文字を残すための序論的考察

The Character as the Cultural-Heritage

當山　日出夫*

Resume:
デジタルドキュメンテーションにおいては，コンピュータをつかう。その検索メタデータや解説にかかわる部分については，「文字」に強く依存する。しかし，このコンピュータの文字は，紙に書かれた墨筆のような安定性が無い。むしろ，規格の変更，コンピュータへの実装において，常に不安定な状況にある。そして，これは，現在の問題であると同時に，未来への文化資産の継承に大きな問題点ともなる。本発表では，この問題点を指摘する。

1．はじめに

本稿は，あえて，いわゆる JIS 規格「JIS X 0208」の範囲内の文字で「記述」してある。厳密には，0208 規格，0213 規格，それぞれに，さらに歴史的経緯がある。本稿では，〈0208〉〈0213:04〉の名称を使うことにする。本稿執筆の環境は，Windows Vista で，〈0213:04〉で「ディスプレイ表示」を見ながら書いている。しかし，この原稿が，JADS においてどのように処理され現実の研究会の「プリントされた予稿」となるのか，〈0208〉で「印字」なのか，それとも〈0213:04〉で「印字」なのか，筆者には関知し得ない。本発表は，このような，身近にありながらも意外と見過ごされている，「文字の同一性の保証」という点について考えてみたい。

2．デジタルアーカイブ

近年，急速に使われるようになってきたことばとして「デジタルアーカイブ」がある。このことばについては，いわゆる「アーキビスト」（アーカイブズに従事するひとたち）のほか，人文学（特に，人文情報学），また，情報工学においても，多用されるようになっている。

本発表においては，デジタルアーカイブにおける，「文字」の問題をとりあげる。以下，とりあえず，「デジタルアーカイブ」とは，デジタルによる，種々の資料の保存，管理，と考えておく。デジタルライブラリ，デジタルミュージアムなどをふくんで，総括的に考える。

3．アーカイブと文字

二つの問題にわけて考える必要がある。

(1).テキスト(文字で書かれたもの)それ自体を，テキストとして，つまり，文字コード化することによって保存するもの。

たとえば，史料編纂所のデータベース，が代表であろう。また，日本の文学作品でも，多くの作品が，テキストデータベース化されている。国文学研究資料館の日本古典文学大系ＤＢ，など。また，インターネット上でも，古くは『万葉集』『源氏物語』から，近代の文学作品（「青空文庫」）にいたるまで，数多くの作品が，テキストデータとして，利用可能になっている。この種のデジタルアーカイブにおいて，「文字」が重要な意味をもつことは，言うまでもなかろう。

(2).画像データのアーカイブではどうであろうか。通常，問題になるのは，当該の画像データの解像度や，形式，色空間，などである。

しかし，これを，ドキュメンテーションの視点から見ると，「文字」の重要性が認識できるはずである。また，さまざまな検索メタデータも，基本的には，文字によって記される。つまり，デジタルアーカイブにおいて，その適切な保存，運用，管理のためには，「文字」は必須である。

通常，「文字」についての議論は，テキストが対象とされてきた。たとえば，『古事記』が，文字コード〈0208〉で，記述可能かどうか。作家の名称として，「森鴎外」の「鴎」の字を「区鳥」

*とうや　まひでお（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）
原稿受理日：2008/10/26

〈0208〉で書くか,「區鳥」〈0213:04〉で書くかなどの議論である。

だが,これに対して,「ドキュメンテーション」「キューレーション」における「文字」の問題は,あまり論じられてこなかったように思われる。以下,本発表では,この方向からの,「文字」の問題を考えてみることとする。

4.文字のいったい何が問題か

文字が問題であるといっても,さらに,いろんなレベルで考えることができる。
大きくは,次の２点の問題がある。
(1).文字コード系
(2).文字の字体

5.文字コード系の問題

日本国内の事情に限定すれば,現時点(2008年)で,社会で一般的に使用されているコンピュータで搭載の文字について,簡単に整理すると,
(1).旧 Windows XP で使用の, JIS X 0208　第一・第二水準の漢字。
(2).現行の Windows Vista, MAC OS X で使用の, JIS X 0213:2004　第三・第四水準までの漢字。
通常は,このどちらでエンコードしたかが問題となる。

だが,現実には,日本の規格(JIS)以外にも,ユニコード(Unicode)の文字が使える。いや,厳密には,今のコンピュータは基本的に内部処理としては,ユニコードで動いているのだが,ディスプレイ表示やプリントアウトについては,それぞれの規格(日本の場合は JIS)のフォントで表示している。

