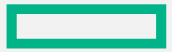


Hewlett Packard
Enterprise

文字コードの話

Noriyoshi Shinoda

December 21, 2021

SPEAKER
篠田典良（しのだのりよし）

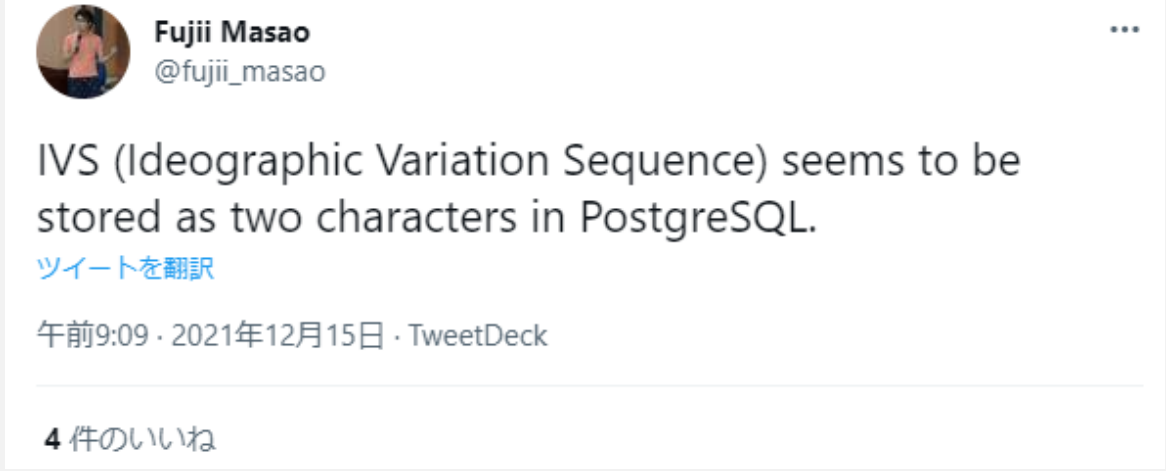
- ✓所属
  - ✓日本ヒューレット・パッカード合同会社
- ✓現在の業務
  - ✓PostgreSQL をはじめ、Oracle Database, Microsoft SQL Server, Vertica 等 RDBMS 全般に関するシステムの設計、移行、チューニング、コンサルティング
  - ✓Oracle ACE
  - ✓オープンソース製品に関する調査、検証
- ✓関連する URL
  - ✓「PostgreSQL 虎の巻」シリーズ
    - ✓http://h30507.www3.hp.com/t5/user/viewprofilepage/user-id/838802
  - ✓Oracle ACE ってどんな人？
    - ✓http://www.oracle.com/technetwork/jp/database/articles/vivadeveloper/index-1838335-ja.html

SPEAKER
篠田典良（しのだのりよし）

- ✓PostgreSQL Unconference #15 (2020/07/30)
  - ✓検知できない破壊の話
- ✓PostgreSQL Unconference #20 (2021/02/02)
  - ✓プロセス障害の話
- ✓PostgreSQL Unconference #26 (2021/08/24)
  - ✓エラーが出ない話
- ✓PostgreSQL Unconference #29 (2021/12/21)
  - ✓文字コードの話
- ✓スライドはこちら
  - ✓https://www.slideshare.net/noriyoshishinoda

