ディジタルドキュメント (7)

高久雅生

2014年5月29日（木）3・4時限

本日のお品書き

- （第3回レポート課題の回収）
- （前回の復習）
- （前回講義での質問・要望への回答：7件）
  - レポート評価、出版、HTMLとXML
- テキスト処理とドキュメント形式
  - ドキュメント形式（フォーマット）とは
  - テキストフォーマットとは
    - 文字コード
  - 電子メールのドキュメント形式
  - 軽量テキストフォーマット

（前回の復習 = ふりかえり）

- WWW (World Wide Web)
  - 歴史
  - ハイパーテキスト、ハイパーメディア
- HTML
  - 歴史と標準化
  - 要素、タグ、属性
  - ハイパーリンク
  - 構造と見栄え
- XML
  - メタ言語
  - 整形式

質問･要望へのお返事 (1-1)

- 授業の最初のほうに話されていたレポート課題についてだが、やはり納得がいかない部分が多かった。「大学生としての最低限のレポート」というのはもっともであるし、第1回レポートの「書誌事項をしっかり」というのはそれにのっとってまだ納得できるものであったが、第2回レポートの文章に関してはそもそものレポート課題の内容がデジタルドキュメントとしての論文を見て、提供形態、形式、出版プラットフォームを調べてくるというもので、論理立てるも何もないと感じてしまった。すべて自分の課題の意図を読み取る力の無さや、普段の仕様を満たすだけのプログラム系レポートに慣れすぎているせいだとは思うが、とにかく、第2回レポートに関しては不満しか残らなかった。不満しか書いていなくて申し訳ないのですが、第2回レポートについては「何を書いていいのか分からない」と提出の前から、自分の周りの人は皆、話していたので、要件関係なく、出し方も悪かったと思います。

質問･要望へのお返事 (1-2)

出題形式に分かりづらい点があるかとは思いますが、レポートの出題意図が十分に明らかでないと感じた場合、教員に確認、質問してください。
また、返却後の採点に疑義がある場合も早めに教員に確認するようにしてください。

質問･要望へのお返事 (2)

- レポートを提出期限以降に提出したものは、採点しない、もしくは、最低評価に準ずるものにしてもらわないと、期限の意味が無く、解説を聞いてから作成したレポートの方が評価が高くなることがあり得るのはいかがなものでしょうか。

原則として、遅れ提出を受けたものは、他のものと同様に採点して返却しますが、遅れの程度に応じ、大幅に素点を割り引いています。したがって、遅れ提出の方が評価が高くなることはありえません。

質問･要望へのお返事 (3)

- あと何回レポートが課されるかが気になります。

あと1～2回の予定です。

質問･要望へのお返事 (4)

- 講義とは直接関係無いが、私費出版の定義について、詳しく知りたいと思う。印刷所に原稿を持ち込んで印刷してもらうのが私費出版ならば、同人誌等も含まれるのだろうか？ いわゆる二次創作は難しいかもしれないが、一次創作等、本人に著作権があるようなものは納本されるのだろうか。

同人誌･二次創作であろうが、納本義務に関しては無関係のはずです。いわゆる出版社側が責任を持って納本することになっていますが、罰則規定がないため、商業出版社であっても納本率は100%ではないと問題にはなっていますが。自身で作成した出版物については寄贈もご検討を。

Ref: 木川田朱美, 辻慶太.国立国会図書館におけるポルノグラフィの納本状況. 図書館界, 2009, Vol.61, No.4, pp.234-244. http://ci.nii.ac.jp/naid/110007819783

質問・要望へのお返事 (5)

- 太字にするなど、同じ目的のタグが複数あるのはなぜか?

歴史的な経緯と後方互換性のために残っているものです。当初はブラウザ上の表示のための物理的スタイル指定（b, i要素など）のみだったのですが、テキスト読み上げ機能や端末制限を考慮した論理的スタイル指定（strong, em要素など）に一般化された経緯があります。

質問･要望へのお返事 (6)

- スライド45において、XMLは「Webでの経験を元に…」の説明の時、この経験での問題点を口で言っていたと思いますが、イメージが出来なかったので、具体的なものは何か?

