

「京都と広島と幻の犬」

坂口 裕靖

「ちわーっす」「お、久しぶり。今日はちょっと涼しいね」「またそういうこと言うー。ここの気温、今 36 度ですよ」「あー ... 40 度な気温とか、50 度超えの車内とかに慣れてたから、なんか涼しい気になってたわ ...」「猛暑日相当ですよ。暑いとしか言えないじゃないですか」「そうだねー。すでに立秋過ぎてるけど、秋はまだ先かな」「これからが本番でしょう。今年は早魓っぽいですし、またコメが大変かもしれませんね ...」「ちょっと先が思いやられるね ... とところで、最近何やってんの?」「検索を、ちょっと。」「検索? そんなの google 様で十分じゃないの?」「まあそうなんですけどね ... 世の中には google 様に公開できない、公開したくない種類の情報ってのもあるわけで、そうすると閉じた検索環境が必要なんです」「ふーん。まあ分かるけど、買ってくればいいんじゃないの? 買ってこなくてもさ、そういうば mysql とか、全文検索機能あったじゃない?」「まあそうなんですけど ... つまんないじゃないですかあ」「つまんないっ

て ... あれじゃないの、実績あるやつ使った方がいいんじゃないの?」「確かに実績はありますが、オーバースペックだったり、インターフェースがややこしかったりすると、原理からやった方が色々手軽ですし、第一設定して終わりじゃ、ヤリガイないじゃないですか」「やりがい ... まあちゃんとできるんだったらいいけど、自分でこねくって動かないってのは勘弁してほしいかな」「わかります。まあ最悪全文 grep すればいいので、あんまり気にしてないです」「そうなの? ... まあ、ウチの案件じゃないならいいけどさ」「そこは大丈夫。御社のです」「そうなの!? まあ、私の担当じゃなさそうだからまあ ... でもさ、検索ってそれなりにノウハウが必要なんじゃないの?」「まあ言い出すとキリがないのは確かですね」「そうでしょう? だったらやっぱり、それなりに使われているシステムの方が、お願いする方としては安心かな」「まあ勘所は時代とともに変化して行きますからね。

『幻の犬問題』って聞いたことがあります?」

「... いや、聞いたことないけど」「でしょうね。ワタシが勝手に呼んでるだけですから」「あら ... で、それ何?」「これはですね、最近のシステムだとほとんど気にする必要ないんですけど、昔のデータをいじる場合は注意する必要があるんですよ」「... ならまあ、関係ないか」「まあまあ、そうおっしゃらず。これはですね、euc-jp、もしくは jis なデータを扱うときに発生するんですよ」「イーユー ... なんだっけそれ? ヨーロッパ関係?」「えー、最近だと end user computing の略語として使われることが多いですが、今問題にしているのは Extended Unix Code の方ですね。JIS の漢字を複数バイトにエンコードしたやつです」「... 聞いたことないな」「最近ではみんな utf8 ですからね。その昔、漢字をコンピュータで扱うには大きく 3 種類のやり方があって、一つはいわゆる JIS コード、次に Microsoft 由来の Shift-JIS、あとはこの euc でした。その後 unicode が勃興して」「勃興 ...」「ええ、もはや人類は unicode 以外をサポートすることをやめ

