

デジタル Essense エッセイ

「ヤスクニ問題」

坂口 裕靖

総務省により平成23年度に行われた、「市区町村が利用する外字の実態調査」によれば、1,386の市区町村から合計1,166,536文字の「外字」が提出されたとのこと。外字だけで100万字越えてるといのが、文字のデータ化がいかに困難であるかを物語っているのではないのでしょうか。100万ですよ、100万。unicodeなんて、サロゲートペアを使ってもたったの111万字ちょい(1024x1024+65536-2048)しか収容できないです。もっとも、同調査で同定した結果、934,812文字は住基ネット統一文字14,852種類・戸籍統一文字6,709種類に同定でき、不明な漢字が52,294文字、不明なかな文字が97,791文字、記号とかが81,639文字と、合計231,724文字が本来の意味での「外字」であったとのこと。そもそも「文

字」と言ってるものが何を表しているのかは、結構深遠な問題をはらんで、よく考えるとぐらぐらしてくるような概念になってることに気づきます。実際今こうやって書いている原稿は「文字データ」の連なりであって、「文字」の連なりとは微妙に異なります。「文字」の連なりは「コトバ」に対応するはずですが、文字の字体・字形によりニュアンスが変わって来たり、字体そのものが何かを表している場合、意味が同じだからといって字体を変形すると、データとしての意味がなくなってしまう。その昔の8bitパソコンだと、そもそも漢字ROM自体がオプションだったりして、漢字を扱うだけでも一苦勞でした。漢字が扱えるといっても第一水準しかサポートしなかったりして、そうすると「キレイ」という意味で「綺麗」と書きたいのに、「綺

の文字が第二水準で定義されてて入ってないから、仕方なく「綺麗」と書くしかなかったり、という不自由さを存分に味わってました。まあ「綺」自体は常用漢字に含まれていないものの、人名漢字には含まれているためギリギリセーフ、「麗」は中学校で習う常用漢字なので使用はOK、というわけで「綺麗」がある意味正しいわけですが、この文字列が表すコトバは「綺麗」に対応するものの、文字面は全く異なり、斜め読み時の印象も違って来ます。「文字」のややこしさを最も表しているのは名前じゃないでしょうか。名前を構成する文字の並びには「コトバ」としての意味は(名付け親の想いはともかくとして)特になく、それぞれ万葉仮名から続いている、文字として認識される図像の並びと、それに対応する音の並びが、そのモノヤコトとの対応関係を

ワンポイント パスワード マイゲート

埼玉に住んでると、結構色々なところで引き落としのために埼玉りそな銀行の口座を要求してくることが多く、先日仕方なく口座を開きました。その際、もちろんヤスクニ問題のチェックを行ったわけですが、結果は残念ながらバツ。「裕靖」で口座開設書類を申請することはできるが、システムがこの漢字を扱うことができないため、印刷物等は「裕ヤス」という表記になる。そのかわり、自筆の名前については「裕靖」を使っても●構わない●。一方、「裕靖」で口座開設書類を申請してもらえば、システム上で扱える漢字であるため、印刷物も問題なく出て来る。ただし、今後自筆の名前については「裕靖」でないと●受け付けない●、とのことでした。ここで公的証明書の表記はすべて「裕靖」だが、口座の「裕靖」との突合に支障を来すのではないかと確認したところ、問題ないということ。つまり、「裕靖」で入力されたレコードと手書き文字列との突合は「裕靖」のグリフのみで行い、「裕

靖」は拒絶する一方、公文書における「裕靖」は包摂基準により「裕靖」として解釈し、受容するという。なんじゃその二枚舌。しかも備考等に「裕靖」と記載することはできないんだって。これ、「上司に確認」して頂いた結果ですから、同行の正式な回答ということでした。早く銀行のシステムに unicode やら戸籍統一文字やらが使えるようになって欲しいものです。実際公文書で使えるのだから!! ところで同行のインターネットバンキングサービス「マイゲート」があるのですが、その初回アクセス時のパスワードが「口座番号+(キャッシュカードの)暗証番号」という数字列だということで、心底驚きました。一応そのページはEV SSLで保護されてはいるようですが、暗証番号をブラウザが保存する危険を犯すべき理由がわかりません。今時ランダムな番号生成するなんてカンタンですが... 説明コストの方が高価なんじゃろか。

持ちます。そうすると、字体の違いがあった場合に、同じものを指すかどうかは非常に微妙な問題をはらむこととなります。ここいらへんが最も顕著に現れるのが公的文書でしょう。公的文書内で固有名詞や地名などに言及する必要がある場合、何を指して何を指しているかの問題は大変シビアになってきます。ある文字列が指し示す何かを確認できる相手がいれば良いのですが、常に必ず正しい返答を返してくれるようなメカニズムはこの世のどこにもありません。となると、文章に記載された図像としての文字の並びを持って、対応するものであると定義していくしかないはず。人間生まれてから死ぬまでの間、色んな所に移動する可能性があります。その移動の履歴は戸籍とか住民票とかで追っていくことになるのですが、(少なくとも従来では)各個人にグローバルユニークなIDが割り振られているわけではありません。となると、レコードの突合を行うためには、複数の属性(住所、氏名、年齢、性別等々)が等しいことを持って同一性を仮定するしか無いでしょう。この時、由来が記載されずにデータが変更されていたら、きちんと突合できるでしょうか? まあ、できないから年金問題が発生するわけですが。さて、パソコン以前の環境では、「漢字は面倒だから極力少なく」という風潮だったわけです。