主として以下の2点が大きかったようです :

- 文字コード等の国際化: 文字コード指定の枠組み
- 整形式: 終了タグの省略を許さないなど

- 文字コード指定の枠組み
  - XMLの場合：ファイル冒頭での文字コード指定

    <?xml version=“1.0” encoding=“utf-8”?>
- 整形式: 終了タグの省略を許さないなど
  - HTMLの場合：

    第1段落…

    第2段落…
  - XMLの場合：

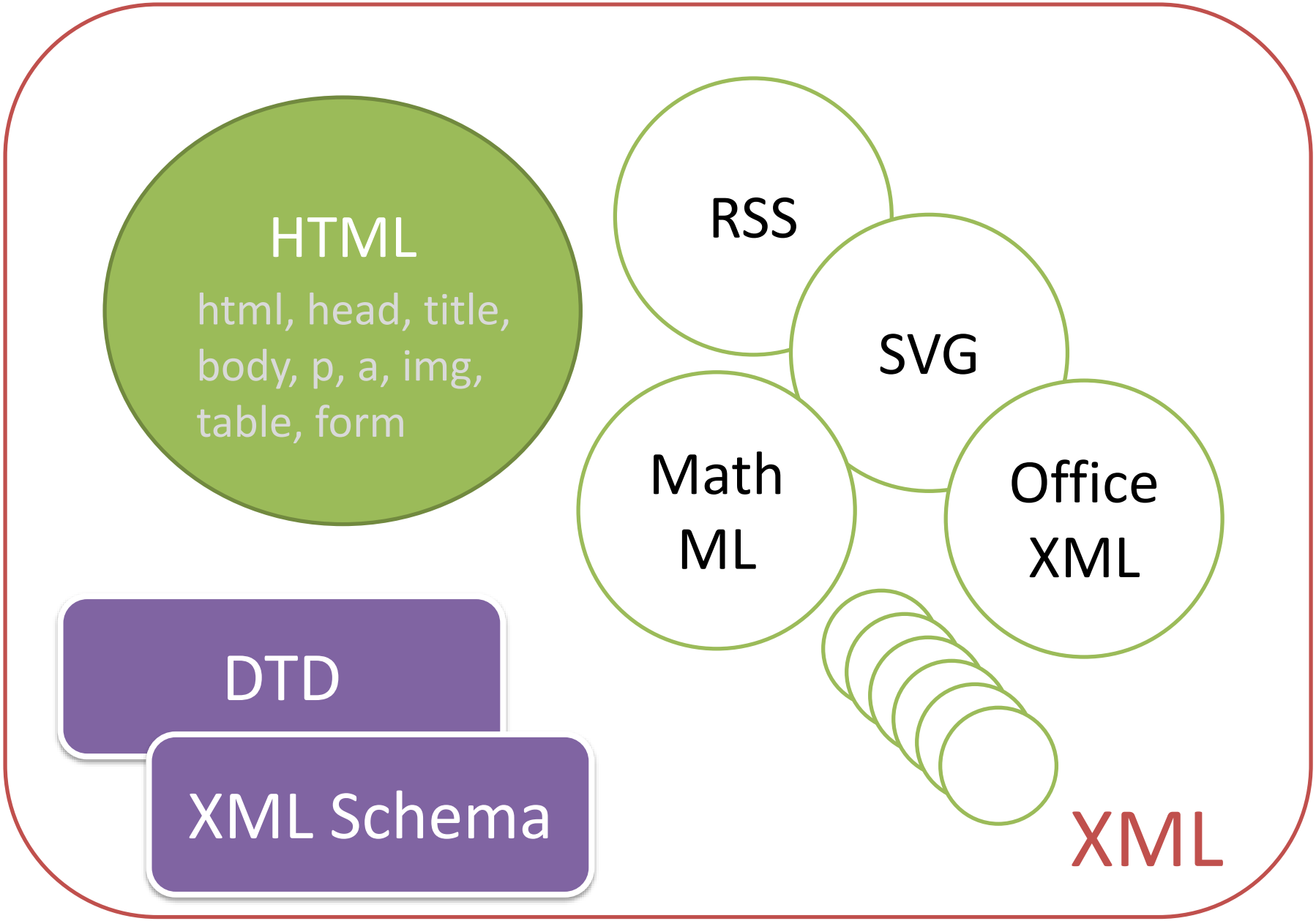
    第1段落…

    第2段落…

質問・要望へのお返事 (7)

- HTMLとXMLは同じものと考えてよいのでしょうか?