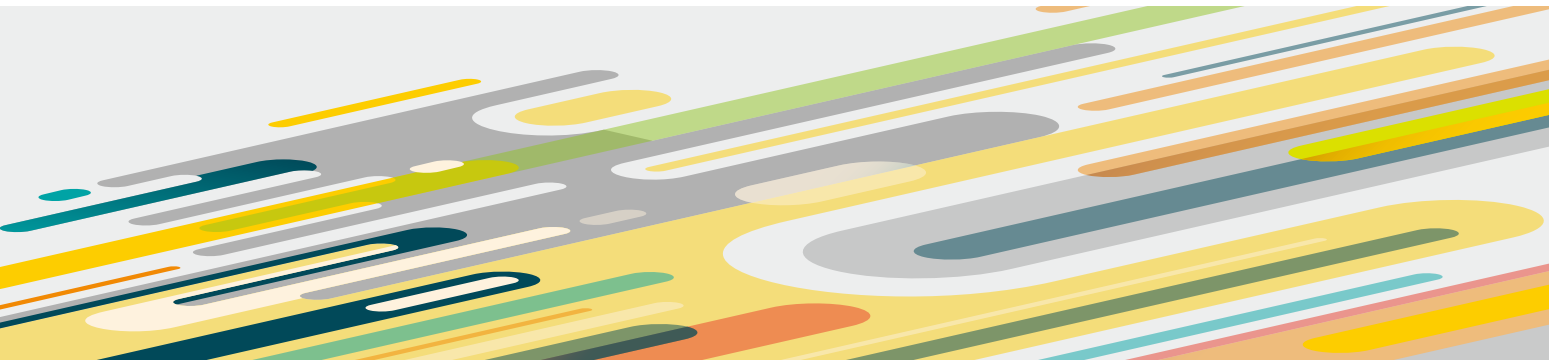
One Point BUZZ WORD

赤いの

なんか最近、タオルが微妙に茶色くなってる気がしたんですよ。アレかな、洗いが悪いのかなとか思って、洗剤の量を増やしたりしてみたのですが改善せず。そしたらアレかな、洗濯槽の裏のカビとかが原因かなとか思って、漂白剤で色々やってみたのですが、一向に取れる気配がありません。仕方ないので、近くのウェルシアで洗濯槽クリーナー買ってきて、試しに回してみたんですよ。そうしたら、なんか明らかに鉄錆色の赤茶色い不気味な水が出るわ出るわ ... これ明らかに鉄サビでしたわ。つまり、選択プロセスのどこかで鉄錆が洗濯物に付着して、茶色い汚れになってたということらしいです。気になったのでステンレスの洗濯槽をちょっと拭いてみたところ、赤錆色の汚れが付着します。どうやらタオルが茶色くなったのは、表面に微細な粉末状に鉄錆が付着

したステンレス洗濯槽で乾燥したからだったようです。そこで洗濯槽の表面を丹念に拭いたところ、いくらこすっても赤錆色の汚れがつかなくなりました。とりあえずこれ以上茶色くなることは防げたようです。一方で気になるのは、この鉄錆がどこで発生してるかですね。お前、ステンレス洗濯槽ちゃうんか、と。ステンレスなのは見た目だけで、裏は絶賛鉄錆だらけなわけ?

まあそもそも耐用限界をこえてぶん回しているのは確かで、最近洗濯槽が回転するときにギシギシ音がしてて、ベアリングかなにかがいかれてる雰囲気濃厚ではあるので、鉄錆ごときでオロオロすることでもないのかもしれませんが ... この洗濯機、とにかく乾燥性能が低く、何をどうやっても乾かないという問題があったわけですが、今度は乾かないだけでなく、汚しやがるという問題が発生しているわけです。そろそろ買い替え時なのかもしれませんね。同じ機種はコリコリですが、スケールファクター的に選択肢が ...



てしまいましたので、事実上デジタルな文字コードは unicode 一強になりました。その影響で、ほとんどの文字エンコーディングは現在 utf8 になってます「そうなんだ。あー、だから聞いたことないんだね」「そうですね、まあここ 10 年ぐらいは utf8 なんじゃないでしょうか」「一昔前の話ってことね」「そもそも、それまで 1 文字あたり 2 バイトですんでたのが、utf8 にすると 3 バイトに増えます。にもかかわらず、圧倒的に utf8 が支持された原因の一つが、この『幻の犬問題』だと思ってます」「ふーん ... で、それなに?」「えーとですね、そもそも漢字っていうのは日本工業規格 ... いや、今は日本産業規格ですか。で規定されてて、そこでは『区点コード』で定義されてます。例えば『幻』は 24 区 24 点、『犬』は 24 区 4 点ですね」「ふーん ...」「句点コードはそれぞれ 1 ~ 94 まであるんですけど、JIS コードはそれを 16 進数の a1 から fe にそれぞれ当てはめます。別の言い方をすれば、10 進数で 32 を足します。そうすると、『幻』は 0xb8b8、『犬』は 0xb8a4 となります。ちなみに『の』は 0xa4ce です」「ほお ...」「euc の漢字は、JIS の漢字コードをそのまま使います。なので、euc だとそれぞれ『幻』は 0xb8b8、『の』は 0xa4ce、『犬』は 0xb8a4 となります」

