青空文庫

デジタルアーカイブ学会第2回研究大会

日本語デジタルテキストの「正書法」を探求した青空文庫：

日本語（による／のための）マークアップの誕生とルールの発展・活用、テキストの品質管理

(大久保ゆう)

14:50-15:10

2018.3.10@東京大学本郷キャンパス

自己紹介

大久保友博（PN:大久保ゆう）
［青空文庫・本の未来基金］

専攻：翻訳文化史［京都橘大学］

フリーランス翻訳家

（1998年［高１］から
青空文庫に参加）

青空文庫とは

1997年創設の
ボランティア主体の
デジタルアーカイヴ

14000作の電子テクスト
(パブリックドメイン含む)

「本を電子化して**誰でも読める**ようにしておくと
面白い」→「青空という**自由な本棚**に集める」

青空文庫のアーカイヴ作業

↓　ルール［青空文庫注記］に従って電子テキスト化

↓　自由にコピー＆翻案可能なものとなる

↓　社会の公共物としての書架と蔵書とルール

→　さまざまな活用への展開

原文入力ルールの採用と公開

●1997年当時の青空文庫

「青空文庫工作員マニュアル」

視覚障碍者読書支援協会（BBA）
「原文入力ルール」に基づく

青空文庫

初期ルールの一例

【ルビ】

✓闇《やみ》の中を跳梁《ちょうりょう》するリル

✓表情?豊《ゆた》かな

【傍点注記】

✓天界の牧羊者［*「天界の牧羊者」のすべての文字に傍点］

【外字注記】

✓［#「うしへん」に建、16-2］陀多《かんだた》

✓森鴎［#「メ」の代わりに「品」、115-7］外

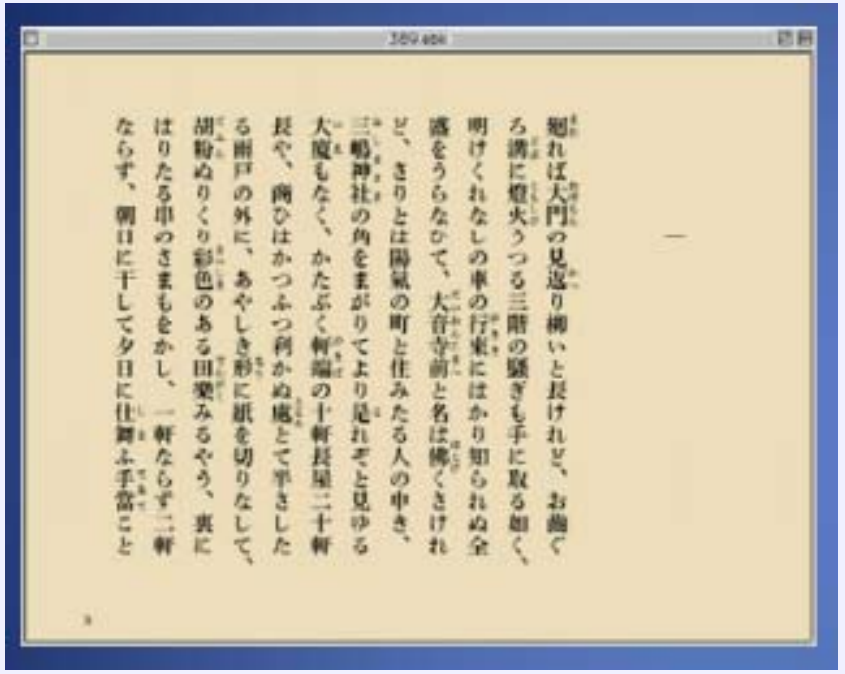
青空文庫

電子書籍のマークアップとして

電子書籍ファイル制作用の
組版指示として
[まだ体系化されていない]

↓

JIS文字コード収録漢字選定
のための資料整理から
マニュアルの整備へ



一

知れば大門の見返り柳いと長けれど、お歯ぐろ溝に燈火うつる三階の騒ぎも手に取る如く、明けくれなしの車の行来にはかり知られぬ全盛をうらなひて、大音寺前と名は仏くさけれど、きりとは陽気の町と住みたる人の申き、三嶋神社の角をまがりてより是れぞと見ゆる大廈もなく、かたぶく軒端の十軒長屋二十軒長や、商ひはかつふつ利かぬ風とて手きしたる格子の外に、あやしき形に紙を切りなして、胡粉ぬりくり彩色のある田楽みるやう、裏にはりたる串のさまもをかし、一軒ならず二軒ならず、朝日に干して夕日に仕舞ふ手当ごと

