# 情報処理の概念

#6 HTMLにおける構造と表現、アーカイブ

Yutaka Yasuda, Kyoto Sangyo University

## Webと電子出版

- まず出版過程の電子化から
  - 1450 年頃：グーテンベルグの活版
  - 1960 年代：電算写植の導入
    - 日本では 70 年代に新聞CTSで活版から移行 (Cold Type System、Hot な鉛を使わない)
  - 1980 年代：DTP の登場
    - 1986 年：Macintosh / LaserWriter
    - WYSIWYG システムとPostScriptの出会い
- 成果物の電子化 - 紙との決別
  - 1990 年代：CD-ROM 出版、PDF、Web等多様なメディア

## PDF と電子出版との関係

- 出版過程の電子化の一段階
- 特徴：紙のイメージを保持
  - 紙を出力対象とした印刷技術の変遷の最終形態（かもしれない）
  - 最後に紙のイメージを電子的に表現する
  - レイアウトを完全に保持して再現
- 難点：機械可読性が弱い
  - データとして扱えるという意味で可読だが
  - 本来ドキュメントがもっている文章の意味や構造を無視して文字の並びだけを扱う傾向がある

## PDF と対比した Web

- 共通項
  - 紙ではなくデジタルデータとして出力
  - 文字情報については機械可読
- PDFの弱み
  - 文書の構造などを汲み出せない（タグつきPDFも Acrobat 5 から用意されたが）
  - 文字情報は印刷のため
  - 可読とはいえ再利用性が低い
  - 機械で読むのは再利用、加工のためでは？

## 情報加工、再利用

- 例：PDFドキュメントから第 3 章だけ抜き出しなさい
  - どこからどこまでが該当部分かわからない
  - 人間は見たらわかる（意味を理解している）
  - 機械にはわからない（アラビア語Web page を見た時に似る = 文字は見えるが再利用できない）
  - ページの切れ目を変えて再構成できない
  - あくまで「見ため」を残しているのであって、文章の構造は消えている

## 情報加工、再利用

- 例：サブタイトルが「印刷技術」の章を抜き出しなさい
  - タイトルが飛び跳ねてたら？
  - 「見ため」としての文字は残るが文は消える
- じゃあWebはできるのか？
  - Yes, HTMLがテキストの構造を記述するから
  - SGML の本来の価値に注目

## Webにおける構造と表現

- HTML (Hyper Text Markup Language) とは何か
  - SGML (Standard Generalized Markup Language) の一つの例
  - コンピュータ間でデータを交換するために
  - 情報の属性を記述する
- 本を SGML で記述する
  - これがタイトル
  - これがパラグラフ
  - ということがわかるように

SGML による記述例

```
<BOOK>
  <HEAD>
    <TITLE>情報処理の概念</TITLE>
    <AUTHOR>安田豊</AUTHOR>
  </HEAD>
  <BODY>
    <ABSTRACT>
      情報処理技術の様々な応用、すなわちコンピュータやネットワークの
      利用が 進んだ結果、、、、
    </ABSTRACT>
    <SECTION>SGMLについて</SECTION>
    <SUBSECTION>文法</SUBSECTION>
    <PARAGRAPH>SGML はタグと呼ばれる、、、</PARAGRAPH>
    <PARAGRAPH>このとき、ドキュメントは、、</PARAGRAPH>
    <SUBSECTION>目的</SUBSECTION>
      .....
  </BODY>
</BOOK>
```

構造が残されていることに注目
これなら「三章を抜け」も可能

## 情報加工、再利用

- 溢れる情報
  - 情報発信者の激増
  - 通信環境の改善、能力アップ
- 新しいモデル
  - 将来流れる情報はまず機械が読む
  - 個人向けに再編成してから読む
  - 興味のあるニュースだけ集めるシステム
  - 「新しい本の情報を見つけたらABSTRACTだけ集めて見せてくれ」
  - 「ABSTRACTにこのキーワードがあるものだけ」

## アプリケーション例

- 機械翻訳
  - Webページ自動翻訳
- ロボット型検索エンジン
  - HTMLの機械可読性が活きている
  - 一次情報はまず機械が読むという感覚
- HTMLのまずさが問題に
  - 視覚的表現に重点が移行

## HTMLのまずさ

- 理想
  - 構造を表現すればそれなりに見せてくれる
  - 構造の記述と好ましい表現の両立
- 現実
  - より良い見た目のために記述を工夫する
  - 構造の表現が崩れても構わない
  - 一文字ずつ離して配置する
  - 絵で文字を代行させる（見出しなど）
  - プログラムで表現（Java, Flash など）