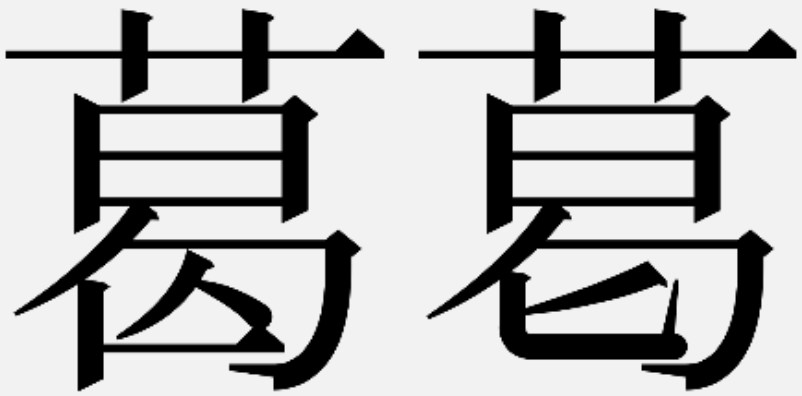
はじまりは？
Twitter

✓ 藤井さんの検証



文字コードの話
葛飾区と葛城市

異字体シーケンス
Ideographic Variation Sequence



- ✓ 漢字異字体シーケンス
  - ✓ 例は「東京都葛飾区」と「奈良県葛城市」の字体
  - ✓ ほぼ同一字形の文字で同一とみなされることが多い
  - ✓ 文字コードに異字体セレクタ（Variation Selector）を付与できる
  - ✓ すべての異字体がコード化されているわけではない

文字	違い
葛 葛	文字コードは U+845B、異字体セレクタで区別
骨 骨	フォントが異なるだけ（文字コードは U+9AA8）
齋 齊	元々コードが異なる（文字コードは U+9F4B と U+9F4A）

異字体シーケンス
Ideographic Variation Sequence

- ✓ UTF-8 の場合
  - ✓ UTF-8 では 7 ～ 8 バイトになる

文字	UCS	UTF-8	備考
葛	U+845B + U+E0100	E8 91 9B F3 A0 84 81	葛飾区
葛	U+845B + U+E0101	E8 91 9B F3 A0 84 80	葛城市
辻	U+8FBB + U+E0100	E8 BE BB F3 A0 84 80	
辻	U+8FBB + U+E0101	E8 BE BB F3 A0 84 81	
鯖	U+9BD6 + U+E0102	E9 AF 96 F3 A0 84 82	
鯖	U+9BD6 + U+E0103	E9 AF 96 F3 A0 84 83	

PostgreSQL
PostgreSQL 14

✓文字エンコードの指定は UTF8
✓2文字と認識される
  ✓LENGTH 関数は文字数を返す

```
postgres=> CREATE TABLE ivs1(c VARCHAR(10));
CREATE TABLE
postgres=> INSERT INTO ivs1 VALUES (CONVERT_FROM('¥xE8919BF3A08481'::bytea, 'UTF8'));
INSERT 0 1
postgres=> SELECT c, LENGTH(c) chr, OCTET_LENGTH(c) byte FROM ivs1;

 c    | chr    | byte
------+--------+--------
 葛   |      2 |      7
(1 row)
```

PostgreSQL
PostgreSQL 14

✓文字の区切りは3バイト＋4バイト

```
postgres=> SELECT OCTET_LENGTH(LEFT(c, 1)), OCTET_LENGTH(RIGHT(c, 1)) FROM ivs1;

 octet_length | octet_length
--------------+--------------
            3 |            4
(1 row)
```

✓Shift_JIS へのコード変換は失敗

```
postgres=> SELECT CONVERT_TO(c, 'SJIS') FROM ivs1;

ERROR:  character with byte sequence 0xf3 0xa0 0x84 0x81 in encoding "UTF8" has no
equivalent in encoding "SJIS"
```

MySQL
MySQL 8.0.27

✓文字コードの指定

文字コードの指定	1文字のバイト数	説明
utf8mb4	最大 4 バイト	
utf8mb3	最大 3 バイト	旧バージョンのデフォルト

✓2文字と認識される

✓LENGTH 関数はバイト数を返す

```
mysql> INSERT INTO ivs1 VALUES (x'E8919BF3A08481');
mysql> SELECT c, CHARACTER_LENGTH(c) chr, LENGTH(c) byte FROM ivs1;
+--------+------+------+
| c      | chr  | byte |
+--------+------+------+
| 葛     |    2 |    7 |
+--------+------+------+
1 row in set (0.00 sec)
```

