2008－CH－79 (4)
2008／7／25

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 －

Characters and Digital-Archives

當山 日出夫（TOUYAMA Hideo）
htoym@kcn.ne.jp

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）
Ritsumeikan-University Global-COE（DH-JAC）

文字とアーカイブとは、きわめて密接に関連する。だが、従来この方面からの考察はあまりなされてきていない。本稿では、その重要性を指摘するとともに、論点の整理の一端をこころみる。
(1).紙の文書では、それをアーカイブして残すことが、文字を残すことであった。
(3).今のわれわれが、どのような文字で、デジタル文書を見ているのか、これもまた、文書のアーカイビングの重要課題である。
(2).デジタル化文書では、それが分離するので、それぞれに保存の方策が必要である。

The character and the archive are related closely in the emergency. However, until the present, we aren't considered from this area. In this paper, I point out the importance.
(1). We do a paper document in the archive. We are to leave a character.
(3). We in now see a character with the digital document. This, too, must be done in the archive.
(2). The character and the document data must be left in the future.

【00】はじめに

この原稿は、Windows Vista、MS-Word 2007、JIS X 0213:04、MS 明朝・MS ゴシック・Century メイリオ、という環境で書いている。では、この文書『情報処理学会研究報告 2008-CH-79-4』(金沢文庫)が、なにがしかの形態でデジタル化保存されたとき、将来(すくなくとも数十年後)、同じ「文字」の文書として見ることが可能であろうか。技術的な問題もさることながら、まず、このような問題点の発想を、今のわれわれは、共有し得ているであろうか。本発表では、根本的な発想の次元から、文字とアーカイブの諸問題について、論点の整理をこころみるものである。

【01】文字とアーカイブの二つの論点

文字とアーカイブについては、基本的に、二つの方向から考えなければならないと、筆者は認識している。

(1).現在の文書が、アーカイブとしてデジタル保存された場合、同じ文字(字体・グリフ)で、見ることが可能であるかどうか。

(2).現在、われわれが使っている文字、21 世紀初頭のコンピュータ環境で見ている文字がどのようなものであるのか、これを、将来にわたって残し、記録・保存する必要性について、どう考えるべきであるのか。

以上の二つの論点がある。簡潔に表現すれば「デジタル・アーカイブの文字」と「文字のデジタル・アーカイブ」である。

それ加えて、現在(2008 年)では、コンピュータの文字としては、JIS X 0208 と JIS X 0213:04 とが共存している。現実的には、Windows XP と Windows Vista の問題でもある。また、さらに、現在、「(新)常用漢字表」の改訂が問題になっている。「常用漢字表」の改訂は、JIS 漢字規格に影響を及ぼす可能性がきわめて大きい。いま(2008 年)からしばらくは、コンピュータの文字において激変期にあると言っても過言ではない。このような問題をふくめて、上記の問題について、いささかの論点の整理をこころみる。

【02】デジタル・アーカイブの文字

この研究会(CH-79)が、アーカイブ特集(第 2 回目)となっているように、現在の日本社会において、「デジタル・アーカイブ」の構築は緊急の課題である。それには、以下のようないくつかの問題点がある。

(1).「アーカイブ」におけるデジタルの意味

(2).「アーカイブ」における文字の役割

02-1：デジタルアーカイブの保存性の問題

いわゆる「アーカイブ」「アーキビスト」の側の人たちが、「デジタル」(特にその保存性)にどのように取り組んでいるかという問題。

この問題については、現時点では、きわめて悲観的な印象をもたざるをえない。日本アーカイブズ学会を中心とする人たちが、文書のデジタル保存について、きわめて否定的な見解を持っていることは、関係者にはよく知られている。その否定的根拠の主な理由は、デジタル化資料の、保存の安定性にある。

具体的には、以下のような理由による。紙(中性紙)であれば、100 年以上の耐久性がある。しかし、デジタルのデータは、どうであるか。CD-R や DVD は、そのままで 100 年の耐久性があるであろうか。あるいは、ハードディスクを 100 年間にわたって連続的に稼働させることは可能であろうか。これらは、現在のコンピュータ技術においては、否、であろう。

