「くずし字データセット」データ作成（基本仕様）

「日本古典籍くずし字データセット」の各データについて、国文学研究資料館（以下「当館」という。）がどのような仕様で発注しているか、仕様書の内容を提供することにより、「くずし字データセット」に汎用性をもたせていきたいと考えています。

業者に発注する場合はもちろんのこと、個人でデータを作成する場合の作成ルールとして、ご活用ください。

朱書き部分は当館の部署、規程等に関わる部分及び本基本仕様に対する注（＄表示）ですので、各発注者の事情に合わせて変更ないし削除願います。

１．件名

くずし字データセット及び翻刻テキストデータの作成業務[＄1]

＄１　「翻刻テキストデータ」は、人文学オープンデータ共同利用センター（ＣＯＤＨ）から公開されている「日本古典籍くずし字データセット」には含まれていませんが、くずし字の字形データを作成する作業には、翻刻テキストデータを作成する作業を含みますので、効率性を考慮して同時に発注しています。なお、「翻刻テキストデータ」は、ＣＯＤＨから公開されている「日本古典籍データセット」で提供されます。

２．概要

【国文学研究資料館（以下「当館」という。）】から提供する歴史的典籍の画像情報（以下「原本画像」という。）に基づき次の５種類のデータを作成する。

（１）原本画像から頁毎に個々の文字が原則として正立しているよう角度補正を行い、資料以外の箇所をトリミングした画像（以下**「原本補正画像データ」**という。）を作成する。

（２）原本補正画像データから、一文字一文字に該当する文字の画像を切り出し、画像ファイル（以下**「字形画像データ」**という。）を作成する。

（３）作成した字形画像データに関する付帯情報をテキストファイル（以下**「付帯情報テキストデータ」**という。）として作成する。

（４）字形の判別が困難等の理由で字形画像データを作成しなかったものについて、その記録をテキストファイル（以下**「判別困難テキストデータ」**という。）として作成する。

（５）作成した付帯情報テキストデータ及び判別困難テキストデータから、原本画像の版面に合わせた翻刻文の電子的テキスト（以下**「翻刻テキストデータ」**という。）を作

成する。

3．対象作品

『(作品名)』他●点（別紙のとおり）

字形画像データ　　●●●，●●●字※1

※1　文字数はあくまで概算であり、作品毎に文字数は変動する。

4．納入期限

平成　　年　　月　　日（　）

5．納入場所

【国文学研究資料館古典籍共同研究事業センター事務室（以下「当館担当者」という。）】の指定する場所

6．納品物

次の（1）から（5）を各一式、ＤＶＤ－Ｒに格納して納品すること。

（1）原本補正画像データ

（2）字形画像データ

（3）付帯情報テキストデータ

（4）判別困難テキストデータ

（5）翻刻テキストデータ

7．前提条件

（1）【当館担当者】が提供する原本画像を基に作業を行うこと。

（2）作業に必要となるソフトウエア及び機器類は、受注者において準備すること。

（3）受注者は、作業予定工程表（任意様式）を契約締結後、速やかに【当館担当者】に提出すること。

（4）作業終了後、当館から提供された物品等（電子データ含む）は、【当館担当者】に返還すること。

8．詳細仕様

（1）基本要件

①　原本画像から、詳細仕様（2）の指示に従い頁毎に角度補正とトリミングを行い、原本補正画像データを作成する。

②　詳細仕様（3）の指示に従い、対象となる一文字一文字について、補正後原本画像から対応するくずし字の字形を切り出し、字形画像データを作成する。

③　字形画像データに関するファイル名、Unicode のコードポイント等を、詳細仕

様（４）の指示に従い付帯情報テキストデータとして作成する。

④　字形画像データの作成対象でありながら、歪み等の理由により字形画像データが作成できなかった場合は、詳細仕様（５）の指示に従い判読困難テキストデータを作成する。

⑤　詳細仕様（６）の指示に従い原本画像の版面に合わせた翻刻テキストデータを作成する。

⑥　作成した上記の５種データは、作品毎に作品フォルダを作成して格納する。補正後原本画像は、作品フォルダの直下に補正後原本画像フォルダ（”images”）を作成して格納する。字形画像データは、作品フォルダの直下に字形画像データフォルダ（”characters”）を作成し、その直下にコードポイント毎のフォルダ（"U+3042”等）を作成して、コードポイント別に格納する。付帯情報テキストデータ、判読困難テキストデータ及び翻刻テキストデータは、作品フォルダの直下に格納する。

（２）原本補正画像データの作成

①　原本画像のうち見開きで撮影されているものについて、頁毎に別のファイルとして作成すること。

②　頁毎に別のファイルとして作成した画像は、元の原本画像のファイル名に右頁なら”_1”、左頁なら”_2”を最後尾に付けてファイル名とする。※２

（例）200003076_00004_2.jpg

※２　元の原本画像が、見開きによる画像でない場合は、元の原本画像のファイル名をそのまま、原本補正画像データのファイル名とする。

③　資料内の文字が原則として正立しているように角度補正を行い、資料以外の箇所をトリミングする。

④　解像度の変更は行わない。

（３）字形画像データの作成

①　詳細仕様（２）で作成した原本補正画像データに基づき、次の（a）から（h）の条件に基づき字形画像データを作成するものとする。作成対象は Unicode にコードポイントが存在する文字のみとする。なお、作成対象でありながら、字形に歪みがある文字や、字形の判別が困難な文字等で字形画像データの作成が困難であるものについては、【当館担当者】に協議のうえ判断すること。字形画像データを作成しなかった字形について、その記録を詳細仕様（５）の判別困難テキストデータとして作成すること。