「... なんとなく見えてきたぞ」「そうなんです。『幻の』というデータは 0xb8b8a4ce という並びになって、これ頭から 2 バイト目と 3 バイト目をひろって『0xb8a4』になります。『幻の』の中に『犬』として読めるビットパターンが含まれてるわけです」「ふーん ... それがどう影響するの?」「つまりですね、文字の境界を意識しないでビットパターンを検索すると、とんでもないデータにマッチしてしまうという問題が発生するんです。これを私は『幻の犬問題』と名付けました」「なるほどね ... でもさ、それがなんで utf8 だと大丈夫な

の?」「utf8 の場合、『シングルバイト文字』と『マルチバイト文字の最初の文字』と『マルチバイト文字の 2 バイト目以降』がそれぞれ区別できるんですよ。さらに、マルチバイト文字は最大 4 バイトにしかありません。だから、検索したいビットパターンが文字境界を超えてたまたまマッチしてしまうという、『幻の犬問題』が発生しません」「なるほどね ...」「まあマルチバイト文字列使わないカルチャの連中にしてみれば、8bit clean にするだけでマルチバイト文字に対応できるのが便利で、utf8 が支持されたんじゃないですかね。日本語圏でも、データ量が 1.5 倍になるデメリットはあまり問題にならなかったと思います」「なるほどね ... でもそうすると、他には問題ないの?」「『京都問題』がありますね。あるいは『人参問題』といってもいいでしょう」「... なにそれ?」「短い文字列を検索するとき、それより長い文字列の一部にマッチしてしまうことがあります」「ほお ... つまりアレか、『東京都』に『京都』が含まれるってことか」「御名答!」「... じゃ、ニンジンはなんなの?」「これは『外国人参政権』とかですね」「がいこく ... あ、なるほど」「ここいらへんは漢字文字列そのものの性質なんで、基本的には回避不能ですね」「ふーん」「基本的に短い文字列の検索の方が難しいですよ。他に

になんか情報があれば、まあ区別もつくんでしょ。あるいはなんかこう、除外ができれば」「除外ね ... そうか、京都にマッチして、東京都にマッチしないとかか」「そうです。ここいらへんは『C 問題』とい

う。そりゃまあ、プログラミング言語 C ができた頃は検索とか夢の夢だったでしょうけど、こんなに検索しづらい言語ないですよ」「... そんなもんかね」「です。今どき検索を意識してネーミングするのって大事ですよ。なるべく少ない文字列で、かつ他と被ってなく、名前の一部に他のものを極力含まないとか」「... ややこしいんだね」「まあそれだけ検索が生活に溶け込んでくるということでしょう。今現在東京にいるとわかってる端末で『京都』と検索されたのなら、多分東京についての記事を知りたいわけじゃないだろう、とか推認できるでしょうけど ... まあその推認が正しいかどうかは微妙ですよ」「まあ、でもそういうのやってるんだろね、google 様は」「似たようなのは『北広島問題』ですかね。例えば広島駅は広島県、北広島駅は北海道、とか」「あー、これに『京都問題』がからむと余計ややこしいね」「当初のアルゴリズム検索ではここいらへんの問題もありましたが、最近はあんまり気にならないですから」「色々やってるんだろな」「でしようね。まあ雲の上の話です。クラウドだけに」「すてらのなびこーん」

Hiroyasu Sakaguchi
フリー IT エンジニア

映像スタジオ施工

多様化するデジタル映像環境に対応、映像スタジオ施工なら豊富な実績、直営システムに依る徹底したコストダウンを実現する



MA室ブース各種編集室

新設、リニューアルに関わらず何でもご相談ください。

～映像・音響専門で
47年～

(映像・音響・防音・建築・設計・施工)

一級建築士事務所

高橋建設株式会社

本社 〒216-0032 神奈川県川崎市宮前区神木1-7-8

TEL 044-853-0547 FAX 044-852-1588

(社) 日本ボストプロダクション協会会員 / (社) 日本音楽スタジオ協会会員
(社) 日本音響学会会員

<http://www.takahashi-kensetsu.co.jp>
info@takahashi-kensetsu.co.jp