しかし、この問題点については、次のように反論できる。デジタル化データは、メディア変換とコピーを繰り返すしか保存が保証できない。そうである以上、これを認めた上で、そのコスト(人的・資金的)を、アーカイブの運用・保存に、組み込んで考えればよいのである、と。

したがって、この論点については、現時点で見解の相違はあっても、いずれ、解決する可能性がある。

また、現在、デジタル環境で発生し流通する文書・記録がある。例えば、国や地方自治体の HP など、である。これは、公文書、と認定すべきと筆者は考える。このようなものは、もはや、デジタルでしか保存し得ないものである。さらに言えば、アーカイブの基本原則である、文書の原秩序維持は、デジタル文書(HP など)において可能であるかどうか、再検討の余地がある。

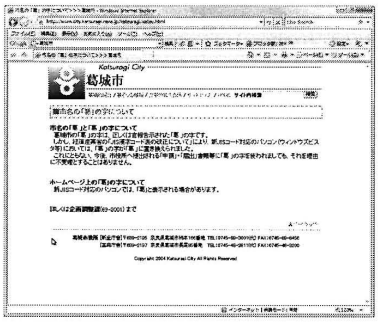

図-01　葛城市

02-2：デジタル公文書というべきものの存在

「デジタル公文書」と称さなければならないものが現実に存在している。「葛城市」「葛飾区」の HP は、インターネットでの閲覧を前提に作成されている。このなかには、市・区の正式名称としての文字についての言及がある。また、「岡崎市」のポルトガル語の HP は、現在、この市における人々と行政のあり方(ブラジルからの労働者の存在)を端的にしめす公的な記録と考える。現実に、どの程度のポルトガル語話者が居住したいたかのデータと別に、ポルトガル語の HP を行政として用意しておく必然性があったということ、およびその内容の記録として、デジタル公文書としてアーカイブの必要がある。

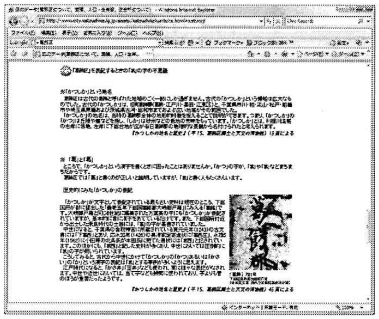

図-02　葛飾区

図-03　岡崎市

02-3：デジタル化文書には 2 種類ある

以上、デジタル化文書について、その保存を中心に考えてみた。しかし、さらに考えると、文書のデジタル化には、2 種類あることに気づく。

第一は、既存の文書の画像データ化である。この代表的なものとしては、国立公文書館のデジタルアーカイブがある。国会図書館の、近代デジタルライブラリーもある。

これは、いくら文書作成がコンピュータ(ワープロ)に依存するようになっても、残ることである。少なくとも、文書に、印鑑(あるいはサイン)が必要とされる限りは、最終的な文書は、紙の形態で残ることになる。そのデジタル画像化と保存、ということになる。

第二は、印鑑やサインが必要とされない文書であるならば、ワープロ文書データのままで残すということが可能になる。あるいは、PDF 化である。

この場合には、使用の文字コード、あるいは、使用のフォントが、問題となる。例えば、今、筆者は、MS 明朝で、この原稿を書いている。そして、プリントアウトし、論集のカメラレディ原稿となる。だが、これを、別のフォントで表示・印字することは、いとも簡単である。つまり、文書の文字については、フォントを指定しない限り、同じ見え方を保証することができない。

02-4：デジタル化文書における文字

コンピュータ環境による文字の見え方が違う、このようなことは、ワープロ文書に限ったことではない。今の社会で日常的に必須となっている電子メールでも、その表示フォントの設定によって、見え方は異なる。等幅の MS ゴシックで表示するか、プロポーショナルの MSP ゴシックで表示するかで、文書レイアウトとしての、見え方は同じではない。