No.
HTMLというのが一定の規定からなる有限のタグ・要素集合から構成されるのに対し、XMLはどのようなタグ・要素にも許容しうるという大きな違いがあります。言い換えると、「XML = メタ言語 vs HTML = インスタンス言語」という立場の違いが厳然として存在します。


HTML
html, head, title, body, p, a, img, table, form
RSS
SVG
Math ML
Office XML
DTD
XML Schema
XML

ドキュメントフォーマット (1)

ドキュメントフォーマットとは？

- デジタル文書の内容を機械的に解釈するための「決まりごと」「様式」「形式」
  - データをどのように並べるか（配置）
  - 内容解釈のためのデータ・情報は、デジタル情報として、符号化（エンコード）される
- 例：
  - プレインテキスト？
    - 文書内容を文字コードにより符号化し、テキスト（文字の連なり）として解釈できるようにしたもの。
  - 画像形式？
    - 表現すべき画像要素を、ピクセル・描画要素単位の情報として符号化し、2次元画像として解釈できるようにしたもの。

ドキュメントフォーマットの切り口 (1)

- テキスト (text) vs バイナリー (binary)
  - ビットデータ
  - 文字コードによる解釈
  - 外字
- フォーマットの指定・識別・判別
- シンプルコンテンツ vs 複合メディア
  - 埋め込みコンテンツ
  - ハイパーリンク
- メタデータ
  - 埋め込みメタデータ
  - 外部メタデータ記述
- 文書レイアウト
  - ページ概念
- 文書内の書式要素
  - 見栄え / スタイル
  - フォント
- ファイル vs ストリーム
  - データの保存・蓄積と配信

ドキュメントフォーマットの切り口 (2)

- オープンフォーマット
  - 移植可能性 / ソフトウェア独立性
  - Free / proprietary
- 標準化
  - デファクト標準とデジュール標準 (“de facto” vs “de jure”)
- 文書フォーマットのバージョン
- フォーマット変換
  - プレインテキスト → HTML
  - LaTeX → PDF
- 圧縮
  - 可逆 vs 非可逆
- セキュリティ
  - パスワード
  - 電子署名
- 長期保存
- デジタルフォレンジック

