DOCUMENTS ON WEB

Learning Units

11.1 The internet and the world wide web

11.2 Documents and the world wide web

Learning Goals

- The basic technology used to build the internet
- How the world wide web uses the internet
- How documents are specified using HTML
- The distinction between presentation and structure of documents
- How documents are specified in XML

Motivation

- We examined in the last module how documents are formatted and printed using a computer.
- Documents in that context meant results computed by a computer
- In a more general context one should look at also documents which are to be disseminated via the world wide web.
- Besides dissemination one should also consider possibility of reading values from documents stored in remote computers and processing them for various purposes
- The need to exchange documents electronically and processing them have gained importance since the emergence of e-commerce

Motivation (Contd)

- To understand the need to distribute documents electronically we should first understand how computers are connected together and communicate in an orderly fashion among themselves
- Thus we will first examine very briefly the internet and the world wide web which uses the internet infrastructure

Computer Networks

- Now-a-days no computer has an isolated existence
- Computers in an organization are interconnected by local area networks (LAN)
- Home computers are connected to Public Switched Telephone Network (PSTN) which provide a connection to an Internet Service Provider (ISP)
- LANs of organizations connected to LANs of other organizations via PSTN using routers

Logical Network-internet

•Internet is the network of networks and interconnects millions of computers all over the world

•Internet is used to exchange electronic mail, exchange files and log into remote computers

•Common set of rules used by computers connected to the internet to communicate - called Internet Protocol (IP)

•Each computer connected to the internet has a unique address called IP address

•IP address is 4 bytes long

•IP addresses are a scarce resource

Internet-continued

• IP address converted to strings of characters which are easy to remember

• Group of characters combined as domains

• In the address rajaram@serc.iisc.ernet.in

in – Top most domain –country code

ernet – Internet Service Provider (ISP) in country

serc –Department within organization-name assigned by organization

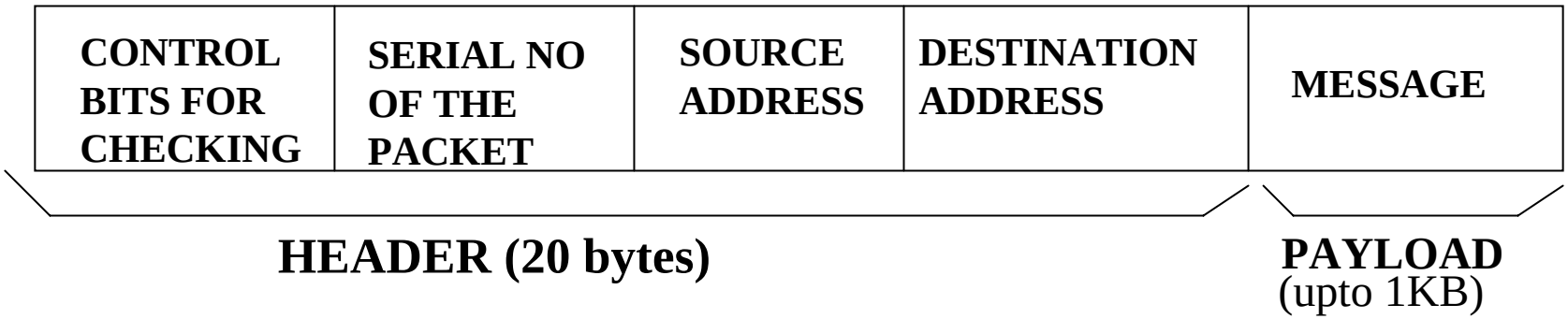
rajaram – Name of the user in department – assigned by department

• Hierarchy of addressing facilitates expansion

Internet -Continued

•Internet breaks up messages sent from source to destination into a number of packets

•Packet structure :



•Packets need not be of fixed length. Maximum length of a packet is 1 KB

•Message packetised to allow different packets to go along different paths - called packet switching

Packet Switching – Advantages And Disadvantages

- Each packet can pick free (cheapest) path to take
- Finally packets reassembled using serial no.
- Packet switching less expensive and adaptive as faulty paths can be avoided
- Major disadvantage of packet switching is the difficulty in predicting time taken by different packets to travel from source to destination