さらに、文字コードが異なれば、同じ文字(字体)が見えない、ということが起こる。

筆者の研究対象とした文字として、「祇」「葛」がある。

・「祇」は、0208 では「ネ氏」、0213:04 では「示氏」となる。

・「葛」は、0208 では「ヒ」であり、0213:04 では「人」である。

これらの漢字は、地名としてよく使う。「祇」は、京都の祇園が有名である。しかし、地名としては、祇園信仰の全国への広がりにともなって、全国各地に「祇園」の地名・駅名・学校名など、がある。「葛」は、奈良県葛城市では「ヒ」を公式な文字とし、東京都葛飾区では「人」を公式な文字としている。それぞれの HP には、地名文字についての解説がある。HP などは、アーカイブの対象ではない、などと言うことはできない。これらの HP で伝えている内容は、この市や区が、どのような漢字(字体)で書くべきか、その根拠と判断を示した、公的な文書であると、認定すべきである。

では、これらの漢字をふくむ地方自治体の文書(公文書)は、どのように見えればよいのであろうか。あるいは、将来、どのように見えるのであろうか。ここで、デジタルアーカイブの文字が、課題となる。

ここで、だからデジタルのアーカイブは信用できない、という理屈はなりたたない。アーカイブの基本理念が、現在の文書を記録として未来に残す、ということであるならば、今、現実にわれわれがパソコンのディスプレイで見ている、その文書・HP を、将来に残して保存するということを、根本から否定することになるからである。文字情報として同じ文書であればよい、というものではないはずである。人間が、文書を書くとき、手書きであれ、ワープロであれ、なにがしかの価値判断のもとに、文字を選んでいる。複数の字体(異体字)があれば、そのうちどれを選んで文書を書いたか、このことも、また、文書に内在する歴史資料としての価値である。

「葛(ヒ)」か「葛(人)」か、そのどちらを選んでいるのか、それを、HP でどのように説明しているのか、また、それを見る人は、どのように見ているのか、これは、文書の文字についてアーカイブすべき事項である。

以上を総合して述べるならば、アーカイブ(デジタル)においては、その表示の文字(字体・字形)をもふくんだものでなければならない。

【03】文字のアーカイブ

次に、逆の方向から考えてみる。デジタルの環境における文字のアーカイブである。

(1).今、われわれが使っている文字を、アーカイブとして、将来に残す必要はないのか。

(2).文字を、記録し・残す必要があるとするならば、それは、どのような手段によって可能であるのか。

デジタル化文書は、それを「テキスト(画像ではなく)」として残す場合、それに使用した、「文字」とともに保存しなければならない。でなければ、将来にわたって、同じ文書を、同じように見える、ということの保証が得られない。

文字について、なにがしかの付随データを保存するとき、次のような属性をあつかうことになる。(1)コード系　(2)文字セット　(3)フォント　(4)グリフ

以下、簡単に私見を述べる。

(1).コード系：現在のコンピュータによる文書データであるならば、具体的には、JIS コード、シフト JIS コードか、ユニコードか、ということになる。多くの場合、コンピュータの文字については、このレベルで語られることが多い。

(2).文字セット：冒頭に述べた通り、筆者は、この原稿を、「JIS X 0213:04」の環境で書いている(日本語入力は、ATOK2008)。日常的に書く文書や、論文(文字についてのものが多い)、あるいは、電子メール、ブログの文章、など同様である。このとき、筆者の方針として、基本的に、JIS 第 1･2 水準の範囲内で書くようにしている。第 3･4 水準の文字は、原則使わない方針でいるし、ユニコード(Ext.A)も、使用しない。どうしても使用するときは、その旨を、明記した上でメッセージを書く。(この文書では、後述の「鷗」だけが例外的に、第 3 水準文字になる。)