プレインテキストとは？

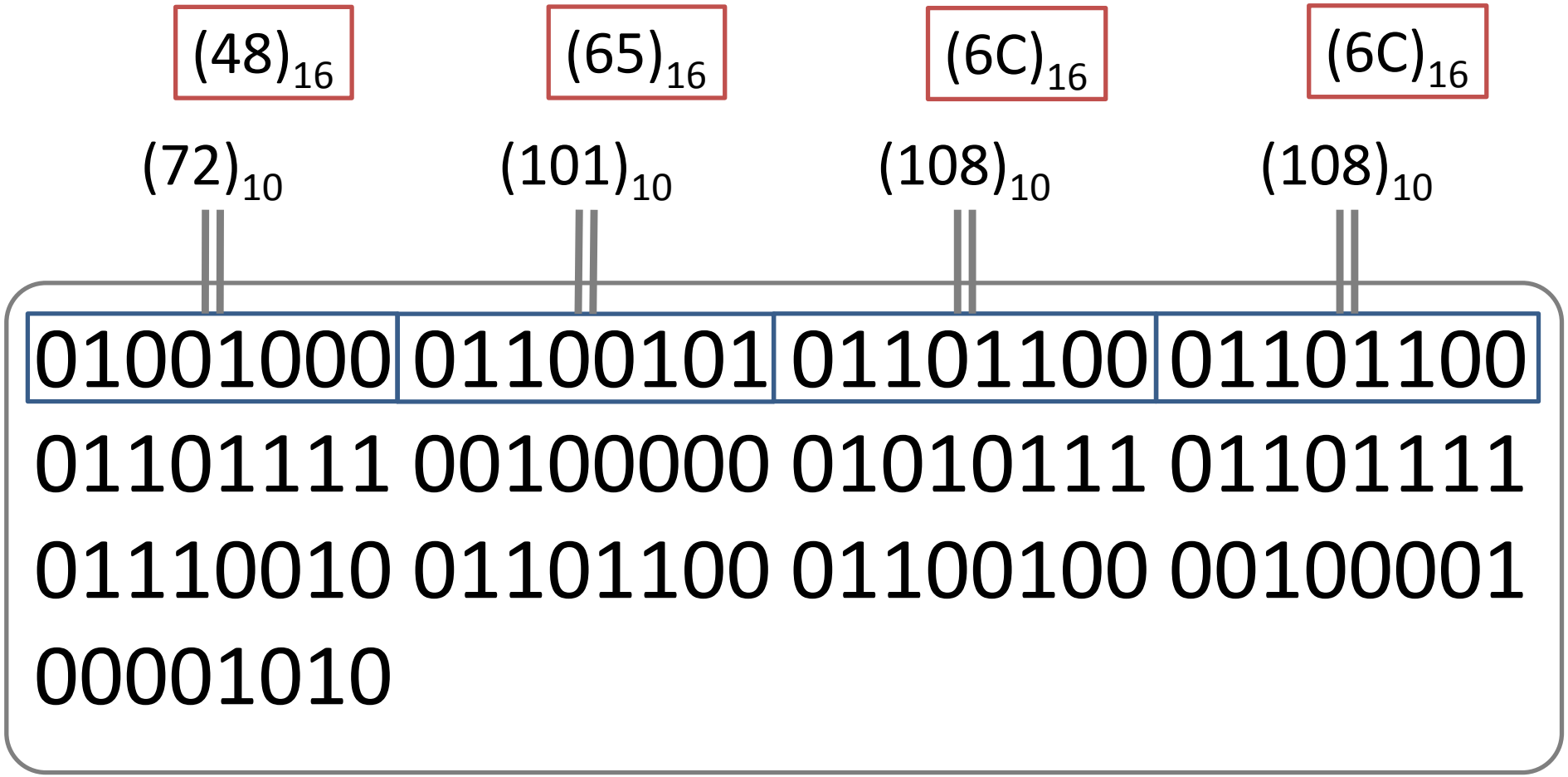
- 文書内容を文字コードにより符号化し、テキスト（文字の連なり）として解釈できるようにしたもの。
- もっとも基本的なドキュメントフォーマットの一つ
  - ドキュメントフォーマットの基礎
  - （別の定義：バイナリフォーマットではない）
- 特徴：
  - ほぼあらゆる環境で特別なソフトウェア無しに用いることが出来る
  - テキストフォーマットだけでは、書式要素を保持しない
  - 複合的なオブジェクトとの関連付けはそれ単体ではできない

プレインテキスト読解例 (1)

01001000011001010110110001101
10001101111001000000101011101
10111101110010011011000110010
00010000100001010

プレインテキスト読解例 (2)

- 先頭からバイト(byte)単位で解読してみよう。
  - 1バイト = 8ビット (bit)



プレインテキスト読解例 (3)

