

情報社会論

(252S0541)

講義概要

第7回目

メールや web での文字（日本語等）の取り扱い

符号化文字集合: 計算機で扱う文字を決めた集合。1 バイト系 (ASCII や 拡張 ASCII), 2 バイト系 (JIS X 0213) の他, 多バイト系として Unicode などがある。

文字符号化方式: 実際に符号化集合の文字を計算機で扱うための符号化法。エンコーディングとも呼ばれる。Shift-JIS, ISO-2022-JP, EUC の他, 各種の UTF などがある。

7 電子メールの仕組み (2)

7.1 Web メールを使ったメール処理

IMAP (Internet Message Access Protocol): サーバ上にあるメールフォルダを管理出来る機能を持つサーバとクライアント間の通信手順。現在は仕様は **IMAP4 rev.1 (Internet Mail Access Protocol version 4, revision 1)** [2003 年, RFC3501]^a が最新。POP よりも高機能の手順, 複数のクライアントで 1 つのメールサーバを利用するのに適している。

^a1996 年に RFC2060 として定められた。

[IMAP 利用の Web メール⁵⁰の構成要素]

- PC: インターネット用のブラウザがあればよい。
- Web メール用ホームページのある WWW サーバ
- IMAP 対応のメールアドレスがある受信用 SMTP サーバ
- 送信用 SMTP サーバ

www サーバ, 受信用 SMTP サーバ, 送信用 SMTP サーバは同一でもよい。

[Web メール処理の流れ]

- サーバ上のメール用のホームページにログオンし, ブラウザを利用してメールを作成。これをサーバに渡して送信。
- いくつかの SMTP サーバを経由して, メールは最終の宛先サーバに配信される。
- サーバ上のホームページでメールを読んだり, 管理する⁵¹。

[Web メールの特徴]

PC 側では, インターネット用のブラウザがあるだけで良く, 特別なメーラーをインストールする必要がない。また, メールを PC には取り込まないので, どの PC からでも利用できる。メールの機能は, メール用ホームページに任されるので, 個人が独自に機能を変更や追加は出来ない。個人の PC を持たない場合でも, メールアカウントが取得できれば公共の PC からでも利用できる。また旅行先等や出張先からでも利用できる。

7.2 符号化文字集合と文字符号化方式

電子メールは元々 7 ビットアスキー文字のみで構成されるが, それ以外の文字を利用する必要性から色々な工夫がなされた。

⁵⁰Web メールは IMAP を利用している場合が多い。

⁵¹Web メールソフトが, IMAP を使って, サーバ上のメールフォルダを管理する。

コンピュータで扱う文字集合内の各文字には一意の符号点 (一般的に非負整数列) を割り当てる。集合の要素である各文字に符号点が関連付けられた文字集合を**符号化文字集合**と言う。符号化文字集合には、例えば JIS X 0201, JIS X 0208, ISO/IEC 10646(UCS-2, UCS-4) 等がある。

符号化文字集合を定義した段階では、単に使用可能な文字と符号点が関連付けられただけの状態であり、コンピュータ上でどのようなバイト列として表現するかは定義されない。符号点からバイト列に変換する方法は、**文字符号化方式**という別の概念として定義される (Shift JIS や UTF-8 等)。ASCII 等、符号化文字集合と文字符号化方式の分離が明確ではないものもある。

※ウィキペディア「文字集合」より抜粋。

7.3 日本語の使用

符号化文字集合 (計算機で扱う文字を決めた集合)

- 1 バイト系 (ASCII や 拡張 ASCII) : アルファベット等
- 2 バイト系 (JIS X 0213 など) : 漢字など
- 多バイト系 (Unicode など) : 各国語対応

JIS X 0213 (新拡張 JIS コード)