本文中の※は、底本では次のような漢字（JIS外字）が使われている。

※々（くの字点）と 𠮟

※（まぶた） 𥃳

※（はこ） 𣏌

青空文庫

1998年末改訂ルールの一例

【ルビ】

✓耳まで火照《ほて》って来る

✓一応｜何時《いつ》もの

【傍点注記】

✓胡麻塩おやじ［＃「おやじ」に傍点］

【外字注記】

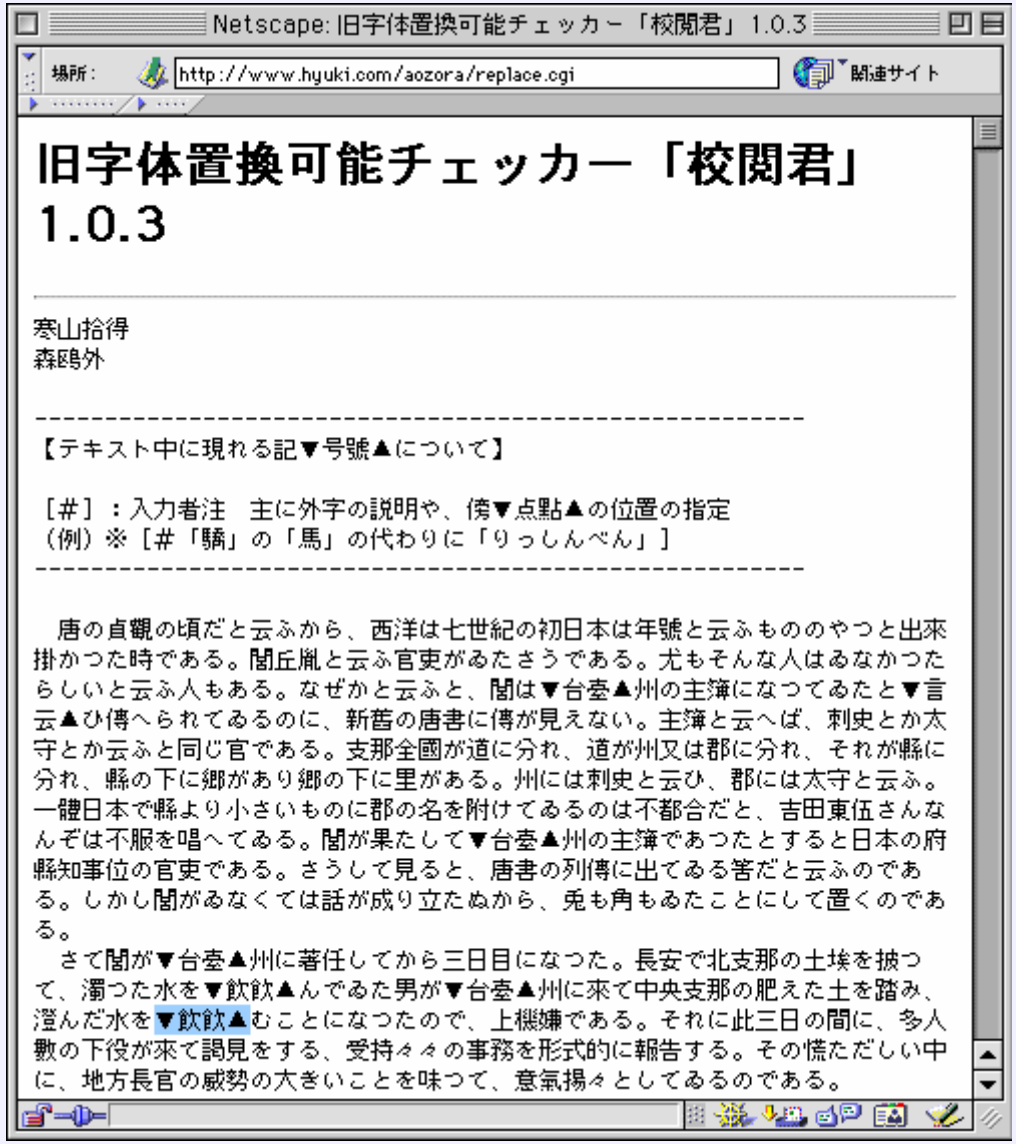
✓喉を掻き※［＃「※」は「てへん＋劣」、読みは「むし」、30-16］って

✓１［＃「１」は底本では○付き数字］インターネット

JIS X 0208範囲内での作業とツールの発展

JIS X 0208の範囲で
入力可能な旧字を探す
校閲君のチェック結果

あやしい文字と代替候補
を「▼▲」で挟んで
示してある



青空文庫

JIS X 0208範囲内での作業とツールの発展

あお・あおへん【 青 】 目次に戻る