つまり、Windows Vista(Ultimate)で、Word 2007 で書く環境では、JIS の第 3･4 水準、さらには、Ext.A まで、確実にあつかえる。しかし、潜在的にその可能性があるということと、現実に、どの範囲の文字を使用するかは、別次元のことがらである。これは、文字コードにかかわる範囲での「文字セット」についてのことである。

だが、文字コードにかかわらない「文字セット」もある。教育漢字・常用漢字・人名漢字、などである。もし、公文書として、「常用漢字」の範囲内で書く(一部、固有名詞を除いて)という方針であるならば、文書について、その旨の属性情報が必要になってくる。もし、常用漢字に厳しく限定するならば、「イタリア」を「伊」とは略記できない。「伊」は常用漢字外である。しかし、第 1 水準漢字である。

このようなことは、文書のアーカイブにおいて、その属性情報として、記載し記録に残す必要はないのであろうか。どのような「文字セット」の範囲で書いた文書であるか、ということは、その文書内のことば(語)の表記に影響する。

(3).フォント：筆者は、この文書を、MS 明朝・MS ゴシック･Century・メイリオ、で書いている。だが、この文書データの文字列を、エディタにコピーすれば、このようなフォントにかんする情報は、消えて無くなってしまう。単なる、コード化された文字の連続でしかない。この文書(Word 2007)を、デジタルのままで保存して、はたして将来、同じように見える保証があるだろうか。それを確実にするためには、フォント全体を保存しておかなければ不可能である。

(4).グリフ：さらに細かなことをいえば、MS 明朝であるからといって、同じではない。「0208」と「0213:04」とでは、「祇」「葛」「辻」などの字は、違って見える。さらに、別のフォントに変えれば、文字の見え方(デザイン)は異なる。最終的には、個々の文字ごとに、どのような見え方であるのか、グリフのレベルでの情報保存が必要になる。

【04】紙の文書とデジタル文書

以上に指摘したことがらを考えるとき、文書の保存において、「紙」というものの優位性が見えてくる。「紙」は、それ自身で、デジタル文書における、ハードウェアとソフトウェアのほとんど全部の機能をふくんでいる。CPU であり、メモリであり、記録媒体(ハードディスクや CD-R など)であり、ディスプレイでもある。だから、「紙」のものだけを残せばよい、というわけにはいかないと、筆者は考える。それは、以下の理由による。

(1).はじめからデジタルでしか発生しない文書というものがある。これは、デジタルで保存するのが妥当である。この点については、アーカイブの立場からは、残すべきデジタル文書を選んで、それをプリントアウトして、ということも考えられないではない。しかし、この場合でも、プリントアウトする際の、レイアウト(ページ設定)や文字の問題は、必然的に発生する。

(2).「紙」の文書の保存と利活用のためのデジタル化の必要。通常のアーカイブの概念では、その対象となる文書は、それ一点きりである。したがって、その保存においては、可能な限り慎重でなければならない。しかし、その利活用(閲覧)は、文書を傷めることでもある。そのためには、「紙」の原本の保存と同時に、デジタル閲覧も、十分に考慮にいれるべきである。

(3).「紙」のまま保存し、利活用に供するとしても、そのためには、検索のためのメタデータが必要になる。少なくとも現時点において、これはコンピュータに依存することになる。もはや、紙カードの時代ではない。この種のデータを構築し、かつ、公開することによって、それぞれのアーカイブ機関(資料館・文書館、さらには図書館や博物館など)との、横断的な検索、資料の総合的な利用が可能になる。

すくなくとも、利用者(一般市民、研究者)の立場にたって考えてみるならば、ある文書が、行政機関などの情報公開によって見ることができるのか、アーカイブとして公開されているのかは、関係ないであろう。また、歴史的文書であれば、アーカイブ(資料館・文書館)と、図書館・博物館の違いはない。たまたま、どの機関・組織の所有であるのか、歴史的経緯の結果でしかない。

