Part III: Survey of Internet technologies

- Content (e.g., HTML)
  - kinds of objects we're moving around?
- References (e.g, URLs)
  - how to talk about something not in hand?
- Protocols (e.g., HTTP)
  - how do things move around the net?

Part III.A: Content

- Labeling content: MIME, charsets
- Document formats: HTML, XML, XSL
- Image formats: GIF, JPEG, TIFF
- Multimedia: audio, video, music
- Interactive content

Why standards for content?

- Interoperability
  - Multiple implementations
- Preservation
  - Can you read Word 2.4 files?
- Global communication
  - Standards designed for consistency over features

Content Packaging: labeling data

- MIME:
  Multipurpose Internet Mail Exchange
  - Originally designed for mail
  - now used by other protocols
- Allows
  - Multiple media
  - Multiple character sets
  - Multiple languages

Internet Media Types ("MIME types")

- Standard way of naming data formats
- Hierarchical structure with parameters
- Applications use MIME to decide how to interpret data (instead of file extension)

If you can't get everyone to use the same file format, at least get them to say what format they used.

MIME Major Types

- `text`: sequences of characters
- `image`: bitmaps in various forms
- `audio`: sounds in various forms
- `video`: animations
- `message`, `multipart`: special purpose
- `application`: catch-all

MIME subtype

- Standard registry: “**`image/tiff`**”, “**`application/postscript`**”
- Registry rules: security, both standard & private (vnd)
- “**`application/vnd.ms-word`**”

Character sets: terminology

- Character: semantic (“A”, “capital alpha”)
- Code: number assigned to a character (63)
- Byte: 8-bit quantity
- Glyph: drawn shape

Simple map for A-Z 0-9

Complexities: accents, ligatures, Asian

More terminology

- Character set: assignment of characters and codes (ASCII, Unicode, JIS)
- Encoding: way of representing sequence of codes as bytes (UTF-8, UCS-2)
- Font: assignment of glyphs to codes combinations
- **`charset:`** character set + encoding

Charsets in the Internet

- Allow local (optimized) representation
- Labeling data with the charset used!
  - Instead of “user adjust browser”
- Support a minimum charset for interoperability (UTF-8)
- Other common values for “charset” include:
  - ISO-8859-1 (Western European)
  - Shift-JIS, EUC (Japanese)
  - Big5 (Chinese)

Internet Document formats

- HTML, SGML and XML
- Page layout: PDF
- proprietary application formats (word, wordperfect, etc.)

SGML and XML

- Standard Generalized Markup Language
- An ISO standard (ISO8879:1986)
- A way of writing
  (ways of writing documents)
- DTD (Document Type Definition)
  defines elements and rules about them
- XML (from W3C) is simplification

Markup: saying things about parts

- Semantic markup

  ```
  <part-no>N1025B</part-no>
  ```

- Structural markup

  ```
  <H1>N1025B</H1>
  ```

- Presentation markup

  ```
  <font face=aslan>N1025B</font>
  ```

HyperText Markup Language (HTML)

- An application of SGML (more or less)
- A way of writing text

  that includes links

  and (mainly) structural markup

  with some other things (like images) embedded.

HTML design goals

- *lingua franca* for the web
- Hypertext views of existing documents
- Simple, scaleable
- Platform independent
- Support for visually impaired
- Interoperability with common editors

HTML standards

- 1994: 2.0 (baseline) RFC 1866
- 1996: 3.2 (tables, forms, presentation)
- 1998: 4.0 (style sheets, lots more) W3C Recommendation

HTML/4.0

- More complete tables
- File Upload
- Internationalization
- Embedded objects
- Extensions
- Style sheets

XML: SGML simplified

- Primarily: simplify SGML
- Fix up ‘naming’
- Tools just now being deployed
- Being used inside protocols as general data representation

Style sheets

- Separate presentation information
  - **`<H1>`** should be bold, TimesRoman, 36 point
- Multiple styles for single document
  - print, display, handheld
- Developments
  - Cascading Style Sheets (designed for web)
  - Document Style Semantics and Specification Language (designed for SGML)
  - eXtensible Style Language (new development)

MHTML

- How to send HTML in email?
  - Include the images without changing URLs
- created new “multipart/related”
  - works for more than HTML
  - doesn’t require rewrite

“Active Content”

It’s a program! It’s a script! It’s a document format!

- Create documents that embed computation that control the document’s display
  - Pros and cons for this approach
  - *Postscript does this, PDF doesn’t*
- Dynamic HTML
  - Cascading Style Sheet… plus ...
  - JavaScript (ECMAScript)
  - control points for Document Object Model (DOM)
- Java applets as a document format

Page layout on the Web

- Postscript
  - Designed for printer control
  - **`application/postscript`**
- Portable Document Format (PDF)
  - Useful for screen presentation and printing with exact layout
  - **`application/pdf`**

Images on the Web

- **`gif`**: Graphics Interchange Format
  - – 8-bit color, transparent areas; patent cloud
- **`jpeg`**: Joint Photographic Expert Group
  - – lossy compression for photos, not line art
- **`tiff`**: Tagged Image File Format
  - – issues over tag standardization
- **`png`**: Portable Network Graphics
  - – calibration, hypertext links

Making content accessible Web Accessibility Initiative (WAI)

- guidelines for text labels as well as images
- avoiding audio tracks or providing subtitles
- using content negotiation
- cultural differences

More web content-types

- Desktop applications
  - Word, Excel, etc.
- 3-D renderings
  - VRML, etc
- Active content
  - Java
  - JavaScript, Document Object Model

Video formats on the Web

- MPEG
- QuickTime
- AVI

Audio and Music

- **`audio/basic`**
- Audio ‘files’ of limited use
- MIDI and music unevenly deployed
- Real time streaming media
  - combine protocol and format
  - create ‘codecs’ for processing

Summary: Content standards

- XML is most significant recent development
- Evolution along many fronts
- Market tension for proprietary extensions:
  - “free” viewer, pay for encoder
- Platform, ability, context, language, independence is major difficulty