2000 年に制定された日本語文字コード規格。^a

11,223 個の漢字・記号文字が 2 バイト 7 ビット表現, 2 バイト 8 ビット表現でそれぞれ定められている。

^a2004 年に第 2 次規格 **JIS X 0213:2004**, 2012 年に第 3 次規格 **JIS X 0213:2012** が出た。

文字符号化方式 (実際に符号化集合の文字を計算機で扱うための符号化法, エンコーディング) 文字符号化方式 (**Character Encoding Scheme**) とは, 符号化文字集合で文字に対応付けられた非負整数値を, 実際にコンピュータが利用できるデータ列 (バイト列) に変換する符号化方式を指す。文字符号化体系, 文字符号化スキーム (CCS) とも言う。以下の表にあるような, いくつかの組み合わせがある。例えば, 日本語の文字の場合は, 主に **JIS X 0208** を符号化文字集合として, その符号化法には, **ISO-2022-JP**, **EUC-JP**, **Shift-JIS** がある。

この JIS コードは, 1993 年 RFC1468 で電子メールで ISO-2022-JP の名称で使用されるコードとして規定された。

電子メールでの日本語コードは JIS コードが基本

[コンピュータの文字コードの例 (符号化文字集合と文字符号化方式の組合せ)]

文字集合 character set	符号化方式 character encoding scheme	符号化表現
ASCII	(なし)	ASCII
(JIS ⁵² で定めた日本語のセット) JIS X 0201 + JIS X 0208 JIS X 0213 ⁵³	Shift-JIS ISO-2022-JP EUC	Shift-JIS ⁵⁴ JIS EUC-JP ⁵⁵
KS X 1001 (ハングルと漢字のセット)	EUC	EUC-KR
GB 2312 (簡体字中国語のセット)	EUC	EUC-CN
Unicode ⁵⁶	UTF-8, UTF-8N	UTF-8
	UTF-16BE, UTF-16LE	UTF-16
	UTF-32BE, UTF-32LE	UTF-32

日本国内ではコンピュータによって色々な日本語コードが使われているが、メールを送信する際にほとんどの場合は、メーラーが JIS コードに自動変換してから送信している。受信側は受信したメール内容を、そのコンピュータで使用している文字コードに変換して読む。

7.4 MIME（多目的インターネットメール拡張）

メールは7ビットアスキー文字から構成される。これ以外の文字やコードメールで送信するためには、すべてのコードを7ビットアスキーコードに変換する必要がある。このための変換規約として、多目的インターネットメール拡張 MIME(Multipurpose Internet Mail Extensions) が定められた。現在のものは、1996年に定められたもの⁵⁷で、RFC 2045, RFC 2046, RFC 2047, RFC 4288, RFC 4289, RFC 2049 から構成される。MIME は、7ビットへの変換への規則だけでなく、その内容・形式について細かく規定し、メールで種々の処理が出来るようにしている。現在では、Web でも利用されるようになった。

通常、メーラーでは表示形式として、(1) プレーンテキスト形式、(2)HTML 形式 (HTML メール) の2種類が選べるようになっている。もちろん、スマートフォンでメールを見る時には、基本的に (Web メールとして) ブラウザを利用しているので、HTML 形式で表示していることになる。

- MIME ヘッダを構成するフィールド

フィールド	内容
MIME-Version	MIME のバージョン。現在は 1.0 のみ
Content-Type	内容の形式を表す。
Content-Transfer-Encoding	符号化の方法を表す。
Content-ID	MIME の実体 (entity) を示す ID
Content-Description	MIME の実体に関する記述。
Content-Disposition	MIME の実体の性質
Content-Language	MIME の実体の言語

- Content-Type の例 (タイプ/サブタイプ [; charset=文字コード 等のパラメータ])

Content-Type: text/plain; charset=iso-2022-jp

Content-Type: text/html;name="test.html"

Content-Type: image/jpeg;name="test2.jpg"

Content-Type: image/gif;name="test3.gif"

Content-Type: application/vnd.rn-realmedia;name="test4.rm"

Content-Type: audio/mpeg;name="test5.mp3"

Content-Type: MultiPart/Mixed;Boundary="--Next_Part(Sat_Jul_21_15:08:39_2001)--"

