



The next sessions Title:

“Using MySQL in a Japanese Environment...”

よつぎ

Please kindly send a copy of any taken photos and videos to: [valentin\\_nils@be-known-online.com](mailto:valentin_nils@be-known-online.com)



The next sessions Title:

“Using MySQL in a Japanese  
Environment...”

... and avoiding common pitfalls



# Using MySQL in a Japanese Environment...

- 0) Introduction
- 1) Google Numbers
- 2) Required Definitions
- 3) General Considerations
- 4) Setting Character Sets and Collations
- 5) Summary
- 6) Questions and Answers

... and avoiding common pitfalls



# 0) Introduction

Does everybody hear me loud & clear ???

;-)

(Short intro about myself in English / Japanese)



# 1.0) Google Numbers (World)

# of mentions on the web according to google.com\*

- MySQL 17.6 M (+16.5%)
- Oracle 12.7 M (+6.7%)
- PostgreSQL 2.97M + 0.67M (-25%)(postgresql + postgres)

# of links to site according to google.com\*

- MySQL.com 131 T (+137%)
- Oracle.com 30.7 T (+114%)
- Postgresql.org 13.7 T (-5%)

\*as of 10th April 2004 and 5th Sep. 2003



# 1.1) Google numbers for Japan 1/2

# of mentions on the web according to google.co.jp\*

- MySQL 318 T
- Oracle 970 T
- PostgreSQL 383 T + 45 T (postgresql + postgres)

# of links to site according to google.co.jp\*

- Mysql.gr.jp 132
- Oracle.co.jp 1110
- Postgresql.jp 295

\*as of 10th April 2004



# 1.1) Google numbers for Japan 2/2

Softagency.co.jp	394	(Distributor)
jp.Xoops.org	591	(CMS)

other observations about MySQL activities in Japan :

- MySQL documentation in Japanese now complete ;-)
- Strong Partnership relations
- Local MySQL KK presence & homepage would further strengthen business relations



## 2.0) Required Definitions

**Q:** What is a character set ?

**A:** Set of allowed symbols and encodings

**Q:** What is a collation ?

**A:** Set of rules for comparing characters\*

**Q:** Why using a collation ?

**A:** To organize / sort your data.

\*characters must exist in the same character set



## 2.1) A 10-letter mini Character Set, ... (Sample)

あ あい いいう う っ え えお お

Sound: a A i l u U e E o O

Value: 0 1 2 3 4 5 6 7 8 9

---

車 Sound: Kuruma (car)  
Value: 5236



## 2.1) ...its Binary Collation,...

(Sample)

...Where 'A' < 'E' /\* true \*/

-Reason: Encoding value 'A'=1 is less than encoding value 'E'=7

...Where 'U' = 'u' /\* false \*/

-Reason: Encoding value 'U'=5 is more than encoding value 'u'=4

---

**\*Binary Collations are case sensitive (ending \_bin)**

**f.e** **sjis\_bin**  
**ujis\_bin**  
**utf8\_bin**  
**ucs2\_bin**



## 2.1) ...and its non-Binary Collation\* (Sample)

...Where 'A' < 'E' /\* true \*/

-Reason: Encoding value 'A'=1 is less than encoding value 'E'=7

...Where 'U' = 'u' /\* true \*/

-Reason: 'u' is converted to 'U' before comparison

...Where 'U' < 'kuruma' /\* false \*/

-Reason: Encoding value for 'kuruma'=xxxx is not part of the same character set



## 2.1) ...and its non-Binary Collation\* (Sample)

**\*non-Binary Collations are case insensitive (ending \_ci)**

**f.e** `sjis_japanese_ci`  
`ujis_japanese_ci`  
`utf8_general_ci`  
`ucs2_general_ci`



## 2.1.1) Useful additional facts

- MySQL supports 30+ character sets and 70+ collations.
  - Mysql> SHOW CHARACTER SET;
  - Mysql> SHOW COLLATION;
  - 1x Character Set = X x Collations ...
- **...BUT 1x Collation = 1x Character Set**

\* from Version >= 4.1.x



## 2.2) Character Sets for the Japanese Environment

4.0.x	4.1.x	Collation	Description	Max length
Ujis	Ujis	ujis_japanese_ci	EUC-JP Japanese	3
Sjis	Sjis	sjis_japanese_ci	SHIFT-JIS Japanese	2
--	<b>UTF8</b>	<b>utf8_general_ci</b>	<b>UTF-8 Unicode</b>	<b>3</b>
--	<b>UCS2</b>	<b>ucs2_general_ci**</b>	<b>UCS-2 Unicode</b>	<b>2</b>

