試験問題データ：テキスト化のための仕様

2014/4/21

# 1. 全般

ファイル名

* テキスト化したファイル名は、元のPDFの名前と同じにする。名づけ規則は次の通り。
  + センター試験の場合：Center-{年度（西暦4ケタ）}--{本試験と追試験の区別}-｛科目名｝.txt
  + 代ゼミセンター模試の場合：YozemiCenter-{年度（西暦4ケタ）}--{模試の回番号（半角数字）}-｛科目名｝.txt
  + 代ゼミ東大模試の場合：YozemiTodai-{年度（西暦4ケタ）}-{文系・理系の区別}-{模試の回番号（半角数字）}-｛科目名｝.txt
  + 本試験と追試験の区別は、本試験の場合はMain、追試験の場合はSupp1と表記
  + 文系・理系の区別は、文系の場合はBun、理系の場合はRiと表記。
  + 科目名は次の通り。
    - Kokugo（国語）
    - Eigo\_hikki（英語筆記）
    - SuugakuIA（数学IA）
    - SuugakuIIB（数学IIB）
    - SekaishiB（世界史B）
    - NihonshiB（日本史B）
    - ButsuriI（物理I）
  + ハイフンはすべて半角。年度と本試験・追試験別を区切るハイフンは二本（--）とする。

文字のテキスト化

* 文字コードはUTF-8で統一する。
* UTF-8で表示できない文字があった場合は、<unk>タグを使ってその情報を記述する。（詳しくは、問題構造アノテーション仕様書ver.8の「11.テキストの装飾」参照）
* 大問番号、小問番号、参照用データのセクション見出し（Ａ、Ｂ、Ｃなど）、選択肢の丸付き番号およびそれらへの参照を除き、ファイル中の数字、アルファベットはすべて半角とする。
* 句読点は、国語問題を除き、「，」（全角カンマ）と「。」を使用する。
* 国語問題は、「、」と「。」を使用する。
* 英語問題中の英文は、カンマとピリオドの後は半角スペース1つ分を開ける。

テキスト化対象外の部分：以下の部分は不要。

* 表紙、裏表紙、下書き用紙、草稿用紙、計算用紙、解答ページ
* ページ番号、ページのヘッダの情報
* 「試験問題は次に続く」などの文言

# 2. ファイル先頭

* 冒頭にファイル名（拡張子なし）を入れる。
* その後に、年度、本試験・追試験の別、科目名を以下のように入れ、これをタイトルとする。元データにこれに類する情報があっても、テキストデータには入れる必要なし。

例：

2013年度　本試験　世界史B

2013年度　追試験　数学IIB

# 3, 番号

* 大問番号は、全ファイルで【　】（墨付きカッコ）に全角数字で入れる（例：第１問　→【１】）。国語の場合のみ、漢数字を入れる（例：【一】）。
* 小問番号は、「問」の後に、1ケタならば全角数字、2ケタならば半角数字で入れる（例：「問３」（1ケタなので全角）、「問17」（2ケタなので全角））。アルファベットの場合は半角とする。
* セクション見出し（世界史、日本史、英語に見られるＡ、Ｂ、Ｃなど）は全角アルファベットとする。ひらがな・カタカナの場合は全角とする。
* 選択肢番号は、丸付き数字（①、②など）とする。
* 解答欄、空欄、傍線部、下線部その他のラベルは、英数字の場合、すべて半角とする。ひらがな・カタカナの場合は全角とする。
* 上記の各項目への参照（「問９に答えよ」の「問９」、「次のＡは縄文時代の遺跡に関する記述である」の「Ａ」、「下線部bに関して」の「b」など）は、参照先の番号と同じ書式にする。

# 4. 解答欄および空欄

* センターの解答欄および空欄（四角で囲まれた部分）は画像ではなくテキストで番号のみとする。四角は再現する必要なし。
* 解答欄、空欄の前後には、半角スペースを一つずつ入れる。

# 5. 画像

* 問題紙面から抽出する画像は、png形式とする。
* 複数の画像が固まっている場合は、それぞれ切り分けて別画像とする。
* 各々の画像の命名規則は以下に従う。

{ファイル名（拡張子なし）}-{三ケタの数字}.png

例：

Center-2012--Main-SekaishiB-001.png

Center-2013—Supp1-ButsuriI-002.png

* 画像ファイル名の最後の数字は、試験問題内での出現順に振られているのが望ましいが、厳密に出現順になっていなくても問題はない。