Vol.2009-CH-82 No.1
2009/5/23

デジタル文字の共有と継承について

當山 日出夫[†]

人文学では「文字」はきわめて重要である。フォント埋め込み PDF の利用は、確かに便利ではある。しかし、フォント（グリフ）とともに文字があつかえるようになることは、新たな問題点を生み出す。現在、デジタル文字を確実に共有できるであろうか。また、将来にわたって、同じ文字を利用できるであろうか。この問題について、『内村鑑三全集』デジタル版を事例に、考えることにする。

About the common use and the succession of the digital character

TOUYAMA Hideo[†]

Planning of the digital publishing of " UCHIMURA KANZO ZENSHU " is progressing. This is using font embedding PDF. However, the user in the after times can not see the same character. Font embedding PDF invents a new problem about the character. We must proceed with the research about the essence of the character.

1. はじめに

本稿の執筆においては，意図的な問題提起として，次のように作成した．

(1) 基本的な書式の設定は，情報処理学会の規定による．用紙の余白，文字サイズ，などである．

(2) しかし，使用のフォントについては，MS 明朝ではなく，主にヒラギノ明朝を使用する．現在の MS 明朝では，この論文は，筆者の意図どおりに書けない．

(3) 使用文字の範囲は，「0208」および「0213 (04)」を基本とする．しかし，その両方を使用している．また，それでも書けない部分がある．

(4) ただし，これ以外は，MS-Word2007（2003ではない）のページ設定にしたがっている．

その理由は，

(1)「文字」は「ことば」の表記とともに存在する．特に本研究会（CH）においては，「文字」が非常に重要な研究課題である．多言語処理の他，日本語に限っても，どのようなフォントで表記するか，字体について，研究対象となる資料との関係において非常に重要である．

(2) CH 研究会においては，人文学関係の研究者が多く参加している．デジタル環境における「文字」は，はたして，相互に共有でき，将来にわたっても安定して継承可能であろうか．「文字」の共有と継承なくして，人文学におけるコンピュータ利用の発展は望めない．逆に，デジタル技術は，人文学におけるこのような特性を支援するものでなければならない．

(3) フォント埋め込み PDF などによって，より多くの文字をフォントとともにあつかえるようになる可能性がある．だが，これはこれで，新しい問題を生み出す可能性がある．

このように考えたとき，フォント埋め込み PDF による資料・論文などは，いささか問題なしとはしない．情報処理学会が完全にペーバーレス化され，CH 研究会としても，その最初の研究会である．ここで，あえて，この問題について考えてみることとしたい．

† 當山日出夫
立命館大学グローバル COE 日本文化デジタル・ヒューマニティーズ拠点客員研究員

ⓒ2009 Information Processing Society of Japan

2. デジタル版『内村鑑三全集』

現在，デジタル版の『内村鑑三全集』の企画・出版（DVD）が進行している．
その企画の概要は，つぎのごとくである．

(1)．『内村鑑三全集』（全 40 巻，岩波書店，2001 年，第 2 刷）を，完璧に，ワープロで再現して，PDF 化（フォント埋め込み）したものである．使用のワープロソフトは「一太郎（ジャストシステム）」．

(2)．本文についての検索システムを用意して，『全集』の本文（ルビをのぞく）について，検索可能とする．そのために，PDF（フォント埋め込み）から，さらにプレーンなテキストファイルを抜き出して，検索用インデックスを作成するなどの作業が必要である．

(3)．現在（本稿執筆時点），『全集』（全 40 巻）の入力が終わり，PDF 化（フォント埋め込み）および，その検索システムを作成している段階である．

なお，『全集』のデータ入力，および，刊行は，「内村鑑三全集 CD-ROM 版出版会」（事務局：斎藤顕氏）による．その PDF 化，検索システムなどは，精興社が担当．また，『全集』をデジタル化するにあたっての，著作権などの権利関係については，すべて問題無く解決されている．