このような利用者の立場にたって考えてみるならば、検索データのデジタル化とその横断検索は必須であり、この点では、いかに「紙」であることを尊重するにしても、デジタル化から逃れることはできない。

【05】検索のための文字

そして、検索データのデジタル化にあたって、どのような「文字」で記載するのか、重要な課題となる。

その端的な例として、国立公文書館で「森鴎外」を検索してみる。結論からいえば、「森鴎外」では2件の文書が検索できる。しかし、「森鷗外」では、何も出ない。

「鴎」:いわゆる「拡張新字体」、第1水準にある。83JIS 漢字体。

「鷗」:いわゆる「正字体・旧字体」、第3水準にある。78JIS との互換性のために追加。

また、第2水準までの漢字であっても、「熙」「熙」「熙」の区別はかなりやっかいである。「近衛忠熙」「近衛忠熙」「近衛忠熙」それぞれに、検索結果が異なる。

【06】文字とアーカイブ

アーカイブにおいて、文字は重要である。デジタル文書(ワープロの文書ファイル)を保存する場合にはもちろんのこと、仮にアーカイブが、デジタル文書を対象としないとしても、その利活用における検索データとしても、見逃すことができない問題点をはらんでいる。

ここでアーカイブされた文書の利活用のための検索データに焦点を絞ったとしても、次のような課題がある。この場合、一番重要なのは「文字セット」である。

(1).どの文字セットで記述したか。現行の常用漢字では、公文書アーカイブの検索データの作成は不可能である。なぜなら、現行の常用漢字では、都道府県名ですら、書けない。岡山・熊本・

大阪・栃木・山梨、などがそうである。都道府県名の文字を排除して、国、および、これら府県にかかわる公文書が、あつかえるはずがない。

(2).では、それを拡張する範囲として、どの範囲までを許容したか。第 2 水準までか、第 4 水準までか。

(3).いずれの「文字セット」内であっても、その内部に、各種の異体字・新旧字体の組み合わせが複雑に存在する。これらは、統合するのか、あるいは、検索段階で、文字シソーラスを利用することにするのか。その場合、文字シソーラスの内部は、どのようであるのか。公開されなければならない。

(4).そして、アーカイブにとって重要なことは、そのアーカイブ自体が、どのように形成されたのかのプロセスもまた、アーカイブの対象である、という視点である。そのためには、文書・資料の整理・検索のために、どのような文字セットを、どのように利用したのか、そのことも、アーカイブにとって必須の事項であると言わざるを得ない。

【07】文字のアーカイブのためになすべきこと

現在のコンピュータ社会、また、アーカイブの利活用のためには、コンピュータにおける文字というのもが、いかに重要であるかは述べたとおりである。では、その文字を、どのようにアーカイブすることができるであろうか。

(1).規格票そのものの保存。日本に限定してであるが、コンピュータの文字は、JIS(日本工業規格)によって、決められている。そして、それは、原則的に 5 年ごとに更新されることになっている。この規格票の保存が急務である。5 年で更新ということは、5 年以上経過して、新しい規格が作成されれば、用済みで廃棄されることになる。そして、この規格票は、意外なことに、ほとんど保存されていないことが判明している。

(2).フォントデータの保存。しかるき機関・組織が、フォントデータを残さなければならない。現在、一般社会では、Windows XP が、いまだに多く使用されている。XP では、MS 明朝といっても、規格としては「0208」によっている。XP(0208)と、Vista(0213:04)では、見える文字の字体・字種が異なることは既に述べた。デジタルで残せるものであるならば、可能な限り、文書データと同様に、フォントデータも、記録・保存の対象として、策を講じるべきである。

【08】今後の課題：(新)常用漢字表のおよぼす影響

現在、現行の「常用漢字表」にかわる新しい、「(新)常用漢字表」が、審議中である。2010 年をめどに、審議が進められている。「常用漢字」の改訂は、かつての「当用漢字」から現行「常用漢字」への流れをうけて、時代の変化として、当然といえるかもしれない。