いっその事漢字やめちまえ、という勢力もいたようです。それがパソコンが普及し、処理能力が問題なくなった現在では、「どうせ処理能力は充分あるので、無理に同一視することでデータロスを生むぐらいなら、細かな違いをデータで弁別できるようにしたい」方向へと変わりつつあります。包摂基準により字形の違いとして区別できるような文字については、「異体字セレクタ」という仕組みで unicode に取り込まれつつあります。一方で、既存の漢字コード体系との互換性を考えると、同じ字形に複数のコードが割り振られていた場合、あるいは元のコードでは区別されていたものが

unicode では包摂されてしまう場合、相互変換が不可能になってしまいます。このため、相互変換を目的として CJK 互換漢字が設けられるようになりました。で、筆者の名前である「裕靖」の「靖」の字(「靖」の「青」のうち、月の部分が「円」になってる字形)が unicode に CJK 互換漢字の U+FA1C として収録されてるのを、ついこないだ発見しました(笑)。これは同時に「靖」に異体字セレクタ U+FE00 を組合せたものもあり、戸籍統一文字の 482540 でもあります。元は IBM 拡張漢字の一部だったようです。ピバ IBM。今までは JIS 第一水準の人名漢字たる「靖」を仕方なく使ってきたわけですが、unicode にあるのであれば、もはや積極的に使わない理由はありません。そもそも、「靖」の部首は「立」じゃなくて、(下が円になってる)「青」なのです。まあこうやって部首を表現できるのも unicode のおかげですので、足を向いて寝られませんか。というわけで、名前を入力可能な局面で、極力「裕靖」を使ってみようキャンペーンを個人的に絶賛強力推進中。「裕靖」のヤスは旧字体の「靖國神社」のヤスですが、この文字がちゃんと扱えるかどうかのことを「ヤスクニ問題」と呼んでいる訳です。そもそも、入力として unicode を許容しない場合、当然ながらヤスクニ問題は失敗します。失敗の仕方も様々で、何もなかったかのように文字化けして受取り、レコードがボロボロになるシステムがあるかと思えば、バリデーションの過程で使えない文字であることを教えてくれるシステムもあります。あるいは、システムとしては受容してちゃんとレコードに記録されているにもかかわらず、確認のために送信

しているメールのエンコーディングが iso-2022-jp であって、ヤスがトウフやゲタになって表示される場合もあれば、メールは utf8 なエンコーディングなのでちゃんと表示されるかわりに、郵便物や印刷出力が化けるものもあります。賢いシステムだと自動的に包摂基準に従って「靖」に変換して出力してくれるものもあり、なんかこう今現在、文字の扱いがいかにも混乱しているかの一端を垣間見る感じです。日本語の場合同音異義語が大量にあるため、コンテキストを限定しないと音だけで伝えることは結構難しい問題です。書き文字の場合はこの区別をするために当てる文字を変えて誤解を少なくするわけですが、そうやって文字を前提とした固有名詞等の区別がある以上、外字 100 万字な状況を解決することは一筋縄では行かないでしょう。筆者はたまたま戸籍が電子化される前に生まれていたからこういう状況を楽しめる訳ですが、使える漢字なのに使わせてもらえない世代はちょっとかわいそうかも。... あ、だからキラキラネームが幅をきかせてるのか。あれは平成仮名というべき表記法なのかも知れませんが。

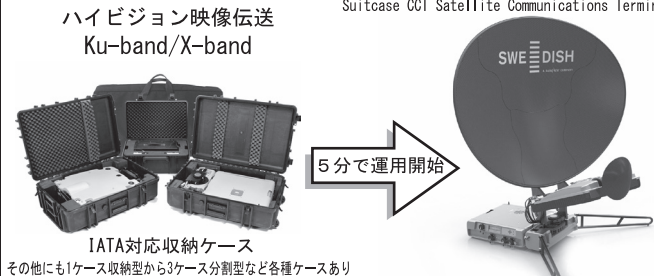
Hiroyasu Sakaguchi
(株) IMAGICA イメージワークス

SWE DISH

緊急報道
ハイビジョン映像伝送
Ku-band/X-band

CCTスーツケース 90cmφ型 2タイプ有り
120cmφ型

衛星通信超小型可搬アンテナ
Suitcase CCT Satellite Communications Terminal



5分で運用開始

IATA対応収納ケース
その他にも1ケース収納型から3ケース分割型など各種ケースあり

エーティコミュニケーションズ株式会社
http://www.bizsat.jp TEL : 03-5772-9125