3. 内村鑑三と書籍版『全集』について

資料のデジタル化（フォント埋め込み PDF）の問題点を述べる前に，内村鑑三および『全集』について簡略にふれておく．

内村鑑三，その生没は，文久 3（1861）～昭和 5(1930)．近代日本を代表する思想家・基督信徒. 札幌農学校で学ぶ. その後, 農商務省に勤務し, 水産関係の調査研究に従事., アメリカに留学し，帰国後，第一高等学校に勤務する．教育勅語奉読式において「不敬事件」を機に辞職.『万朝報』などで執筆活動を行う. 日露戦争時, 非戦論を展開し退社. その後は，独自の「無教会主義」をかかげ，キリスト教布教活動に専心する．

代表的著作としては,『余は如何にして基督信徒となりし乎』,『代表的日本人』,『求安録』『地人論』などがある．また，雑誌『聖書之研究』を刊行．これは，亡くなるまで継続された．その著作は，現在までにいくつかの著作集・全集としてまとめられてきている．最終的には,『内村鑑三全集』（岩波書店，1981～，全 40 巻）に結集されている．この『全集』は，その後，2001 年に第 2 刷が，同じく岩波書店より刊行．なお,『全集』は，編年編集．したがって，ひとつの巻のなかに，書籍・雑誌・講演筆記など，種々の文章がふくまれる．

4. 文字使用の実態

デジタル版『内村』全集（未刊）の制作企画段階で，次のような問題が発生している．

（1）JIS 規格に無い文字をどう処理するか．

Unicode の CJK 統合漢字にも無い字がある．JIS 規格（0208・0213：04）に無い文字については，現在のデータでは，これを「今昔文字鏡」（エーアイ・ネット）に依拠している．

その代表としては，「懶惰」の「懶」の字がある．詳しくは後述．

（2）JIS 規格漢字内において，必要とする字体が無い場合．

たとえば，「葛・葛」などである．これは，JIS 規格の変更によって字体が変わった例として有名．

「葛」は「0208」

「葛」は「0213（04）」　ともに，面区点，1-19-75.

固有名詞などにおいては，どうしてもその字体を使用せねばならない場合がある．『全集』では，人名「葛巻星淵」および「葛飾」で使用．この場合，使用するフォントの規格への対応によって，強制的に二者択一になる．

※この箇所，本稿では，ヒラギノ明朝 ProNW3 と同 ProW3 のフォント切り替えで，表示してある．

（3）Unicode にある場合．

JIS 規格（0213：04）までは無いが，CJK 統合漢字にまで拡大すると，異体字が使用できる場合．「昂・昂」．「志賀重昂」の人名で使用される．

「昂」は，JIS 規格第一水準．1-25-23，U+6602.

「昂」は，U+663B.

この文字は，JIS 規格では包摂することになっている．

ただし，この文字は，実際のデジタル版『全集』のデータを見ると，今昔文字鏡によって入力されている．

ⓒ2009 Information Processing Society of Japan

(4) JIS 規格内における異体字.

「祈祷・祈禱」,「冒涜・冒瀆」などである.「0208」までの範囲内では,「祷」「涜」の,いわゆる拡張新字体の使用となる.しかし,「0213 (04)」では,いわゆる正字体の「禱」「瀆」が使用できる.

以上のようなデータとなっているには,次のような背景がある.

第一に,デジタル版『全集』を忠実に入力したものであること.というよりも,ワープロの編集画面において再現したといってもよいものである.したがって,可能な限り,内村鑑三の漢字の使用法に忠実であろうとしている.「葛」「昂」,などである.

第二には,このデジタル版『全集』の企画は今から 10 年ほど前にさかのぼる.したがって,いまから見れば古い文字規格「JIS X 0208」に依拠している.フォントは MS 明朝.そのため,現在であれば通常の市販のコンピュータで利用可能な「JIS X 0213:2004」の「禱」「瀆」が利用されないでいる.