Packet Switching – Advantages And Disadvantages

- Variable packet delivery time does not matter for e-mail and text files
- Speed unpredictability however reduces effectiveness of audio and video traffic
- Major advantage – diverse machines and LAN's may be interconnected if they use common protocol called TCP/IP

Intranet And Extranet

- A network of computers within an organization using TCP/IP protocol an use all internet facilities such as e-mail, file transfer, remote login etc –called an intranet or corporate intranet
- Two corporate intranets may be interconnected using a leased line from PSTN – such a network is called an extranet
- Extranet between cooperating organizations can provide internet services such as e-mail,file transfer among them

World Wide Web Services

• World Wide Web (WWW) is a world wide multimedia information service available on the internet

• www contains web pages – created using a language called HTML (Hyper Text Markup Language)

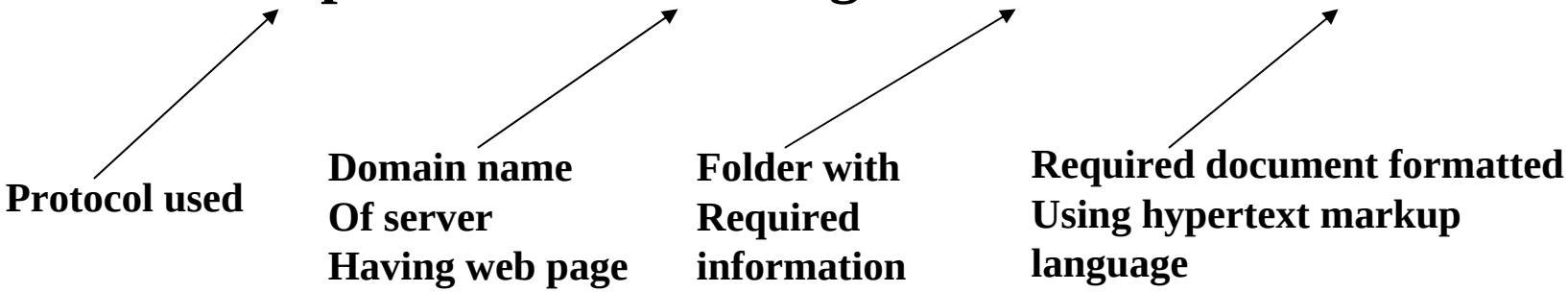
• HTML has features to embed links within web pages to point to other pages – we can thus navigate through links and search for required information

World Wide Web

- Web page located using a scheme known as Uniform Resource Locator(URL)

Example of URL

http://www.freesoft.org/connected/index.html



- Web browsers is a program on one's PC used to search for required information

Search Engines

- Browsers use search engines - that is programs which will navigate web pages using links
- Navigation based on search terms given by user
- All organizations now maintain web pages t o establish their "web presence"
- Web presence important to publicize organization for primarily advertising their services

What Is A Document?

A document has three parts

1. **CONTENT**: The string of characters normally coded in ASCII or UNICODE
   - A document now-a-days also includes, besides text, pictures, audio and video-all bit strings when digitized.
   - We will however be primarily concerned with textual data in this module.

2. **PRESENTATION:** How the data looks like to a human user- presentation may be on a video screen or on paper.

3. **STRUCTURE:** Helps interpretation of data by a computer Information such as type of data (Numeric, Alphanumeric) and its nature, for example an invoice, a purchase order, a recipe etc.

How Are Documents Processed By Computers

- Text processors add special annotations primarily to help format the resulting print outs

  Examples are: Paragraphing, Font selection, Placing titles, pagination,Tabulation etc.

  Examples: WORD, TEX etc.

- These are primarily presentation aids which take raw content and transform them to neat looking documents when displayed on VDU screen or printed on paper.

- They have no idea of the type of document and what they mean.

Text Processing By Computers

- Word processors primarily used for applications such as
  - -Preparation of manuals
  - -Preparation of catalogues
  - -Routine office correspondence
  - -Desk top publishing

- Report Generators (Discussed in Module10) are special variety which use a special language to generate and format reports.

