HTML, CGI & Search Engines

Hypermedia & Multimedia,

Autumn 1996.

Jon Bomme & Kåre Kjelstrøm

Overview

- WWW: History, idea and implementation
- HTML: Links, frames and forms
- CGI & Search Engines

WWW Timeline

1989: Tim Berners-Lee proposes the World Wide Web project.

1990: First webserver and -client ready by December.

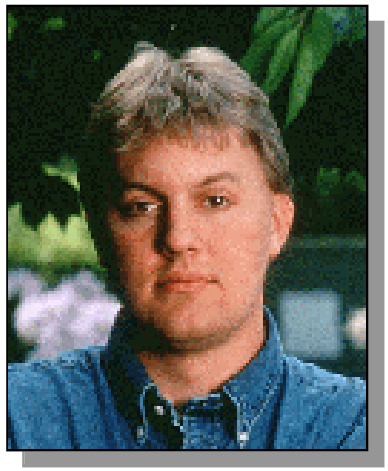
1991: WWW software available for download on the Internet.

1991-93: Protocols de- and refined.

WWW Timeline

1993: Marc Andreessen creates Mosaic at NCSA.

1994: Tim Berners-Lee elected director of the W3 Consortium.



WWW Timeline

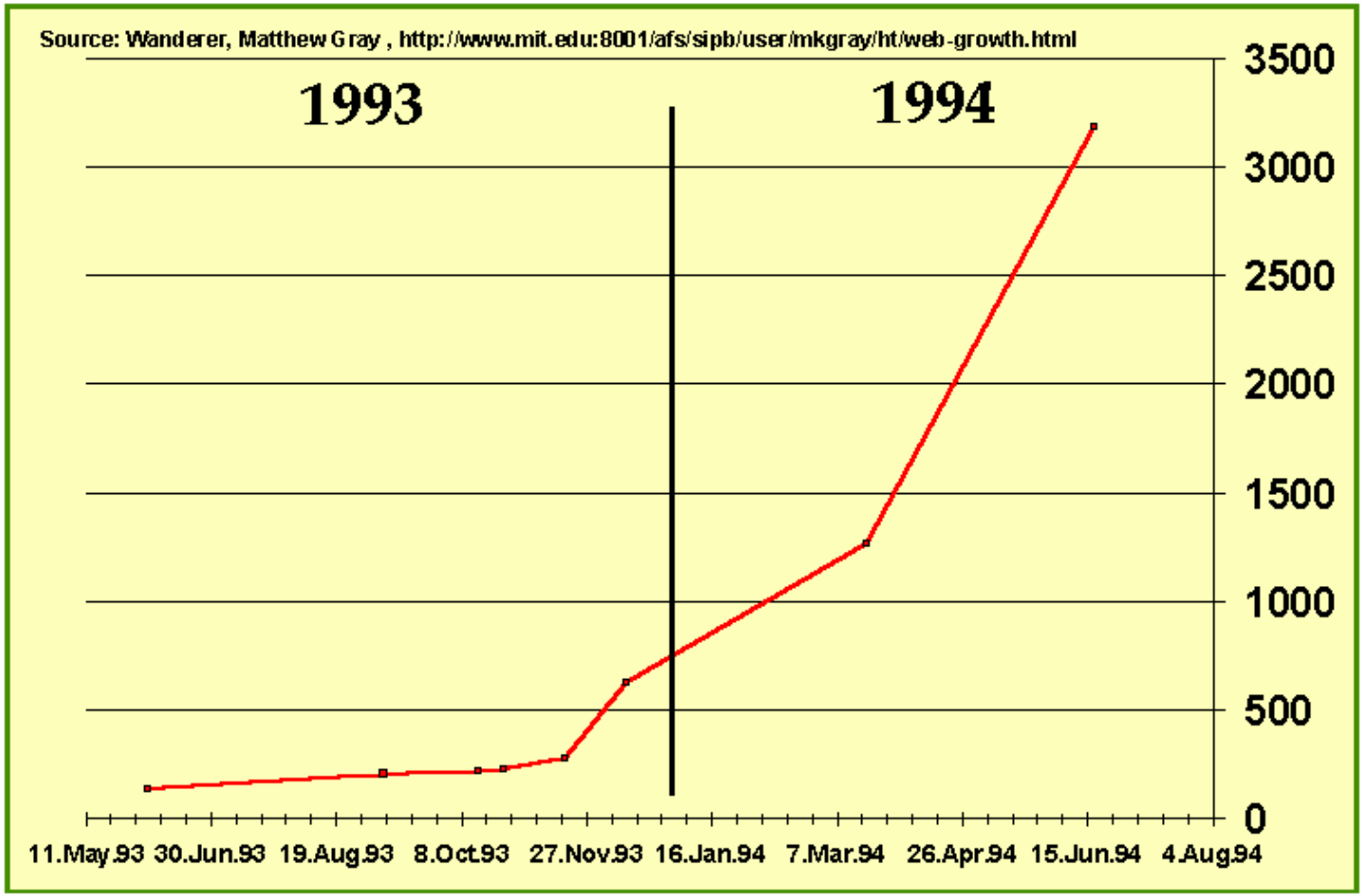
1994: Marc Andreessen & James Clark founds Netscape.

