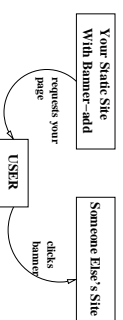
- Fejlen skyldes et forkert link fra en Wimp/Slide — fejlen er nu rettet!

— Vi kan altså se hvilken anden side brugeren kommer fra

— Dvs: vi kan se hvilke sider der indeholder links der ikke virker

### Banner-add click-throughs

- Banner-add ejer har en anden opfattelse af antal click-throughs end din access-log indikerer.



Det er kun access-loggen ved Someone Else's Site der ved hvem der klikkede et banner på din side (via "referrer").

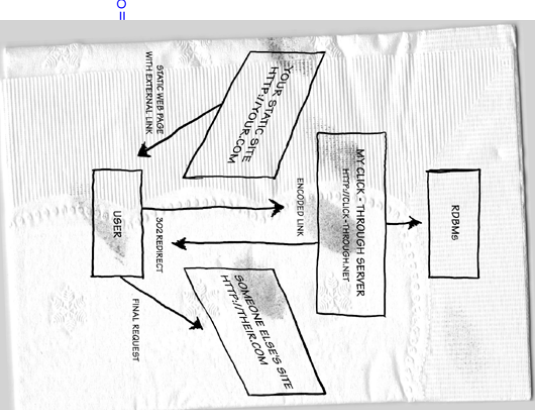
Men det er jo Someone Else's Site der skal betale dig!

- Brug en uafhængig click-through server - se figur.

I stedet for at linke <http://their.com>, så anvend [http://clickthrough.net/your?send\\_to=http://their.com](http://clickthrough.net/your?send_to=http://their.com) fra din side

1. Click til their.com fra din side registreres i databasen

2. Derefter læses en "redirect" til their.com.



## Publicering af billeder på web-sites

Ved en god scanning opnås følgende:

- Originalen (negativ) bliver overflødig
- God billedkvalitet i mange formater
- Et billede kan hurtigt findes ved tekst-søgning (billedbørs)

Brugere ønsker ikke altid billederne præsenteret i høj opløsning!

- Anvend thumbnails til bedre formater — <http://db.photo.net/stock>
- Der skal være mulighed for at se en billedtekst
- Der skal være mulighed for at se tekniske detaljer om billedet
- Der skal være mange formater at vælge imellem
- Det hele skal automatiseres og administreres af en database.

Tekst på en side skal kunne vises også når nogle billeder mangler:

- HTML WIDTH og HEIGHT attributes gør det muligt at tekst læses hurtigt — før billeder
- Programmet WWWis (ftp://www.bloddyeck.com/wwwis/) indsætter automatisk WIDTH og HEIGHT attributter i web-sider

## Opbygning af billedarkiv

```
create table photos (
  photo_id          integer not null primary key,
  photocd_id        varchar(20) not null references photo_cds,
  cd_image_number   integer,      -- will be null unless photocd_id is set
  filename_stub     varchar(100), -- we may append frame number or cd image number
  caption           varchar(4000),
  tech_details      varchar(4000)
);
```

Til hvert billede registreres bl.a. en billedtekst og tekniske detaljer omkring billedet, f.eks. hvilket kamera og film billedet er taget med.

### Eksempel på link til et billede

```
<a href="http://www.photo.net/photo/pcd1765/bearfigh-2.tcl">
</a>
```

## Opbygning af billedarkiv

Se <http://www.photo.net/stock/>.

```
create table photo_cds (
  photocd_id        varchar(20) not null primary key,
  -- bit vectors done with ASCII 0 and 1; probably convert this to
  -- Oracle 8 abstract data type
  jpeg_resolutions char(6),
  -- on which resolutions to write copyright label
  copyright_resolutions char(6),
  copyright_label     varchar(100),
  add_borders_p      char(1) check (add_borders_p in ('t','f')),
  sharpen_p          char(1) check (sharpen_p in ('t','f')),
  -- how this will be published
  url_stub           varchar(100) -- e.g., 'pcd3735/'
);
```

Der er bl.a. information omkring hvilke billed-formater der skal copyright på.

Der anvendes den samme copyright på samtlige billeder på den samme PhotoCD.