- These are primarily for linear texts and not meant for linked text known as hypertext

Documents On World Wide Web

- In the world wide web documents located in many computers are linked
- Each document called a web page. Each web page has a unique path to retrieve it.
- Documents to be used in web pages need special annotations or markups mainly for formatting and for linking them to other documents in the web
- These annotations are called MARKUPS.

Documents On World Wide Web

- As documents on the world wide web are linked to many documents they are called hypertext.
- The markup used to link documents called Hyperlink.
- Web pages are retrieved from the host computers where they are located by a program called web browser running on a client.
- Clients use a communication protocol called **H**yper **T**ext **T**ransfer **P**rotocol (HTTP) to retrieve web pages
- HTTP recognizes a language called **H**yper **T**ext **M**arkup **L**anguage (HTML)

Hyper Text Markup Language

- An HTML Document has the following general layout

```
<HTML Version =“4.0”>        {Version optional}
    <HEAD>
            <!….The headings and their tags are placed here….>
    </HEAD>
<BODY>
            <!….Elements such as text with formatting
                    tags,links,tables,images etc go here……>
</BODY>
</HTML>
```

(! Is symbol used for comments)

Example Of An Html Documents

```
<HTML>
<HEAD>
<TITLE> Description of a book on Information Technology  </TITLE>
</HEAD>
<BODY>
<H1> Introduction o Information technology </H1>
<H2> A first level book in I.T </H2>
<P> Publisher : <I> Prentice – Hall of India </I></P>
<P> Year of publication: <B> 2003 </B></P>
</BODY>
</HTML>
```

Display Of Html Document

When the document is viewed using a browser it will appear as shown below

Introduction to Information technology

A first level book in I.T

Publisher : *Prentice – Hall of India*
Year of publication: **2003**

Explanation Of Tags In Html

- <HTML> tells it is an HTML document
- End of HTML documents is indicated by </HTML>
- <HTML version = “4.0” > version optional
- <TITLE> used to identify the document in the browsers title bar and is stored as the bookmark of this document
- <H1>,<H2> indicate headings. <H1> to <H6> available H1 highest size bold face and H6 lowest
- <P> indicates paragraphing
- <I> Italics and <B> bold face font
- <Observe all tags in this example here has end tag indicated by </tag>
- Stand alone tags are also there in HTML.

Linking Documents

- HTML can link to documents in other files.For Example to link an image we use :

  <IMG src = “mypicture.gif”>

- IMG indicates image and src the source (Observe the tag IMG is standalone and does not have end tag </IMG>

- HTML has feature to list items with serial number or bullets

- HTML can also display tables and forms

- HTML is as rich as some word processors.

Hyperlinking Html Documents

- HTML allows a web page to refer to other web pages

- When a reference link in the page is clicked the browser switches to the referenced site.

- The specification is
  <A href = " http:// www.iisc.ernet.in"></A>
  where A is called **anchor tag.**

- Linking can also be to other files

.

- Automatic conversion of word documents to HTML is possible using a tool

Shortcomings Of Html

- HTML is the earliest markup language which made it possible to retrieve documents stored in the world wide web

- HTML is primarily to facilitate presentation of contents of a web page.

- HTML does not have any means of specifying what the documents represents. Is it an invoice? A purchase order, book description etc.

- It also has no means of specifying the type of data to allow manipulation of data by browser.

- We thus need a markup language which is richer and is more descriptive of structure of a document and what it represents

EXtensible Markup Language

- A document has CONTENT, it has a STRUCTURE and it needs to be PRESENTED for ease of reading

- Word processors and HTML emphasize presentation of content and have no means of specifying structure (or what the data actually represents)

- XML is a new markup language which is capable of specifying what a document really represents

- XML is a proper subset of an international standard known as STANDARD GENERALISED MARKUP LANGUAGE (SGML).It is open standard and not proprietary

Parts Of XML System

- XML defines the structure of a document

- Unlike HTML it has tags which are user defined.This allows easy understanding of the nature of the document and assists in its processing.