1995: Digital opens Altavista to the public on September 15.

1996: The Internet Community runs out of IP-addresses.

1996: Tim Berners-Lee awarded Distinguished Fellowship of The British Computer Society.

WWW Server Growth



URL-Space

Goals:

- Universal access to anything from anywhere.
- Cooperative work in a ‘self managing team’.

Means:

- Integration of existing protocols (ftp, nntp ..)
- Defining the concept of URLs.
- Invention of the HTML-language.

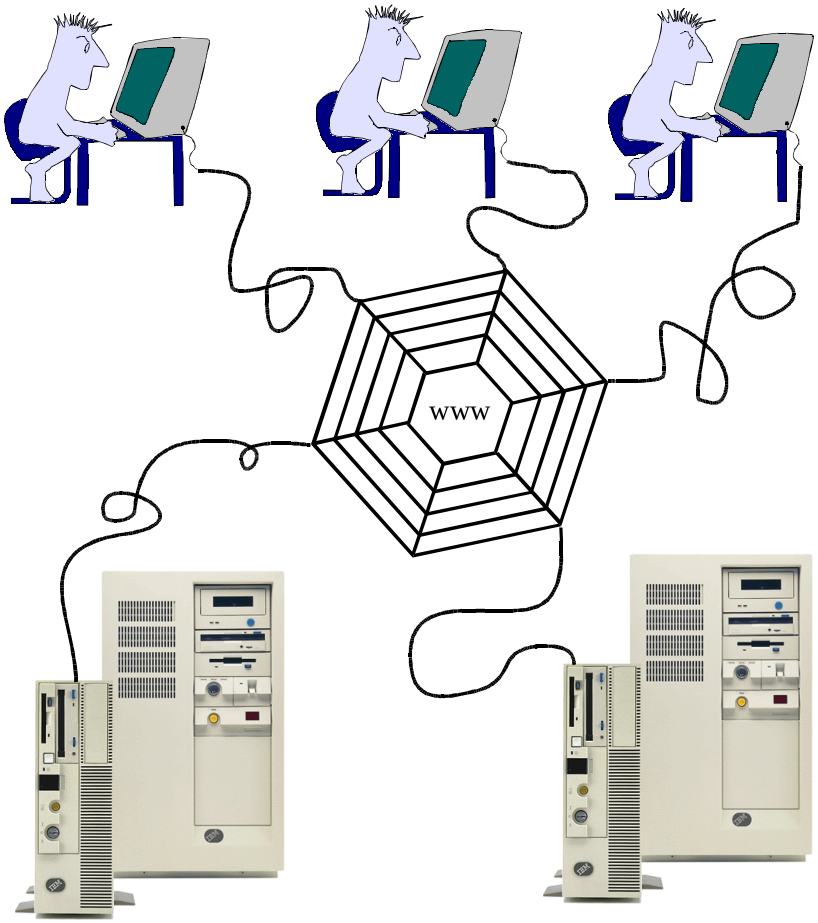
The 7 Issues

1. Search & Query:	Search Engines.
2. Composites:	Frames.
3. Virtual Structures:	No support (CGI)
4. Computation:	CGI, Java.
5. Versioning:	No support.
6. Collaborative Work:	No support.
7. Extensibility:	Good support.