MySQL
MySQL 8.0.27

✓リテラル値に対して CHARACTER_LENGTH 関数を実行すると結果が異なる
- ✓リテラル値で書いた値は varbinary 型になるらしい
- ✓CAST 関数を使って文字列型に変換してから実行すると 2 が返る

```
mysql> SELECT CHARACTER_LENGTH(x'E8919BF3A08481');
+-----------------------------------+
| CHARACTER_LENGTH(x'E8919BF3A08481') |
+-----------------------------------+
|                                 7 |
+-----------------------------------+
1 row in set (0.00 sec)
```

Oracle Database

Oracle Database 21c

✓文字コードの指定

文字コードの指定	1文字のバイト数	説明
UTF8	最大 3 バイト	Unicode 3.0 CESU-8 コード体系
UTFE	最大 3 バイト	Unicode 3.0 UTF-EBCDICコード体系
AL32UTF8	最大 4 バイト	Unicode 12.1 UTF-8
AL16UTF16	最大 2 バイト	Unicode 12.1 UTF-16BE（NCHAR 型用）

✓2文字と認識される

✓LENGTH 関数は文字数を返す

```
SQL> INSERT INTO ivs1 VALUES (UTL_RAW.CAST_TO_VARCHAR2(HEXTORAW('E8919BF3A08481')));
SQL> SELECT c, LENGTH(c) chr, LENGTHB(c) byte FROM ivs1;
C                CHR        BYTE
---------- ---------- ----------
葛                2          7
```

Oracle Database

Oracle Database 21c

✓CSV ファイルのロード(SQL*Loader)は失敗する

```
$ sqlldr userid=scott/{password} control=ivs1.ctl log=ivs1.log
$ cat ivs1.log

SQL*Loader: Release 21.0.0.0.0 - Production on 水 12月 15 09:43:46 2021
Version 21.3.0.0.0
…
レコード1: 拒否されました。- 表IVS1,列Cでエラーが発生しました。
マルチバイト・キャラクタでエラーが発生しました。

表IVS1:
  0 行は正常にロードされました。
…
```

SQL Server
SQL Server 2019

✓文字コードの指定

✓SQL Server 2019 から UTF-8 の指定可能

文字コード(Collation)の指定	1文字のバイト数	説明
Japanese_XJIS_140_CS_AS_KS_WS_UTF8	最大 8 バイト	char / varchar を UTF-8 に指定

✓1文字と認識される

✓LEN 関数は文字数を返す

```
1> INSERT INTO ivs1 VALUES(CONVERT(varbinary(10), 0xE8919BF3A08481));
2> SELECT c, LEN(c) chr, DATA_LENGTH(c) byte FROM ivs1;

c              chr         byte
------- ---------- ----------
葛                1          7
```

SQL Server
SQL Server 2019

✓SQL Server 2019 の照合順序例

✓データベースのデフォルト、テーブル、列単位で変更可能

```
Japanese_XJIS_140_CS_AS_KS_WS
Japanese_XJIS_140_CS_AS_KS_WS_UTF8
Japanese_XJIS_140_CS_AS_KS_WS_VSS
Japanese_XJIS_140_CS_AS_KS_WS_VSS_UTF8
```

✓ 上記照合順序は以下のように分解できます。

照合順序の指定	指定例	説明
言語	Japanese_XJIS_140	SQL Server 2017 以降の日本語文字範囲
文字区別	_CS_AS_KS_WS	大文字／小文字、アクセント記号、カナ、全角半角を区別
異字体セレクタ指定	_VSS	_VSS は異字体を区別する
文字コード指定	_UTF8	char/varchar/text 型の文字コード

照合順序
照合順序の比較

✓様々な RDBMS の照合順序指定
- ✓すべての組み合わせが使えるわけではない

区別	SQL Server	MySQL	Oracle Database	PostgreSQL
大文字／小文字	CI/CS	ci/cs	CI/CS	citext extension
アクセント	AI/AS	ai/as	AI (CI 含む)	unaccent extension
カナ／かな	KS	ks	VS？	-
半角／全角	WS	unicode	M	-
IVS	VSS	-	-	-

THANK YOU

Mail: noriyoshi.shinoda@hpe.com
Twitter: @nori_shinoda