- 文字コードと照合
- ASCIIコード表を基にしてみる

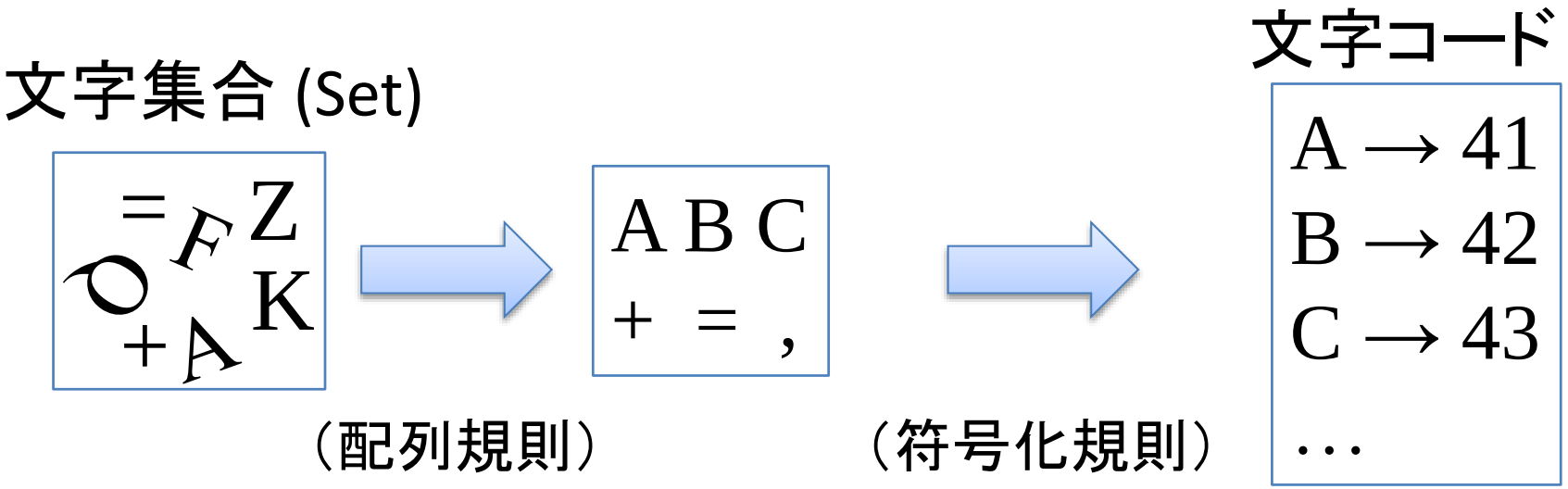
H e l l o　W o r l d ! ¥n

48 65 6C 6C 6F 20 57 6F 72 6C 21 64 0A

01001000 01100101 01101100 01101100
01101111 00100000 01010111 01101111
01110010 01101100 01100100 00100001
00001010

※文字コードとは？

- 文字を表現する集合を配列し、それを符号化する方式を定めたもの
  - 文字集合 (Character Set)
  - 文字符号化方式 (Character Encoding)



ASCIIコード表

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	¥	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	M	n	o
7	P	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

文字コードの実際

- 欧米語の場合
  - ASCII
  - ISO-8859-1
- 日本語の場合
  - JIS X 0208
  - JIS X 0201
  - …
- その他
  - Unicode

- よく使われる文字コード
  - ASCII
  - UTF-8
  - Shift_JIS
  - EUC-JP
  - ISO-2022-JP
  - ISO-8859-1

文字コードにおける制御文字

- 計算機上で表現される「文字」には、情報交換の役割があり、たとえば、下記のような特殊な役割をもつ制御文字が含まれている:
  - 空白文字
  - タブ
  - 改行文字
  - 削除記号（ BS, DEL ）
- これらの文字以外の表現は、歴史的な文脈で必要とされたという経緯に加え、テキストのみで文書構造を表現するためにも用いられる。

0010110100101101001000000000101010111001111000101011010111010110
0100000101100101110110111000000101110000010000000101000010011010 1
1000010111001101100001011011110010000001010100011000010110101101 1
0000101101011011101010010100100100000000100000001000000010000010 10
0010101010000011001000110000001100010011001111000111101011110011 0
1001011011111101110101001001100101111000011110111101100011111001 0
0011000010111001111011001111011000101001001100101110110000110111 0
0101001001110101010100100110111101010010010110111101001001011111 1
1010000110100011000010100010111100101111001000000010000001100001 11
1011110110001111100100011000010111001111011001111011000101111111 1
0111101011110111110001101101001101101110111110111100001100101011 1
1001110100101111000011010010111000111101001011010001110100101101 0
0010101110001010011010110101111001101011011111001111000010100010 1
1110010111100100000001000000100010101101101011000010110100101101 1
0000111010001000000110110101100001011100110110000101101111010000 0
0011100110110110001101001011100110010111001110100011100110111010 1
0110101101110101011000100110000100101110011000010110001100101110 0
1101010011100000000101000101111001011110010000000100000001010100 01
1001010110110000111010001000000001100000011001000111001001011010 01
1100000110101001110010010110100110001001100110011100100110100000 0
1010

「文字化け」とは?

- 文字コードを解釈するソフトウェアが元の文字コードを復号化できない場合に発生する不整合
  - 元の文字コード : Shift_JIS
  - ソフトウェア側 : UTF-8

プレインテキストに基づくドキュメントフォーマット

- プレインテキストに基づくドキュメントフォーマットには多くの事例がある
  - Eメール
  - HTMLファイル（ウェブ）
  - SGML / XML
  - LaTeX