4. 靖 ※［＃「正＋睛のつくり」、第4水準2-91-92］
7. 靚 ※［＃「靜のへん＋見」、第3水準1-93-75］
8. 靛 ※［＃「靜のへん＋定」、第4水準2-91-94］

←外字注記辞書
2007年第五版
［記述の標準化］

【一例】

✓※［＃「登＋おおざと」、第3水準1-92-80］
✓※［＃丸1、1-13-1］
✓※［＃「目＋争」」、U+7741、ページ数-行数］

青空文庫

注記の発展とめぐる論争

【アクセント注記】　（上が1998年12月時点；下が改訂後）

✓ae ao，ae ao，eo，aeo eo！［＃この行の「e」はすべてアクサン（´）付き］

✓〔ae‘ ao，ae’ ao，e‘o，ae’o e‘o〕！

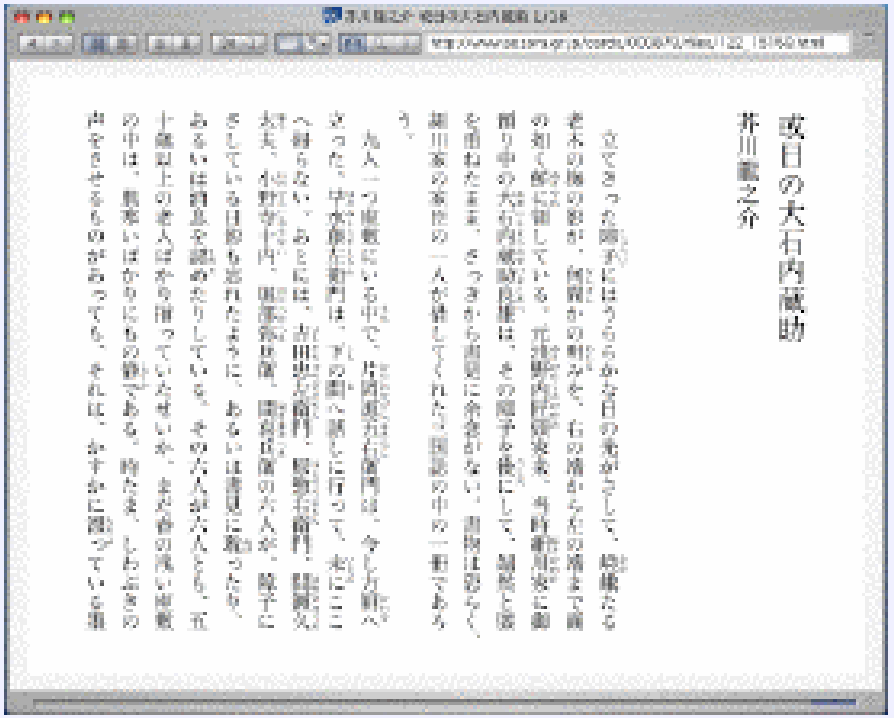
変換例：　aé ao, aé ao, éo, aéo éo!