The WWW Architecture

- Distributed Client/Server model.
- Servers provide data as hypertext or in any other format.
- Clients take care of querying the server for hypertext and then displaying this.
- Each URL generates a connection that is closed when data has been transferred.

The WWW Architecture



The structure of HTML

- HTML is SGML.
- Consists of tags and plain text.
- Tags structure: `<name arg1=value arg2=value ...>` … `</name>`
- Some tags don’t have endtags.
- The W3 consoritum defines the HTML standards.

The structure of HTML

```
<HTML>
  <HEAD>
  </HEAD>

  <BODY>
  </BODY>
</HTML>
```

Information about the document

The actual content of the document

Links in HTML

- Links are unidirectional.
- Links are embedded in the sourcecode i.e. no link-object exists.
- Links can refer to anything, even nonsense.
- Links point from within a document to within another document.
- Links are untyped.

Links in HTML

Links are made via the anchor-tag <A>. Some examples:

```
<A href=title.html>..</A>
<A href=title.html#index>..</A>
<A href=http://...>..</A>
<A href=mycgi?a=no&b=yes>..</A>
```

Links in HTML

The <LINK> tag: A largely unsupported feature.

- An element in the <HEAD> section.
- Defines relationship with other documents.
- Gives the ability to organize documents in a hierarchical or linear way.
- Browser can use information directly.