5. 文字の問題

実際の具体例として「懶惰」を例にあげてみる.「懶惰 (らんだ／らいだ)」は,内村鑑三が非常によく利用する語である.ただ,内村は,この「懶」の字の右の旁を「頁」と書く字体 (異体字) を使用している.この「懶 (頁)」は,

(1) JIS 規格「JIS X 0213:2004」にはふくまれない.

(2) Unicode の CJK 統合漢字にはない.

現在のデジタル版の企画としては,Windows XP を基本に,UTF-16 (ただし,サロゲートペアをのぞく) を使用ということになっている.つまり,この文字は「無い」字である.

(3) その対応策として,出版会としては,「今昔文字鏡」を利用した.

このこと自体は,この時代における対応として,特に問題があったというわけではない.しかし,データの利用という点においては,現時点においては,また,将来における継続的な安定利用という視点からも,いささか問題なしとはしない.

図 1.『基督信徒の慰』国立国会図書館近代デジタルライブラリー

図 2.デジタル版『全集』PDF (フォント埋め込み) の該当箇所.

図 3.当該箇所をワープロ「一太郎 2009」にコピーした状態.

図 2 は,PDF を画像化して一部を切り出したものである.しかし,これを,オリジナルの PDF (フォント埋め込み) から,コピーしてワープロ (あるいはエディタ) で表示すると,字が化ける.このことは,今昔文字鏡が,JIS 規格内の同一コードポイントの文字に対して,別の字体を与えて表示するシステムであることを理解していれば,現象としては納得できる.しかし,利用にあたっては,困ることになる.

つまり,以下のような問題が生じる.

(1) まず,何よりも,内村の漢字の用法として「懶 (頁)」を使用していることを知らなければならない.通常の「懶」では,検索できない.デジタルの世界では,検索不可能な語は,存在しないことになってしまう.

(2) PDF (フォント埋め込み) であれば,それ自体を検索可能である.Acrobat の検索機能だけでも,かなりのことが出来る.だが,このとき,「懶惰」の語は,この表記の文字のままでは検索できない.「艶惰」と入力しなければならない.

(2) これは,あらかじめ,このデジタル版『全集』のデータは,今昔文字鏡で作成してあることを知っていなければならない.そして,その対応関係が,「懶 (頁) =艶」であることが判明していなければならない.

(3) 実際の利用において,論文を書くとき,該当箇所をワープロにコピーしたら,字が化ける (「艶」になる),このことに気づかなければならない.

(4) さらに不都合なことには,内村は,「艶」という字を使用している.「懶 (頁) =艶」ということを知っていたとしても,本来の用字である「艶」と区別して処理しなければならなくなる.たとえば,『求安録』(全集第二巻, p.199).

以上のような問題点をかかえる現在のデジタル版全集の企画の方針にしたがうかぎり,対応策としては,次のいずれかになる.いずれにせよ,また別に,かなりの問題を生み出すことになる.

(1) 本文の校訂をあらためて,通行の「懶」に変える.これは,あくまでも,内村の用字に忠実であろうという方針に反することになる.

(2) 今昔文字鏡 (あるいはインデックスフォント) のライセンスをクリアして使用する.この場合,プレーンなテキストで字が化ける,検索に支障が出る,などの問題が発生する.

(3) この文字の箇所だけを画像データにする.この場合は,検索ができない.また,論文に引用しようとした場合,字が抜け落ちるなどの支障がでる.