**情報処理の概念**

安田豊 / 2004.11.10

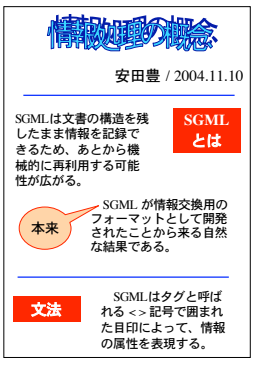
•SGMLとは

SGMLは文書の構造を残したまま情報を記録できるため、あとから機械的に再利用する可能性が広がる。

本来 SGML が情報交換用のフォーマットとして開発されたことから来る自然な結果である。

•文法

SGMLはタグと呼ばれる＜＞記号で囲まれた目印によって、情報の属性を表現する。



## HTMLのまずさ

- 長さの問題
  - 見た目上の理由でページを分けてしまう
  - 短すぎるページ構成となる（細分化されすぎ）
  - HTMLは一文書で完結する設計
  - 本来は巨大なマニュアル本を一つのSGML文書で表現するような設計目標だった
- 構造をどこで表現するか？
  - 建前：一つの文書内でタグによって表現
  - 現実：リンク関係によって表現
  - 「画面一枚の情報に、書くべき構造なんて無い」

## HTMLのまずさ

- Googleの的確な候補表示はどこから？
  - 必要なキーワードを含んでいるページのリンク関係を見て、
  - 人気があり、
  - 入り口と思われるページを割り出す
- それでも機械可読であることの重要性
  - 本来の設計目標とは違っても、機械可読である限り工夫は可能
  - まだまだ Web を有効に利用するためのアプリケーションはある

## HTMLのまずさ

- 完成、版という概念がない
- 利点
  - 即時性は高い
  - 融通も利く
- 欠点
  - リンクが切れる
  - 固定できないため、相互参照に意味がない
  - 情報が失われる
- HyperText は本来そうではなかった

## HyperText のアイディア

- 1981, Literary Machines - Ted Nelson
  - Xanadu - 完成していないプロトタイプ
- 出版すると同時に固定され、改変不可
  - 改訂版は簡単に出せるが、旧版も残る
  - リンクが切れず、意味も変わらずに使える
- 明確な文書の境界線
  - ページ単位ではなく文書単位で出版（登録）
  - 外部参照（リンク）と引用（トランスクルージョン）の使い分け
  - HTMLでは他の文書へのリンクと、自文書の一部分へのリンクに区別がない
- はじめから永続的アーカイブが前提だった

## HyperText のアイディア

- T.B. Lee は ‘89にWebを開発したが
  1. 構造の記述と見た目の表現の混在
  2. 文書の固定とリンクの消滅の関係

  の二点について解決せずに Web/HTML をリリースした
- Web保存計画はその反動である
  - 例えばWARP
- 提案
  - Webではない新しいシステムの開発
  - Web/HTMLでも構造記述と表現の両立を目指す

## その他の電子アーカイブ

- 過去の著作物から積極的に電子化
- 著作権法の期限外のものから
  - グーテンベルグ計画
  - エキスパンドブック
  - 青空文庫

# グーテンベルグ計画

- http://www.gutenberg.org/
- イリノイ・ベネディクティン大学マイケル・ハートが推進
- 1971年開始
- 2001年までに10,000タイトル電子化目標
- 2004.11現在 6000 超ほど
- テキストのみ
  (ASCII 以外に Swedish などもあり)
- XML で楽譜を集める The Sheet Music Subproject も始まっている

Bible のグーテンベルグ例

Bible
Genesis Chapter 1
God createth Heaven and Earth, and all things therein, in six days.

1:1. In the beginning God created heaven, and earth.

1:2. And the earth was void and empty, and darkness was upon the face of the deep; and the spirit of God moved over the waters.
....

単なるテキスト情報のみ

# エキスパンドブック

- ボイジャー が開発 www.voyager.co.jp
- 対象
  - テキスト中心の電子出版
  - 動画、音声なども含めたマルチメディア出版
  - 縦・横組、文字サイズ、行間、字間の指定等さまざまな文字組が可能。
  - ルビ、禁則に対応。
  - 指定した通りのデザインを、WinでもMacでも、読者のマシンで忠実に再現

# 青空文庫

- http://www.aozora.gr.jp/
- 特徴
  - 利用に対価を求めない、インターネット電子図書館
  - 著作権の切れたもの、自由に出せるものを対象
  - テキストとHTML、エキスパンドブックで提供
    現在は XHTML に注力
  - 1997年スタート
  - ボランタリで入力、校閲
  - 収録作品数 4199 本（2004.11 現在）
  - 世界に誇れる日本発のプロジェクトとなるかも
  - ネットワークに散在する力を集めたという意味で極めてインターネット的
    （「むしとりあみ」という誤植連絡窓口の価値）