しかし、アーカイブの視点から見たとき、これは重要な問題でもある。

(1).公文書の類は、常用漢字に準拠する。ただし、現在の常用漢字では、都道府県名さえ書けない。。「(新)常用漢字」は、およそ 200 字ほどの追加になると予想される。

(2).かりに「(新)常用漢字表」が決定されれば、公文書(アーカイブの対象の基本)は、それで書くことになる。しかし、現在の JIS 規格(0213:04)では、「字種」としては対応できても「字体」としては対応できない。現行「常用漢字」との字体の整合性(いわゆる新旧字体)が問題になる。

(3).その結果、JIS 規格の再改訂が余儀なくされる可能性がある。ただ、現行の JIS 規格を温存し

たまま、実装フォントレベルで解決するということも不可能ではない。しかし、この場合でも、規格票の記述は、大幅な見直しが必要になる。

(4).「(新)常用漢字表」、JIS 規格票、フォントの実装、これらには、タイムラグが生じる。現実に、現在の JIS 規格(0213:04)が制定(2004 年)されてから、その実装(Vista の発売、2007 年 1 月 30 日)には、時間差がある。

(5).では、この間の公文書の文字、また、それによるデジタル文書、そして、利活用のための検索データで使用の文字、これらにおいて、どのような対応が必要になるであろうか。

このように問題点を指摘することはできるが、その解答が筆者に用意されているわけではない。だが、このような「文字」という視点から「アーカイブ」を考えることも、今後の重要な課題である、ということだけは確認しておきたい。

【09】アーカイブから学ぶべきこと

これまで、この CH 研究会をはじめとする、各種の人文学とコンピュータにかかわる研究会などでは、さほど意識することなく「デジタル・アーカイブ」の用語を使用してきたように、筆者にには思われる。だが、現在においては、「アーカイブ」の理念から学ぶべきこと（特に未来への責任という倫理観）があると感ずる。また、今の、デジタル技術で何が寄与できるか、さらには、現在のデジタル技術で実現している各種の事象を、どのように「アーカイブ」として残すべきか、本格的に考えねばならない段階にさしかかっていると感ずる次第である。そのなかで「文字」をどのように将来に残すことができるのか、あるいは、残すべきか否か、新たな課題である。

参考文献・ＨＰ

青山英幸.『電子環境におけるアーカイブズとレコード』. 岩田書院. 2005

大濱徹也.『アーカイブズへの眼』. 刀水書房. 2007

小川千代子・高橋実・大西愛(編著).『アーカイブ事典』. 大阪大学出版会. 2003

小川千代子.『電子記録のアーカイビング』. 日外アソシエーツ. 2003

小川千代子ほか(編著).『アーカイブを学ぶ』. 岩田書院. 2007

後藤忠彦(監修)・谷口知司(編著).『デジタル・アーキビスト概論』. 日本文教出版. 2006

三上喜貴.『文字符号の歴史　アジア編』. 共立出版. 2002

安岡孝一・安岡素子.『文字符号の歴史　欧米と日本編』. 共立出版. 2006

野村雅昭.『漢字の未来　新版』. 三元社. 2008.（※旧版『漢字の未来』は、筑摩書房. 1988）

師茂樹.「「デジタルアーカイブ」とはどのような行為なのか」.『CH-66』. 情報処理学会. 2005

永崎宣研.「デジタルアーカイブの弁証法」.『CH-68』. 情報処理学会. 2005

當山日出夫.「京都における「葛」と「祇」の使用実例と「JIS X 0213:2004」」.『CH-70』. 2006

奈良県葛城市　市名の「葛」の字について
http://www.city.katsuragi.nara.jp/katsuragi/katsu.html

東京都葛飾区　「葛飾区」を表記するときの「葛」の字の不思議
http://www.city.katsushika.lg.jp/aisatu/katsushikakunituite.html#katunoji

愛知県岡崎市　ポルトガル語の HP
http://www.city.okazaki.aichi.jp/yakusho/ka2650/tagengo/Home_p.htm