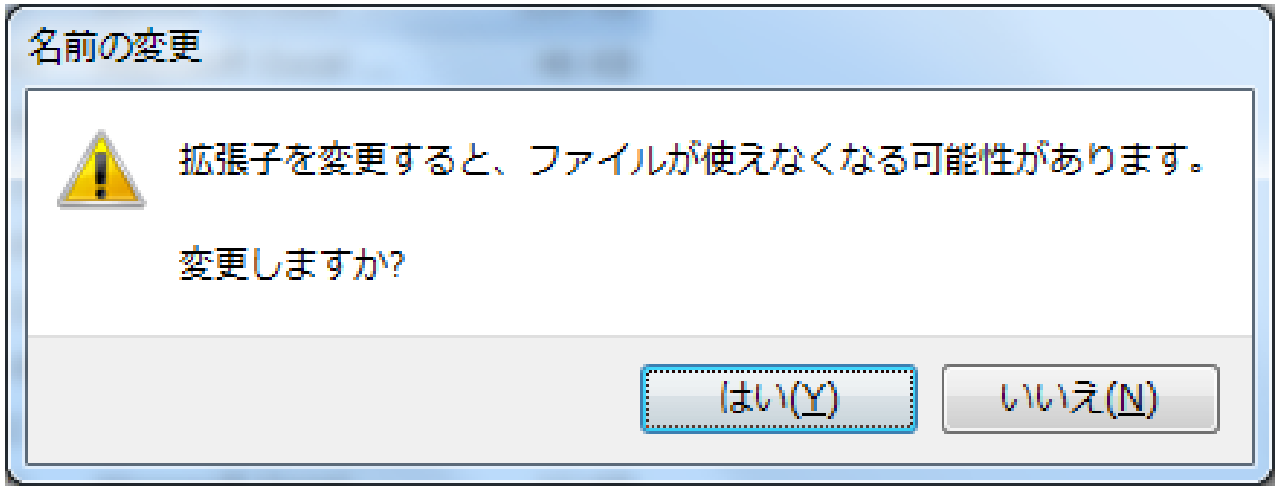
ドキュメントフォーマットの識別，判別

- メタデータとして付与
  - 拡張子
- 内容で判断
  - マジックナンバー（File magic）
- 通信・転送プロトコル上で指定
  - MIME Type (IANA)
  - その他

フォーマットの識別：ファイル拡張子

ファイル名の末尾を「.」（ピリオド）で区切り、ファイル形式を示す2～3文字からなる文字列を付与

- テキストファイル: text.txt
- HTMLファイル： index.html, welcome.htm
- PDFファイル： example.pdf
- などなど。



フォーマットの識別：ファイルマジックナンバー

- 古典的なヒューリスティック手法
- ファイルの冒頭に分かりやすいフォーマットを示すバイト列ヘッダを挿入しておく方式
- 例：
  - XML形式：<?xml version=“ …
  - GIF形式（画像）：GIF87a または GIF89a

フォーマットの識別：MIMEタイプ (MIME type)

- Multipurpose Internet Mail Extensions (MIME)
- コンテンツ用のフォーマット指定タイプ
  - Internet Media Type, Content Type
- ウェブやEメールでのドキュメント形式の指定に使われる
- IANAによる公式登録リスト：http://www.iana.org/assignments/media-types
- `[type]/[subtype]`
  - `text/plain, text/html, application/pdf, image/gif, vnd.openxmlformats-officedocument.presentationml.presentation`
- タイプ
  - application, audio, image, message, model, multipart, text, video

メールフォーマット

- Eメール：最も古典的なインターネットアプリケーションの一例
- テキスト形式によるフォーマットの例
  - 基本的に、テキストデータのみでやり取りする
    - （一種、テキスト形式への符号化がなされる）
- オープンフォーマット
  - 様々なソフトウェア（メーラー）がメール内容を扱う
  - 表示、作成、送信
- 送信者と受信者は異なる環境にいることを前提
  - 全世界で幅広く分散して使われており、新しい仕様に改善しようとしても一度に一斉に切り替えることは不可能
- RFC 5322 “Internet Message Format” (2008): http://tools.ietf.org/html/rfc5322
  - RFC 2822 (2001) ← RFC 822 (1982)

メールフォーマットの例

Date: Thu, 20 Sep 2001 07:00:00 JST
From: 首相官邸 <koizumi@mmz.kantei.go.jp>
Reply-To: koizumi@mmz.kantei.go.jp
Subject:【小泉内閣メールマガジン 2001/09/20】難局にひるまず立ち向かおう！！
To: koizumi-ml-user@mmz.kantei.go.jp