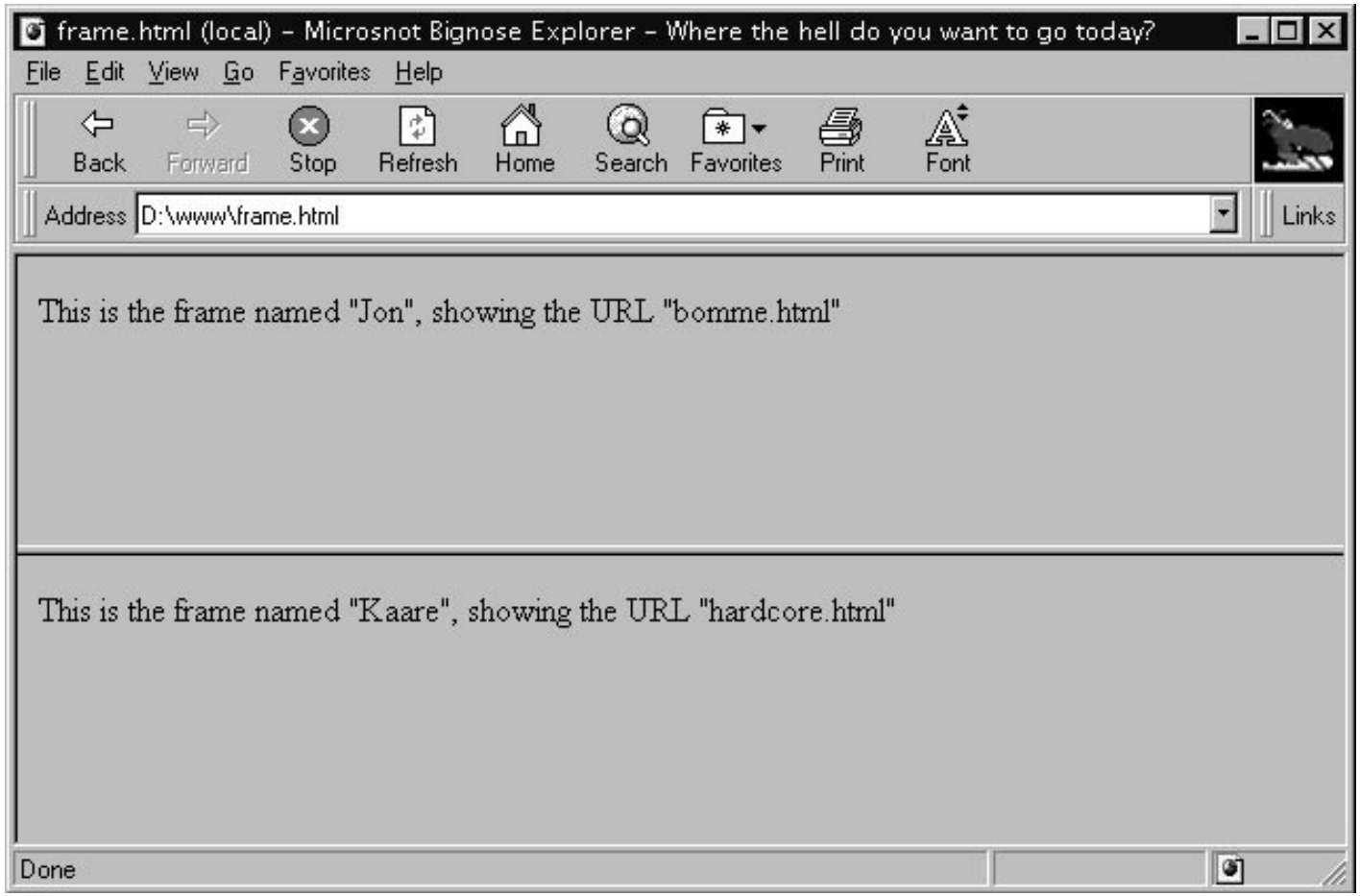
Frames

- Is an HTML 2.0 extension.
- The FRAMESET tag mutually excludes BODY
- A frame can load its own URL.
- Can be given a name, allowing it to be targeted by other URLs.
- Frames can be nested, allowing frames in frames in … Composition!

Frames example

```
<HTML>
  <FRAMESET ROWS="50%,50%">
    <FRAME SRC="bomme.html" NAME="Jon">
    <FRAME SRC="hardcore.html" NAME="Kaare">
    <NOFRAMES>
      Browser not HTML 2.0 compliant!
    </NOFRAMES>
  </FRAMESET>
</HTML>
```

Frames example



Forms

- Interactive mechanism to allow user-input.
- Defines standard input-components: Checkboxes, radiobuttons, inputfields …
- Part of the BODY element.
- Server-side interaction using the CGI protocol.

Forms example

```
<HTML>
  <BODY>
    <H1>Jump to a Search Engine Site</H1>
      <FORM METHOD="POST" ACTION="/cgi-bin/jumpbox.cgi">
      <SELECT NAME="goto" >
        <OPTION VALUE="Lycos">Lycos
        <OPTION VALUE="Alta Vista">Alta Vista
      </SELECT>
      <INPUT TYPE="submit" Name="submit_button" Value="Go To">
    </FORM>
  </BODY>
</HTML>
```

Forms example

CGI

- A Common Gateway Interface :-)
- Allows dynamic generation of URLs and makes computation possible.
- CGI-scripts execute on servers and thus allow interfacing with existing software.
- When clients invoke scripts, they can pass parameters.

CGI - I/O

- Post - pass parameters, using STDIN.
- Get - pass parameters in environment variables.
- Additional information is passed via environment variables.
- New URL is created on STDOUT.

CGI example

```
#!/usr/local/bin/perl
sub GET_FORM_DATA  {
   read(STDIN, $save_string, $ENV{CONTENT_LENGTH});
   @prompts = split(/&/,$save_string);

   foreach (@prompts) {
                ($tmp1, $tmp2) = split(/=/,$_);
                $tmp2 =~ s/\x2b/\x20/g;
                $tmp2 =~ s/%2C/\x2c/g;
                $tmp2 =~ s/%28/\x28/g;
                $tmp2 =~ s/%29/\x29/g;
                $fields{$tmp1}=$tmp2;
   }
   $URL_description = $fields{'goto'};
}
```

```
sub REDIRECT_THEM  {
   if ( $URL_description eq "Lycos" )
   { print "Location: http://www.lycos.com/\n\n";  }

   if ( $URL_description eq "Alta Vista" )
   { print "Location: http://altavista.digital.com/\n\n";  }
}

# -- Main ----------------------------------------------
&GET_FORM_DATA;
&REDIRECT_THEM;
```

Search Engines

- Basically a CGI-script that queries a database.
- Some uses “webcrawlers” for maintaining and updating the database.
- Webcrawlers roam the Web, inserting URLs into the database.
- Engines cannot cover the total domain.

- Scooter indexes 3 million pages per day.
- The indexer crunches 1GB of text per hour.
- AltaVista handles 12 million requests daily.
- The database holds 30 million unique pages.
- The indexer has 10 processors, 6GB RAM and 210GB RAID disk.
- Most powerful machine Digital ever made.