【区点番号5-17と5-86の使い分け・ケヶ問題】

✓※底本は、物を数える際や地名などに用いる「ヶ」（区点番号5-86）を、大振りにつくっています。

注記の再現と構造化テクストへ

情報を残すための注記から
再現するための注記へ
↓
各種ビューワの登場



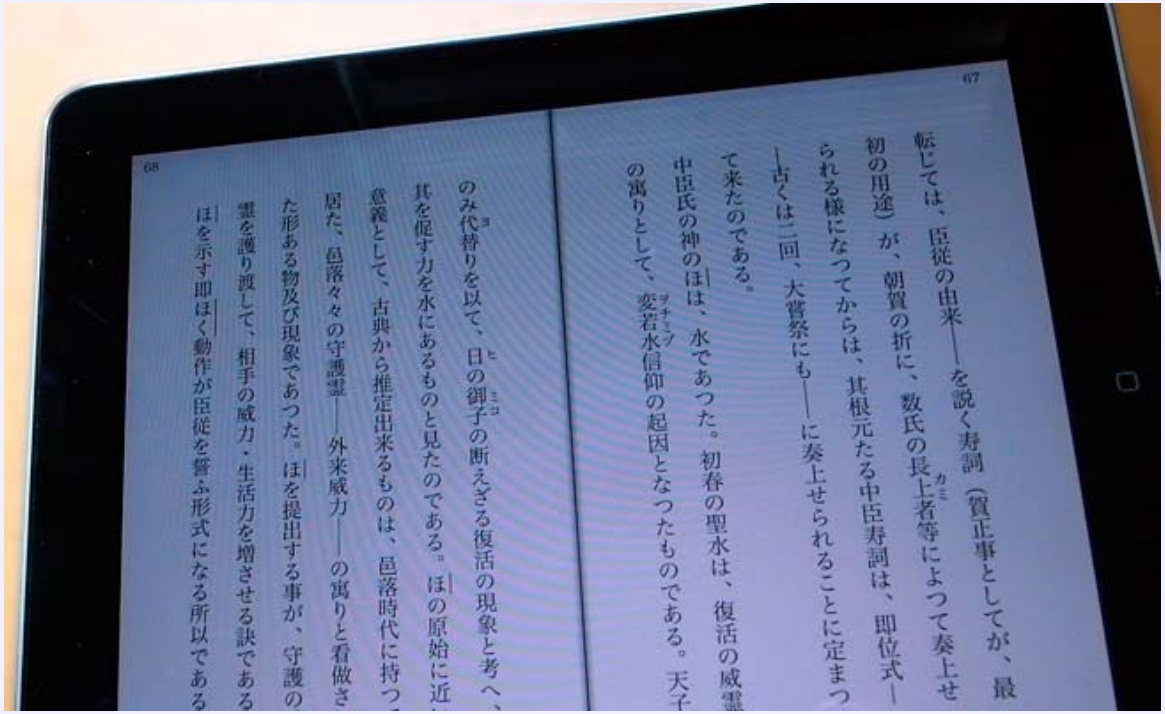
富田倫生 @aobeka　10月14日

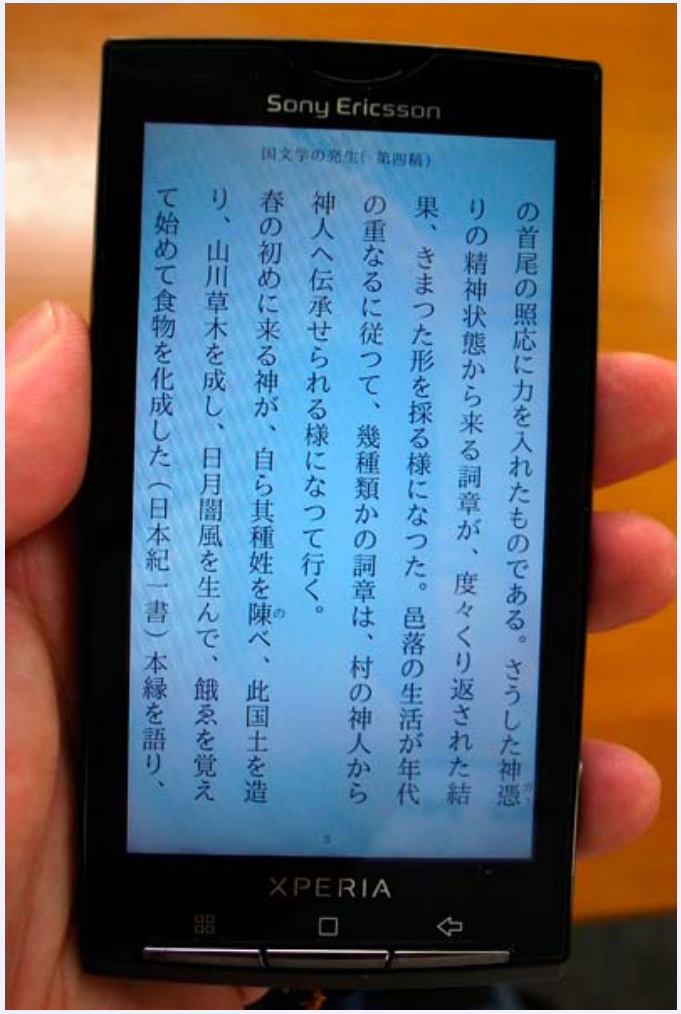
今日は、総務省の昼休み勉強会でお話ししました。青空文庫の蓄積するものは、「本を模したファイル」から、「構造化したテキスト」に変わった。アーカイブの書籍電子化にとっては、必要な変化だったと思う、と言いました。 #aozorabunko