小泉内閣メールマガジン ========== 2001/09/20

★☆　今週のキーワード「補正予算」 ☆★
　小泉総理が平成１３年度補正予算の編成について準備を進めるよう指示をしました。（解説は最後に）

□　目次

［らいおんはーと　～　小泉総理のメッセージ］
●　難局にひるまず立ち向かおう！！

メールフォーマットの例

メールヘッダ

Date: Thu, 20 Sep 2001 07:00:00 JST ⏎
From: 首相官邸 <koizumi@mmz.kantei.go.jp> ⏎
Reply-To: koizumi@mmz.kantei.go.jp ⏎
Subject:【小泉内閣メールマガジン 2001/09/20】難局にひるまず立ち向かおう！！ ⏎
To: koizumi-ml-user@mmz.kantei.go.jp ⏎

メール本文

小泉内閣メールマガジン ========== 2001/09/20

★☆　今週のキーワード「補正予算」　☆★
　小泉総理が平成13年度補正予算の編成について準備を進めるよう指示をしました。（解説は最後に）

□　目次

［らいおんはーと　～　小泉総理のメッセージ］
●　難局にひるまず立ち向かおう！！

メールフォーマットの例（ヘッダ部）

```
Received: from goat.ipc.ulis.ac.jp (goat.ipc.ulis.ac.jp [133.51.184.12])
        by nile.ulis.ac.jp (8.9.3+Sun/8.9.1) with ESMTP id HAA06473
        for <masao@nile.ulis.ac.jp>; Thu, 20 Sep 2001 07:06:10 +0900 (JST)
Received: from mails1.mmz.kantei.go.jp (mails1.mmz.kantei.go.jp [202.212.115.161])
        by goat.ipc.ulis.ac.jp (8.8.6 (PHNE_14041)/3.7W) with SMTP id HAA02715
        for <masao@ulis.ac.jp>; Thu, 20 Sep 2001 07:08:52 +0900 (JST)
Received: (qmail 6408 invoked by uid 113); 20 Sep 2001 07:03:39 +0900
Received: from localhost (HELO mails1.mmz.kantei.go.jp) (sendmail-bs@127.0.0.1)
  by localhost with SMTP; 20 Sep 2001 07:03:39 +0900
Received: (qmail 6325 invoked by uid 200); 20 Sep 2001 07:00:00 +0900
Date: 20 Sep 2001 07:00:00 +0900
From: =?ISO-2022-JP?B?GyRCPHNBajQxRSEbKEI=?= <koizumi@mmz.kantei.go.jp>
Reply-To: koizumi@mmz.kantei.go.jp
Subject: =?iso-2022-jp?B?GyRCIVo+LkB0RmIzVSVhITwlayVeJSwlOCVzGyhKIDIwMDEv?=
        =?iso-2022-jp?B?MDkvMjAbJEIhW0ZxNkkkSyRSJGskXiQ6TikkQTh+JCskKiQmISobKEo=?=
        =?iso-2022-jp?B?GyRCISobKEo=?=
To: koizumi-ml-user@mmz.kantei.go.jp
Message-Id: <20010919220000.6323.qmail@msgs1.mmz.kantei.go.jp>
Mime-Version: 1.0
Content-Type: text/plain;
        charset="iso-2022-jp"
Precedence: bulk
Content-Length: 12860
```

メールサーバ転送の記録

時刻

送信者

返信先の指定

件名

宛先

メッセージID

MIMEバージョン

フォーマット指定

メールマガジン配信指定

本文長

ここまで、メールヘッダ

メールフォーマットの例（ヘッダ部）(2)

- Date
- From
- To
- Subject
- Message-Id
- Received
- Reply-To
- Content-Type
- MIME-Version
- Precedence
- Content-Length

複合型ドキュメントとしてのメールフォーマット

- 複数のドキュメントをテキストフォーマットの中に埋め込む
  - マルチパート：添付ファイルの実現
- MIME：Multipurpose Internet Mail Extensions
  - メール用の交換フォーマットにおいて、添付ファイルやプレインテキスト以外の形式でのやりとりを行うための拡張仕様
- メールヘッダの拡張
  - MIME-Version
  - Content-Type
  - Content-Disposition
  - Content-Transfer-Encoding