この範囲では,次のようなことが問題になる。第一に,古くから構築されてきたものでは,旧来の〈0208〉の範囲内で記述する。というよりも,そのようにせざるをえなかった。

この〈0208〉内部においても,字体の選択の問題がある。

森鴎外は,「区鳥」の字体でしか記述できない。「區鳥」は,ふくまれていない。どうしても,「區鳥」が使用したければ,いわゆる「外字」処理になり,インターネットでの公開データ共有になどにおいて,支障をきたす。

厳密にいえば,〈0208〉規格においても,制定当初は,「區鳥」であった(所謂 78JIS)。それが,

83 年の改訂において,「区鳥」に改められて,現在にいたっている。

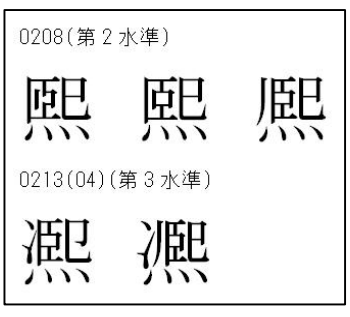
このような例,「区鳥」は,通常の漢和辞典などには掲載されていない。「カモメ」を書き表す漢字そのものが,当用漢字・常用漢字の対象外であったので,公的な規範が存在せず,だが,文字としては,多くの人々が使ってきた文字である。「區鳥」は,旧来の字体(康熙字典を規範とする)であり,それに対して「区鳥」は,拡張新字体と呼ばれる。

常用漢字内において,區→区　歐→欧　が,新旧字体としてある。それにならえば,「區鳥」が「区鳥」の字体で書かれることも,ある種の合理性がある。

では,文字コードの範囲を限定して,〈0208〉規格の範囲内で記述するという方針であったとしても,この内部に存在する異体字処理は,どうであるのか。

同様な例は,つまり,上述の「區」「区」のような,新旧両字体の組み合わせが,〈0208〉の内部においても存在する。たとえば,「音楽」の「楽」は,「音楽・音樂」,両用の表記が可能である。

また,いわゆる新旧字体の違い意外にも,「峰・峯」などの例(異体字)の組み合わせがある。「熙」の字は,〈0208〉においても,「熙」「熙」をふくめると３種類の異体字が存在する。また,「宝」についても「寶・寳」があり,この微細な字体の違いを,通常の表示(ワープロなど)で,区別することは,かなり困難である。このようなことは,〈0208〉規格内に,このような面倒な異体字がある,ということを知っているから,なんとか対応できる。



第二に，そこにさらに，2007 年 1 月に発売された，Windows Vista では，〈0213:04〉第三・第四水準の文字が追加され，およそ 10000 字以上の文字が，実装された。つまり，購入した時点での状態では，この文字を使うことになる。

これは，たしかに便利になった点もある。これまで，〈0208〉規格では存在しないので，やむをえず，何らかの特別の処理をしていた文字が，「共有」の文字として，コンピュータで使えるようになった，このことのメリットは大きい。

例えば，「區鳥」は，〈0213：04〉規格において，正式に登録された文字となった。

だが，その一方で，字が増えることは，メリットばかりではない。実は，この点が，あまり一般には認識されていないように，筆者には思われる。たしかに，コンピュータの多言語対応，より多くの文字が使用可能になることは，一般論としては，歓迎すべきことではある。だが，その一方で，ある種の混乱が，生じることもある。

(1).文字種が増えることに起因する混乱

文字種が増えることは，便利かもしれない。しかし，一方で，さらなる異体字の増加を意味する。「鴎」は，上記のとおりである。新しい〈0213:04〉環境で入力された文字データは，旧来の〈0208〉で入力された文字データと，不整合を生じる。たとえば，人名などでよく使う「崎」が「大」であるか「立」であるか。「大」は，〈0208〉の範囲内。「立」は，〈0213:04〉で追加。

(2).字体の変更が起こることに起因する混乱

「葛・祇・辻」などである。これは，文字をあつかう関係者によく知られている典型的事例である。

「葛」の文字は，従来の〈0208〉では，「ヒ」につくる。だが，最新の〈0213:04〉規格では，「人」につくる。そして，この事例は，次のような問題を生み出している。