- Content-Transfer-Encoding の例

Content-Transfer-Encoding: 7bit

Content-Transfer-Encoding: base64

Content-Transfer-Encoding: quoted-printable

base64 は、すべてのファイルに対応する変換法、quoted-printable は、制御文字や8ビット文字を7ビットに変換する方法。7ビットではない西欧系の文書の変換に利用される。

⁵²JIS=Japanese Industrial Standards Committee の略称、日本産業標準調査会。

⁵³厳密には、JIS X 0213 は、JIS X 0208 の上位互換ではない。

⁵⁴主にパソコン (Windows や Mac) で使われる。

⁵⁵主に Unix で使われる。

⁵⁶万国文字集合 (Universal Character Set)。

⁵⁷URL <https://datatracker.ietf.org/doc/html/rfc2045>

7.5 日本語の取り扱い

- メールで使われる日本語コード
 - － 基本的には JIS コードを用いる⁵⁸。
 - － ヘッダでの日本語の取り扱いは厳密には規定されていないが、最近は MIME に従ったものが多い。
 - － Subject には日本語が使える⁵⁹が、その処理は MUA によって違う。
 - － JIS コードで表す方法。
Subject: ESC\$BF|K\81\$N%5V%8%'%/HESC(B
 - － MIME 的に表す方法。
Subject: =?ISO-2022-JP?B?GyRCRnxLXDhsJE41NSVWJTglJyUvJUgbKEI=?=
From には避けた方がよい?。
 - － メールの本文では普通に日本語が使える。半角カタカナは全角カタカナに自動変換するものが多い。

7.6 メールで出来ることとその危険性

MIME の機能を使うと、MUA に、メールに含まれる種々のファイルを解釈し、実行する機能を搭載することも出来る。これにより、メールでかなり多くのことが実現できる。

例えば、メールで添付ファイルとして種々なファイルを送ることが出来る。ファイルを適当な方法で7ビットアスキーコードに変換し、それを、添付ファイルの規則に従って、メールの文書に含める。受信側のメーラーは、それを解釈し、ファイルを元のものに変換する。ダブルクリックすることで、元のファイルに復元し、あたかも、ディスクに保存されている普通のファイルと似た処理が可能となる。

また、HTML メールとして、画像や音声を含んだ、ホームページ風のメールを送ることも出来る。メールにホームページ用文書と同様なものを添付し、メーラーがホームページ文書の解釈と同様な解釈を行い、実行する。関連する画像ファイル等のマルチメディアファイルも、添付ファイルと同様にアスキーコードに変換したものをメールに含める。

これらの便利さと引き換えに、添付ファイルや HTML メールの危険性が常に付きまとう。つまり、それらを自動実行する設定になっているとウィルスが入り込んだりインターネットに自動的に接続したり情報が漏れたりするのである。

そのため、Windows Update などのセキュリティ更新を定期的実施することやアンチウィルスソフトをインストールし、最新版へ更新することは重要である。また、PCなどでメーラーを使ってメールを表示させる場合には、表示形式として、「プレーンテキスト形式」にしておくことで、添付ファイルや HTML メールの危険性を軽減させることができる。

7.7 多要素認証

多要素認証とは、ID・パスワードなどの「**知識情報**」に加えて、スマートフォンやICカードなどの「**所有情報／所持情報**」や指紋、静脈、虹彩、顔などの「**生体情報**」という認証の3要素の中から2つ以上の異なる認証要素を用いて認証する方法をいう。従来のメールクライアントのPOP・IMAP・SMTP方式での基本認証は「知識情報」に基づくものなので、今後は多要素認証に対応したメーラー⁶⁰に乗り換えたり、Webメールの設定を多要素認証を利用した2段階認証にする必要がある。

⁵⁸MIME で指定した場合、他のコードも使うことは出来る。

⁵⁹半角カタカナは使用禁止だが、自動的に全角カタカナに変換されるものが多い。

⁶⁰2022年10月1日よりExchange OnlineのOutlook, EWS, RPS, POP, IMAP, EASプロトコルの基本認証の無効化が開始されることが2021年9月に発表された。