MIME and the Web

draft-masinter-mime-web-info-01.txt

I-D by Larry Masinter

Presented by Alexey Melnikov

History

- Multipurpose Internet Mail Extensions (MIME) – RFC 2045, RFC 2046, …
- MIME was originally invented for email to extend Internet email messaging from ASCII-only plain text, to include other character sets, images, rich documents, etc.)
- Was later adopted for use in HTTP 1.0

Problems with application of Internet Media Types to the Web

- Lack of clarity about the purpose of MIME in the Web
- IANA registry of MIME types is out of date:
  - Lots of file types aren't registered
  - Those that are, the registration is incomplete or incorrect (people doing registration didn't understand 'magic number' or other fields).
  - The actual content deployed or created by deployed software doesn't match the registration

MIME rules weren’t quite followed on the Web

- HTTP server implementors and administrators didn't supply ways of easily associating the 'intended' file type label with the file, resulting in files frequently being delivered with an incorrect label
- Some popular servers had default configuration files that treated any unknown type as "text/plain" (plain ext in ASCII).
- Browser implementors are liberal in what they accepted, and use what looked like a file extension in the URL and/or magic number or other 'sniffing' techniques to decide file type, without assuming content-type label was authoritative.

Consequences

- servers sending responses to browsers don't have a good guarantee that the browser won't "sniff" the content and decide to do something other than treat it as it is labeled
- browsers receiving content don't have a good guarantee that the content isn't mis-labeled
- intermediaries (gateways, proxies, caches, and other pieces of the Web infrastructure) don't have a good way of telling what the conversation means.

Other issues (1 of 3)

- Differences between Web and Email
  - requirement for use of CRLF as line delimiter in plain text: in practice, web clients didn't restrict content to use CRLF in text/* MIME bodies.
  - Issues with charsets
    - default charset: HTTP specified ISO-8859-1 as the default character set for text/* body parts, not US-ASCII
    - Mislabeled charsets: misuse of iso-2022-jp or euc-jp to signal support for Microsoft extensions
    - Browsers are guessing charsets

Other issues (2 of 3)

- Evolution, Versioning, Forking of Internet Media Types: Internet Media Types do not identify a particular version of a file format, but a family of file formats
  - File formats need to include versioning information
  - Only backward incompatible changes require a new Internet Media Type registration
  - Backward incompatible changes are not always noticed by Media type reviewers, as previous registrations are incomplete/incorrect
  - Liberal processors are sometimes ignoring internal version information

Other issues (3 of 3)

- Fragment identifiers
  - The Web added the notion of being able to address part of a content and not the whole content by adding a 'fragment identifier' to the URL that addressed the data.
    - Originally used for HTML, but how would it apply to other content.
  - Internet Media Type registration template should include this information, but frequently doesn’t

Preliminary recommendations (1 of 2)

- Add fragment identifiers to the Internet Media Type registration template
- Recommend that “applications that use this type” field describes if a particular media type should be
  - for embedding (plug-in)
  - a separate document with auto-launch (MIME handler)
  - or always be downloaded
- Be clear in Security Considerations about scriptable content
- Signify which file extensions are useful for “sniffing”
- A web related draft needs to clarify that magic numbers and file extensions from the IANA registry can be used for sniffing

Preliminary recommendations (2 of 2)

- Update the Internet Media Type registry (fix/correct/add information)
- Relax IANA processes for MIME registries
  - “Internet Media Type registries are hard to update”, and there can be different definitions of the same MIME type.