- Formatting and presentation are not part of XML unlike HTML which has tags for bold face,italics etc.This is delegated to a companion language called XSL (Extensible Style Language)

- Linking documents to create hypertext is also not integrated in XML unlike HTML where tag <A> is a general purpose linking tag. Much more powerful linking is enabled by separating it to a companion language called XLL (Extensible Link Language).

Example Of XML Document

• A purchase order is represented in XML as below

```
< purchase_order >
            < order_no > B55567 </order_no>
            < date>
                        < year > 2004 </year>
                        < month > 10 </month>
                        < day > 9 </day>
            </date>
            <purchaser>
                        <name> ABC Traders </name>
                        <address>
                                    <street> 201 MG Road </street>
                                    <city> Bangalore </city>
                                    <pin_code> 560001 </pin_code>
                        </address>
            </purchaser>
```

Example Of XML Document (Contd)

```
    <item>
            <item_name> C Programming </item_name>
            <item_code> ISBN 81-203-0859-X </item_code>
            <quantity> 50 </quantity>
    </item>
    <supplier>
            <name> P-H India </name>
    </supplier>
</purchase_order>
```

Example Of XML Document

- Observe that the tags used have a syntax similar to HTML. The tags are, however, meaningful to a human reader

- The XML definition clearly brings out the structure of an invoice.

- However to interpret such a document and process it by a computer a companion document called Document Type Definition (DTD) is needed.

- DTD has its own syntax . We give DTD for this XML document in the next transparency.

Document Type Definition (DTD)

- **DTD of XML document of 11.2.15 is given below**

DTD Statements

```
<! ELEMENT purchase order (entry +) >
<! ELEMENT order_no (#PC DATA) >
<! ELEMENT date (year, month ,day)
<! ELEMENT year (#PC DATA)>
<! ELEMENT month(#PC DATA)>
<! ELEMENT day (#PC DATA)>
<! ELEMENT purchaser (name,address)>
<! ELEMENT name (#PC DATA)>
<! ELEMENT address (street,city,pin-code)>
<! ELEMENT street (#PC DATA)>
<! ELEMENT city (#PC DATA)>
<! ELEMENT pin-code (#PC DATA)>
<! ELEMENT item (item_name,item_code,quantity)>
<! ELEMENT item_name (#PC DATA)>
<! ELEMENT item_code (#PC DATA)>
<! ELEMENT quantity (#PC DATA)>
<! ELEMENT supplier (name)>
<! ELEMENT name(#PC DATA)>
```

Explanation Of Document Type Definition

- Each statement in DTD declares the elements of XML program
- <! ELEMENT purchase order (entry +) > states that purchase order is the top level element with one or more entry following it
- 2 statements are introduced at the start of XML definition which specifies the version of XML and the file name of DTD specification
- Assuming DTD is in a file purchase_order.dtd the declarations are
  <? XML Version =“1.0”>
  <! DOCTYPE purchase_order SYSTEM “purchase order.dtd”>
- The tags used in XML definition are then specified.

Explanation Of Document Type Definition

- <! ELEMENT order_no (#PC DATA) specifies that the tag order_no is a string of characters.

- <! ELEMENT date (year,month,day) specifies that the tag date is a higher level tag which consists of three tags- year, month and day.

- The description of each of the next level tags follow,
for example:
<! ELEMENT year (#PC DATA) declares year as a string of characters.

- The rest of DTD is self explanatory

Some Application Of XML

- XML's main use is in creating documents for the World Wide Web which can be retrieved by browsers at client computers.

- User defined tags give several advantages including use in

  - Push Technology – In this application time varying data specified by users e.g. Hourly stock prices of specified shares are automatically sent to the client's browser

  - Online banking – A standard XML format known as financial exchange initiative is used to obtain information such as bank statements.

Some Application Of XML

- Software and database updates

- XML adaptable to many natural languages such as Kannada as it uses Unicode standard.

- Use in Scientific Publications – Markup languages based on XML have been developed for chemistry – CML (Chemistry Markup Language) and MML (Mathematical Markup Language)