添付ファイル付きのメールフォーマット例

```
Received: from SINPRD0310MB391.apcprd03.prod.outlook.com ([169.254.5.109]) by
    SINPRD0310HT002.apcprd03.prod.outlook.com ([10.255.17.37]) with mapi id
    14.16.0311.000; Thu, 30 May 2013 02:38:55 +0000
Content-Type: multipart/mixed; boundary="_002_9BB0C5FDD26EB142B3B76E53820B0FB9020BC402SINPRD0310MB391_"
Content-Transfer-Encoding: binary
From: Masao Takaku <masao@slis.tsukuba.ac.jp>
To: Masao Takaku <masao@slis.tsukuba.ac.jp>
Subject: =?iso-2022-jp?B?GyRCJUYlOSVIGyhC?=
Date: Thu, 30 May 2013 02:38:54 +0000
Message-ID: <9BB0C5FDD26EB142B3B76E53820B0FB9020BC402@SINPRD0310MB391.apcprd03.prod.outlook.com>

--_002_9BB0C5FDD26EB142B3B76E53820B0FB9020BC402SINPRD0310MB391_
Content-Type: text/plain; charset="iso-2022-jp"

これはテストです。

高久

--_002_9BB0C5FDD26EB142B3B76E53820B0FB9020BC402SINPRD0310MB391_
Content-Type: text/plain; name="20130529-klis-tokuron.txt"
Content-Description: 20130529-klis-tokuron.txt
Content-Disposition: attachment; filename="20130529-klis-tokuron.txt"
Content-Transfer-Encoding: base64

jcCSa4nvgUaR5Yp3iUANCg0KgUWOv5biDQoNCi0gll6ObYzjivqJ25L2gsyQaZhlgs2Cx4KkgsiC
wYLEgtyCt4KpgUgNCi0gk/yOjoLMjduCzINlgVuDfYLGjKSLhoNlgVuDfYLNlc+CpoLEguCR5Y/k
lXaCxYK3gqmBSA0KLSCM9paxiPWOdZZdgsaCqYLBgsSDQYOKgsWCt4KpgUgNCg0K

--_002_9BB0C5FDD26EB142B3B76E53820B0FB9020BC402SINPRD0310MB391_--
```

- Content-Type: multipart/mixed; boundary="_002_9BB0C5FDD26EB142B3B76E53820B0FB9020BC402SINPRD0310MB391_”

複数パートの区切り文字を全体で指定

- 各パートでの内容種別
  1. Content-Type: text/plain; charset="iso-2022-jp“
  2. Content-Type: text/plain; name="20130529-klis-tokuron.txt”

     Content-Disposition: attachment; filename="20130529-klis-tokuron.txt";

軽量テキストフォーマット

- 各種Wiki/Blog記法
  - MediaWiki
  - Markdown
- HTML
- LaTeX

Wiki記法

- WikiWikiWeb等により提唱
- 簡易なテキストフォーマット＋スタイル整形
  - 文書要素：段落、見出し、リンク、図表、強調
  - HTMLへの変換により、Webでの発信や共有を前提とした軽量マークアップ言語の典型
- 標準化の欠如
  - ウィキソフトウェアごとに多くの亜流（数百種?）
- 例：
  - ’’’強調’’’
    - → 強調
  - [http://www.tsukuba.ac.jp/ 筑波大学]
    - → 筑波大学

MediaWiki記法

- https://ja.wikipedia.org/wiki/Help:%E6%97%A9%E8%A6%8B%E8%A1%A8
- 世界最大の百科事典Wikipediaの執筆に用いられている記法
- 特徴
  - Web上での発信用途
  - 誰でも記述できるようHTMLを簡易化した記法
  - 図表や数式の表記まで対応
  - HTMLマークアップにもそのまま対応
  - テンプレート（transclude）

Markdown記法

- http://daringfireball.net/projects/markdown/
- Web上で提供されるソフトウェア開発やマニュアル、ブログ等で用いられる軽量テキストフォーマット
- 特徴
  - Web上での発信に加えて、HTML以外のフォーマットのサポートも。
  - 複数の処理系の存在
    - （処理系ごとの互換性は不完全）
- 例:
  - **強調**
    - → 強調
  - [筑波大学](http://www.tsukuba.ac.jp)
    - → 筑波大学

まとめ

- ドキュメントフォーマットとは？
- ドキュメントフォーマットの切り口
- フォーマットの識別、判別
- プレインテキストのフォーマット
  - テキスト / 文字コード
  - ドキュメントフォーマットの例
    - Eメールのフォーマット
    - 軽量テキストフォーマット