（a）図版は対象外とする。ただし、図版内の文字で正立している文字は対象とする。

（b）ルビは、対象外とする。

（c）訓点（ヲコト点、返り点、送り仮名等）は、対象外とする。

（d）合略仮名については、Unicode にコードポイントが存在する「ゟ（より、U+309F）」「ヿ（コト、U+30FF）」のみ作成する。それ以外は対象外とする。ただし、判別

困難テキストデータを作成すること。

(e) 本文中に使用されている記号や括弧なども作成する。

(f) 丁数（頁）の表示については、対象外とする。

(g) 表紙については、対象外とする。

(h) 書き込みについては、対象外とする。

② 字形画像データは、原本補正画像データから切り出す文字に外接する矩形の画像ファイルとして作成する。矩形は原本補正画像データに対して傾きがないように範囲を指定すること。余白は設けない。

③ 字形画像データは、ＪＰＥＧ形式（フルカラー※3）とする。解像度は、当該原本補正画像データと同じとする。

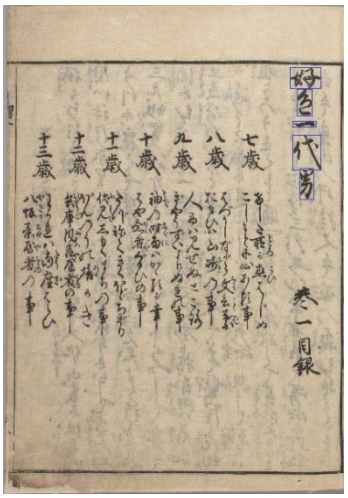
※３ 原本補正画像データがモノクロのものは、モノクロのままとする。

④ ファイル名は、(ⅰ)Unicode のコードポイント、(ⅱ)原本補正画像データファイル名、(ⅲ) 原本補正画像データに対する字形画像データのＸ座標※4※5、(ⅳ) 原本補正画像データに対する字形画像データのＹ座標※4※5の順番に、各項目を”＿（アンダーバー）”でつないだものとする。

※４ Ｘ座標及びＹ座標は、原本補正画像データの左上頂点を原点（Ｘ座標：０、Ｙ座標：０）とし、字形画像データの基準点は矩形の左上頂点（Ｘ座標が最小値である点のうち、Ｙ座標が最小値である点）とする。

※５ 単位はピクセルとする

（例）「好色一代男」



200003076_00004_2.jpg

U+597D_200003076_00004_2_X1741_Y0381.jpg

U+8272_200003076_00004_2_X1789_Y0515.jpg

U+4E00_200003076_00004_2_X1763_Y0709.jpg

U+4EE3_200003076_00004_2_X1758_Y0788.jpg

U+7537_200003076_00004_2_X1805_Y0995.jpg

（４）付帯情報テキストデータの作成

① 付帯情報テキストデータは、ＣＳＶ形式とし、ファイル名は BID(9桁)+”_coordinate.csv”とする$2。字形画像データ１ファイルに対し、１行（１レコード）とする。レコードの並び順は、原本補正画像データファイル＞ブロックＩＤ＞文字ＩＤの順とする。文字コードはＵＴＦ－８とする。別紙（付帯情報テキストデータ　フォーマット）参照。

＄２　「BID(9 桁)」は、当館の書誌ＩＤです。字形データセットには、作品の刊年等の書誌情報を持たせておりません。そのため、書誌情報を持っている「新日本古典籍総合データベース」や「日本古典籍データセット」からデータを参照できるように、作品を特定できる書誌ＩＤ（BID）を「くずし字データセット」の作品のＩＤとして使用しています。「新日本古典籍総合データベース」で公開されている画像であれば、その作品の書誌ＩＤを使用できます。それ以外の作品の場合は、BID と重複せず、書誌情報と関連付けが可能なＩＤ番号を設定されることをお薦めいたします。

② 事項は次の（a）から（o）のとおりとする（括弧内がヘッダ名称）。なお、（a）から（e）をアンダーバーで接続したものが、字形画像データファイル名となる。

（a）Unicode のコードポイント(Unicode)※6（変体仮名は、対応する現在の文字のコードで整理する。）

（b）原本補正画像データファイル名(Image)

（c）Ｘ座標(X)

（d）Ｙ座標(Y)

（e）ブロックＩＤ(Block ID)※7

（f）文字ＩＤ(Char ID)※8

（g）字形画像データのサイズ[幅](Width)※9

（h）字形画像データのサイズ[高さ](Height)※9

※６　踊り字（くの字点）のコードは次のとおりとする。

く(U+3031)