現実的な問題として,テキストから字が抜け落ちたり,化けたりしたら,それはもは

ⓒ2009 Information Processing Society of Japan

や絶望的な状態といってよい．ワープロにコピーして引用した箇所を，書籍版『全集』と見比べなければならないことは（文献研究としては，厳密には必要な手続きではあるが)，学術資料のデジタル化と利活用という視点からは，大きな課題である．そのままでは信用できないテキストを提供することになるからである．

そうであるならば，あえてこのような箇所は，「ゲタ（〓)」で表示するという手段もあり得る．しかし，この方式は，賛否両論がある．

6．フォント埋め込み PDF の問題点

人文学研究のコンピュータ利用では，「文字」の問題が常につきまとう．これは，基本的に，次の２点に要約される．

（1）その文字があるか無いか．字体・字種の問題．

（2）文字化けするかしないか．同じ字体として見えるかどうか．

これは，かつての「0208（78)」から「0208（83)」，また，新たに「0213：04」で，人文学研究者が直面してきた，また現在も苦慮している問題である．この他に，エンコードの方式の問題もあるが，これは，人文学とは離れて技術的な問題としておく．

この問題を一挙に解決するかのごとくに思えるのが，フォント埋め込み PDF である．筆者としても，この方向に反対するものではない．しかし，それで全ての「文字」をめぐる問題点が解決されたわけではない．PDF（フォント埋め込み）が保証しているのは，コンピュータの画面で「見たときの同一性」のみである，といってもよい．これが可能になったが故に，また，新しい問題点を生むことにもなる．あるいは，逆にいえば，デジタルの「文字」が持っている問題点を隠してしまうことになる．

ちなみに，この筆者のこの論文=PDF（フォント埋め込み）をコピー（テキストファイル）して，確実に誰でもが共通に利用可能であろうか．答えは不可能である．

(1)「葛」の字体のつかいわけが，テキストファイルでは表現できない．「0208」によるか「0213（04)」によるかで，強制的に字体は統一されてしまう．

(2) 古いコンピュータで，「0208」だけの環境では，「0213（04)」の第三第四水準文字は見えない．Unicode 統合漢字も問題である．「禱」「濆」「昻」などである．

(3) フォントを埋め込めるが故に，フォントデザインのレベルまでふくめて「文字」をあつかうことになる．これが，一番の問題である．

PDF（フォント埋め込み）は確かに便利である．だが，それは，PDF をディスプレイで表示している限りにおいてである．その限りおいて，微細な字体の違いまで，確実に表現し，保存し，伝達可能である．また，プリントアウトもできる．

デジタルの「文字」の課題は，今後，技術の進歩によってさらに解決されるであろう．しかし，フォントデザイン（グリフ）をふくめて，より多くの字種が使用可能になるということは，新たなパンドラの箱を開けてしまうことになりかねない．ここは，過去のコンピュータと文字の経緯，人文学研究にとって文字とは何であるのか，さらには，そもそも文字とは何であるのか，あらためて考えねばならない．具体的に可視化されたグリフと，抽象的な文字概念の関係について，さらなる検討の必要がある．デジタル文字の共有と継承について考える新たな段階にいたったと考える次第である．

参考文献

1）當山日出夫．2008.「文字とアーカイブ－デジタル・アーカイブの視点からの問題提起－」.『情報処理学会研究報告 2008-CH-79』（金沢文庫）．pp.23-30

2）當山日出夫．2008.「文字を残すための序論的考察」．アート・ドキュメンテーション学会，第1回秋季研究発表会（2008年12月6日，印刷博物館）．pp.7-10

3）當山日出夫．2009.「『内村鑑三全集』デジタル版の文字処理について」.『東洋学へのコンピュータ利用 第20回研究セミナー』(京都大学人文科学研究所附属漢字情報研究センター)．pp.5-18

図 1

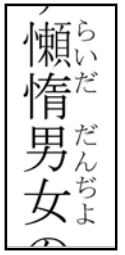

図 2

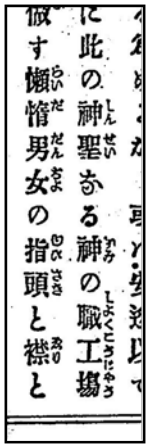

図 3

ⓒ2009 Information Processing Society of Japan