・奈良県葛城市の場合，市の正式名称は，「ヒ」の「葛」。XP で標準。
・東京都葛飾区の場合，区の正式名称は，「人」の「葛」。Vista で標準。

現在，この「葛」については，少なくとも，日本の国内規格(JIS)としては，完全に二者択一で，選択の余地はない。

このようなことは，どう考えるべきであろうか。ちなみに，現在の新聞(『朝日新聞』)は，一般に使用の文字は，「人」の「葛」(葛藤など)。しかし，地名としての，葛城市は，「ヒ」の「葛」によっている。だが，これは，新聞という「紙」に印刷した場合であって，デジタル文字（文字コード，JIS 規格のレベル）では，包摂されてしまう。

6．ユニコードの問題

ユニコード(Unicode)，それ自体が，現在，発展(あるいは，拡張)の途上にあり，その規格の解釈や具体的な実装・運用において，多くの課題をかかえている。

そのなかで，日常的な例で，問題のある字として，「高」について，「くちだか」と「はしごだか」がある。この２字は，JIS 規格においては，包摂されており，〈0213:04〉においても，「くちだか」しか出ない。だが，「はしごだか」の字体は，ユニコード(U+9AD9)に「CJK 統合漢字」として入っている。

(1).ユニコード未対応のコンピュータでは表示できない。また，やや古いタイプのプリンタでは，プリント不可で「・」になってしまう。
(2).日本の規格〈0208〉〈0213:04〉では，包摂されており，同じ文字としてあつかう。
(3).XP･Vista いずれであっても，ユニコード(少なくとも，EXT.A)に対応していれば，表示可能である。

つまり，この「高(はしごだが)」の文字は，安定して見える・印字できることが，現時点では保証されていない。

7．コンピュータの規格以外の文字の制度

当用漢字(常用漢字)・人名漢字・教育漢字など，文字コードではない，文字セットの問題である。戦後，原則当用漢字(常用漢字)で公文書類，新聞などを，書くという方針であった。この影響として,「語い(=語彙)」「駐とん(=駐屯)」「世論(=輿論)」などの，いわゆる「交ぜ書き」「漢字の書き換え」などが生じた。

このような言語政策にかかわる表記のあり方は，単に，文字コードのレベルでは議論不可能である。

まず，その文章がどのような言語政策的「文字セット」によって記述したかの，メタデータが付属していなければならない。

しかし，にもかかわらず，当用漢字（常用漢字）は，固有名詞（人名・地名）を対象外としている。「語い（=語彙)」と書きながらも，人名では,「伊藤」，地名（都道府県名）では「岡山」「山梨」などは，漢字で書かざるをえない。

このような事象は，過去の文献の遡及入力において, 課題となる。原文どおりか, あるいは, 特定の文字コード〈0208〉〈0213:04〉の範囲内に拡張するのか。

8．今後の課題

以上のことを，共時的に考えるならば，コミュニケーションのトラブル, ということである。いわゆる「文字化け」の問題なども，その現象が起こることを知っていれば，相互になんとか対応可能である。だが，歴史的にみたときどうであろうか。「アーカイブ」として，過去の文字を残すことが可能であろうか。あるいは，今，我々がこのような混乱した文字環境(コンピュータによる)にいることを, 未来に伝えることが可能であろうか。今，コンピュータによって使用している文字を, 過去～現在～未来に向けて, 同じ文字を同じ文字として，継承可能であろうか。

さらに端的にいえば，過去に入力されたデータ，現在入力しているデータ，これらが文字によって記述されているならば，将来にわたって検索の可能性を保証できるだろうか。実際には, この件について，有効な解決策があるわけでない。しかし，このような問題を現在の我々がかかえているのだということを,「アーカイブズ」や「ドキュメンテーション」にかかわる人たちは意識しているであろうか。デジタル化されたデータにおいては,「検索不可能＝存在しない」，なのである。また，いわゆる「文字化け」の現象は，共時的にも発生するが，データが蓄積されていくにつれて，通時的な問題として重要になる。

9．まとめ

本学会「アート・ドキュメンテーション」あるいは,「デジタルアーカイブ」「デジタルライブラリ」が，過去の文化資産を未来に残すことをその活動にふくむとするならば，そのためには,「文字」を確実に残さなければならない。

コンピュータの文字は，確実に，時代の流れとともに，変化していく。その流れの動向をみすえながら，どのようにしたらデジタルの世界で「文字」が残せるのか，緊急の課題であると思う次第である。

参考文献

安岡孝一・安岡素子.『文字コードの世界』. 東京電機大学出版局. 1999

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

當山日出夫.「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）. 情報処理学会. 2008

當山日出夫.「文字とアーカイブについての問題提起」.『漢字文献情報処理研究』. 第 9 号. 好文出版. 2008