## Klikbare billeder — server-side image map

Brugeren klikker på et billede og  $x$ ,  $y$  koordinaten returneres til serveren

```
<a href="ismap.tcl"></a>
```

Man skal blot tilføje attributen ismap til tagget <img>

Serveren returnerer  $x$  og  $y$  koordinaten adskilt af et komma:

- Hvis bruger klikker koordinat 0,0 modtager serveren ismap.tcl?0,0

- Hvis bruger klikker koordinat 42,47 modtager serveren ismap.tcl?42,47

Regulært udtryk der finder henholdsvis  $x$  og  $y$  koordinaten:

```
([0-9]*),([0-9]*)
```

Koordinaten 0,0 svarer til øverste venstre hjørne.

*Eksempel* — filen ismap.tcl:

```
return page "ISMap Example"
<html>Please click the picture</html>
<a href="ismap.tcl"></a><p>
<b>x-coord</b>=[get_ismap_coord x].<p>
<b>y-coord</b>=[get_ismap_coord y].<p>
```

Se <http://hug.it.edu:8077/E2001/lec11/ismap.tcl>

```

Procedureren get_ismap_coord — filen ismap.tcl

# get_ismap_coord x returns the x-coordinate
# get_ismap_coord y returns the y-coordinate
proc get_ismap_coord {xy} {
    set form [ns_getform]
    if {$form == ""} {
        return ""
    }
    set form_size [ns_set size $form]
    set form_counter 1
    while {$form_counter < $form_size} {
        set coord [ns_set key $form $form_counter]
        if {[regexp {^(0-9)*} $coord all_x_coord y_coord]} {
            # Found form-variable of type pair, that is xxx.yyy
            if {[string compare $xy "x" ] == 0} {
                return $x_coord
            } else {
                return $y_coord
            }
        }
        incr form_counter 1
    }
    return ""
}

```

- Vi søger gennem alle form-variabler
- Vi stopper hvis vi finder en form-variabel der matcher et par af tal, f.eks. 34,25
- Det vigtige er at forstå hvorledes den kan anvendes!

**Klikbare billeder — client-side image map**

Vi kan opdele et billede i områder:

- firkanter, `rect(x1, y1, x2, y2)`, hvor  $(x_1, y_1)$  er øverste venstre hjørne og  $(x_2, y_2)$  er nederste højre hjørne.
- polygoner, `polygon(x1, y1, x2, y2, x3, y3, ...)`
- cirkler, `circle(x, y, r)`, hvor  $x, y$  er centrum og  $r$  er radius.

Metode: Find nogle områder på figuren og skriv koordinaterne ned.

```


<map name="pz55map">
  <area shape=rect coords="10,90,92,109" alt="Årsend">
  href="\usemap.tcl?text=rectangle\">
  <area shape=circle coords="137,53,10\" alt="\Bøf\">
  href="\usemap.tcl?text=circle\">
  <area shape=polygon coords="98,57,126,57,111,43\" alt="\Didim\">
  href="\usemap.tcl?text=polygon\">
</map>

```

Vi definerer områder med `area`, og angiver *områdetype* med `shape`.

Vi definerer en opdeling af et billede med `map`, som vi giver et navn, f.eks. `pz55map`.

Se <http://hug.it.edu:8077/E2001/lec11/usemap.tcl>

- Web-sites i fremtiden**
- Web-browsere vil findes overalt: telefoner, komputere, køleskabe, tog-stationer...
- Folk vil foretrække web-baserede programmer fremfor desktop-programmer fordi:
- Samarbejde er lettere med web-apps
  - Web-apps frigør folk fra den faste stol!
  - Man skal ikke være sin egen systemadministrator

**“A Future So Bright You’ll Need to Wear Sunglasses”**

- Introduktion til Øvelse 10**
- Konstruktion af et web-baseret pladekartotek**
- <http://www.itu.edu/courses/W2/F2002/Lb/Lb10.html>
- Eksamen**
- Pensum er kolonnen “Reading” på forelæsningsplanen på hjemmesiden, sldes og øvelser
  - Eksamenssæt for E2000 gennemgås ved næste forelæsning:
- [http://www.itu.dk/courses/W2/F2001/exm\\_e2000.html](http://www.itu.dk/courses/W2/F2001/exm_e2000.html)