ぐ(U+3032)

※７　ブロックＩＤは、本文、ルビ、図版中の文字等の区分けで使用する。形式は、先頭に”B”を付けた４桁の数字で表す。例）B0001

原本画像及び左右頁毎に”B0001”から付番する。

※８　文字ＩＤは、ブロックＩＤ毎に付番された文字の順番を表す。形式は先頭に”C”を付けた４桁の数字で表す。例）C0001

※９　単位はピクセルとする。

③ 原本補正画像データと参照用翻刻テキストの文字が異なっている場合は、原本補正画像データに基づき Unicode のコードポイントを入力すること。判断不能な文字については作成を行わず、【当館担当者】に報告すること。

（５）判読困難テキストデータ

① 上記（３）①のなお書きで作成することとした、字形に歪みがあるものや、字形の判別が困難なものについて、【当館担当者】と協議した内容等の記録についてＣＳＶ形式でデータを作成すること。ファイル名は BID(9 桁)+”_report.csv”とする。事項は次のとおりとする（括弧内がヘッダ名称）。別紙（判読困難テキストデータフォーマット）参照。

（a）原本補正画像データファイル名(Image)

(b) X座標(X)

(c) Y座標(Y)

(d) 協議記録：字形画像データを作成しなかった理由(Report)

（６）翻刻テキストデータ

① 作成した付帯情報テキストデータ及び判別困難テキストデータから、原本画像の版面に合わせた翻刻テキストデータを作成すること。

(a) 原本補正画像データファイル名を入力し、改行する。

(b) 字形画像データを作成した翻刻テキストを入力する。原本画像の版面に合わせて改行する。

(c) 判別困難テキストデータを作成した文字も入力すること。Unicode のコードポイントの有無にかかわらず、可能な限り「**=**（不明文字記号）」を使用せず、使用する場合でも可能な限り「**=**（＊「偏」＋「旁」）」のように注を付記すること。合略仮名の「こと」については、「**=**」ではなく、「こと」を入力すること。

(d) 踊り字（くの字点）は、字形画像データ及び付帯情報テキストデータでは、Unicode のコードポイント U+3031（〱）又は U+3032（〲）を使用して作成するが、翻刻テキストデータにおいては、U+3031（〱）は U+3033（/〃）及び U+3035（\〃）に、U+3032（〲）は U+3034（/〴）及び U+3035（\〃）に、それぞれ置換して作成すること。

(e) 原本補正画像データファイル間は、1 行空行を入力すること。

② 翻刻テキストデータは、テキスト形式で作成し、文字コードはＵＴＦ－８とする。ファイル名は、「作品名」+".txt"とする。

９．サンプルデータの作成

（１）作業開始に先立ち、契約締結後７日以内（土・日曜日、祝日を除く）にサンプルデータの作製を行い、【当館担当者】に提出し承認を受けること。サンプルデータの作成対象については、画像情報のおおよそ見開き１ページ分とし、具体的な箇所については【当館担当者】の指示を受けること。

（２）サンプルデータの提出方法（格納メディア）は、ＤＶＤ－Ｒとすること。

１０．瑕疵担保責任

（１）瑕疵担保期間は、納入期限から１年とする。

（２）瑕疵担保期間中に瑕疵が発見された場合は、受注者の責任において瑕疵のない　状態に修復し、成果物の一部または全部を再納入すること。

１１．契約条件

【発注者の所属する法人の規程等を記載して下さい。】

１２．支払い方法

【発注者の所属する法人の規程等に定める支払い方法を記載して下さい。】

１３．完了通知書及び請求書の送付先

【発注者の担当部署を記載して下さい。】

１４．著作権

本件納入物品の著作権（著作権法第２１条から第２８条に定めるすべての権利を含む）は、納入物品の引き渡しをもって、【人間文化研究機構（当館）】に移転するものとし、受注者は次の事項について同意するものとする。

（１）発注者が任意に本件納入物品を改変すること。

（２）発注者が納入物品を任意の表示氏名で公表すること。

（３）受注者は、納入物品について発注者の意に沿わない公表を行わないこと。

１５．秘密保持

受注者は、発注者から提供された情報又は本件により知り得た情報（発注者が秘密　情報と指定した情報に限り、個人情報を含む）を契約履行期間中か否かに関わらず、本件の契約履行以外の目的に使用し、あるいは第三者に提供・開示又は漏洩してはならない。

１６．その他

（１）本件業務に関して疑義がある場合は、【当館担当者】と協議し作業を進めること。

（２）受注者が作製し【当館】に納入した成果物に係る一切の権利は、【当館】に帰属するものとする。

（３）作業従事者は、業務の遂行上知りえた情報を他に漏らし、または、業務を行う目的以外に利用してはならない。また、業務を離れた場合においても同様とする。

（４）請負業者は、定期的に【当館担当者】との打合せを実施し、問題点の確認や改善策の検討に応じること。

（５）本仕様書に記載のない事項について対応する必要が生じた場合、【当館担当者】と協議の上、その指示に従うこと。

以上