posted at 18:25:29

青空文庫

青空文庫の活用について





タブレットやスマートフォンへ

ビューワの発展による変化

- 読込：xhtmlからtxtへ
- 媒体：ＰＣ→ＰＤＡ→タブレット／スマホ
- iPhoneの登場（2007年～）

  テキスト版のアクセス急増（2010年～）

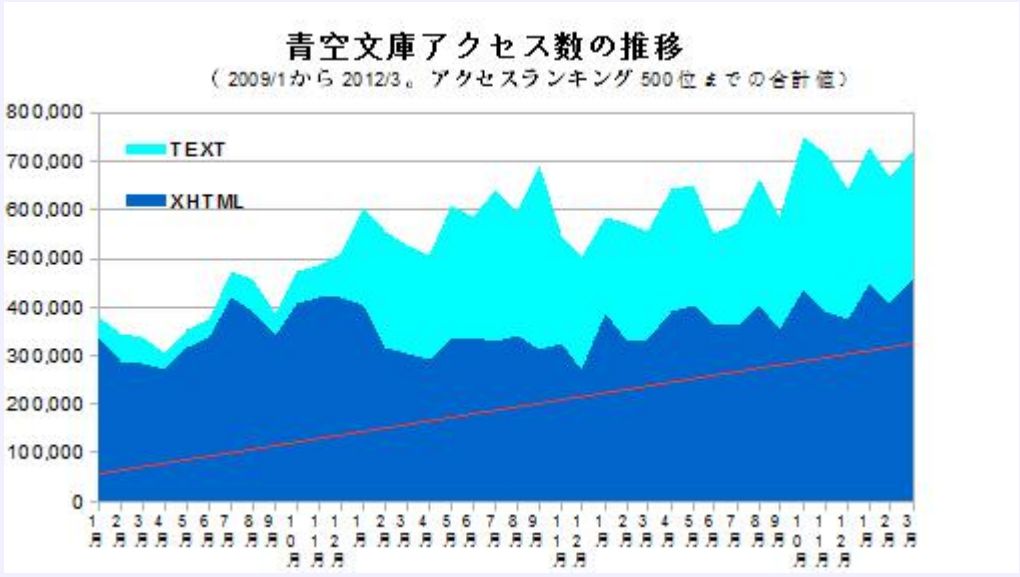


図　アクセス数の推移

（POKEPEEK2011による）

http://www.aozora.gr.jp/aozorablog/?p=351

ビューワの発展と対応

- ✓ 前方参照型 → 開始／終了型

  例：○○［＃「○○」に傍点］

  →［＃傍点］○○［＃傍点終わり］

- ✓ 注記一覧／組版案内の公開（2010年）
- ✓ 構造化と「点検 verification」

間違いだらけの電子テキスト⇩
青井文⇩
⇩
＿夏休みになった。海へ、プールへ、ディズニーへ。澁滞《じゆうたい》がたいへんだ。＿⇩
＿「わぁ、きれいな海。⇩
泳ぎましょう。」⇩
＿この文書にはハンカク（半角。反核じゃないよ）のカタカナが紛れ込んでいます。⇩
＿電気自動車《エレクトリックカー》に乗ろう（笑）。
Hop_onto_an_electric_car.Let's_take_it_out_for_a_spin.⇩
＿あそこにいる奴隷に喝采した。⇩
＿自分白身何をやっているか分からない。⇩

おわりに

- ✓ 青空文庫は、日本語でのマークアップが原則

  → 日本語話者・コンピュータ　双方の可読を意識

- ✓ 青空文庫の注記は、日本語文書の

  電子テキスト化における事例と解決法の蓄積

- ✓ 課題：

  ・公共化しつつあるマークアップの今後の保守と管理

  ・培われてきた知見の活用