**\*\* (new in Version => 4.1.x )**

**\*\* (client side not yet implemented as of 01.03.2004, but hopefully coming soon ;-)**



## 3) General Considerations

- .0 Changes from MySQL Version 4.0.x to 4.1.x
- .1 Displaying Data
- .2 General Database Design
- .3 Datatypes
- .4 Table formats
- .5 Using multi-byte safe commands (operation)
- .6 Export of Data
- .7 Import of Data
- .8 Compatibility with other DBMSs



# 3.0 Character Set related Changes from MySQL 4.0.x to 4.1.x

Dir: ~mysql-install-dir/share/mysql/charsets

Config files:

	4.0.x		4.1.x
	latin1.conf	>>	latin1.xml
	latin2.conf	>>	latin2.xml
	hp8.conf	>>	hp8.xml
	swe7.conf	>>	swe7.xml
	cp1251.conf	>>	cp1251.xml

- 1x XML formatted file = 1x Character Set
- Exception: UJIS/SJIS/UTF8/UCS2 compiled into binary



## 3.1 Displaying Data

- Character Sets and Collations
- Driver limitations
  - JDBC driver (only Fonts provided by JRE or SDK usable)
- OS locale settings (encoding)
  - Unix / Linux: locale command
  - Windows: locale settings
- Application specific settings



## 3.2 How much space for which data ?

- One Byte :  
Basic Latin letters, digits and punctuation
- Two Bytes:  
Special Latin letters (with tilde, macron, acute, grave and other accents), Cyrillic Greek, Armenian, Hebrew, Arabic, etc.
- Three Bytes:  
Korean, Chinese and Japanese ideographs
- Four Bytes (no support yet on Client side)



## 3.3) General Database Design

- Note that Chinese, Japanese and Korean (CJK) ideographs use often three-byte sequences.
- Save up to 66% by choosing the right character set / collation for your data.
- Use VARCHAR instead of CHAR to Save space with UTF8.
- Note: Four-byte characters currently not supported

Comparison	Char (100)	Varchar (100)
max_length	300 bytes	100-200 bytes
Saved space	0%	33-66%



## 3.4 ) Table formats supporting Character Set settings (Version =>4.1.1)

- ISAM (not planned)
- MyISAM
- MEMORY (HEAP)
- InnoDB storage engine \*\*

[NOTE: Code is already in the 4.1 tree, Heikki said.]

\*\* (not yet implemented as of 01.03.2004, but coming real soon ;-)



## 3.5) Operation with multi-byte safe commands

- ALTER, CREATE, DROP, INSERT, REPLACE, SELECT, UPDATE, TRUNCATE
- REGEXP (!!)
- as of 7 April 2004 seem to handle only single byte character set (Maxlen=1)
- >> update of regexp library required



## 3.6 Export of UTF-8 formatted data

As usual



## 3.7 Import of UTF-8 formatted data

- 1) Preparation: Set default character set to UTF-8  
Server/client/connection
- 2a) LOAD DATA LOCAL INFILE + escape character
  - FIELDS TERMINATED BY ,
  - FIELDS ENCLOSED BY “
  - LINES TERMINATED BY \r\n (Windows OS)
- 2b) mysqlimport
- 3) if Steps 1 + 2 done correctly data will NOT  
Gibberish >> jap. “motchibake”
- 4) Check the information LOAD DATA returns



## 3.8 Compatibility with Other DBMSs

For SAP DB compatibility these two statements are the same:

- `CREATE TABLE t1 (f1 CHAR(n) UNICODE);`
- `CREATE TABLE t1 (f1 CHAR(n) CHARACTER SET UCS2);`



## 4) Character Sets & Collations Setup

- .0) Preparation
- .1) Server side
- .2) Client side
- .3) Troubleshooting with an Command-Line Interface
- .4) Operating System Environment



## 4.0.1 How to get information ?

- `>mysql --version (Client)`
- `>mysql --help [|grep character]`
- `>mysql --help [|grep groups]`
- `>mysql --print-defaults`



## 4.0.2 Which commands shows me what ?

Command	Server	DB	Table	Column	Result	Connection	Client
Mysql> \s	X						X
SHOW VARIABLES LIKE '%char%';	X	<b>X**</b>			<b>X**</b>	<b>X**</b>	X
SHOW CREATE DATABASE;		X					
SHOW CREATE TABLE;			X	X			

\*\* from Version >=4.1



## 4.0.3 MySQL 4.0.x status Information

- mysql> \s  
mysql Ver 12.22 Distrib 4.0.18, for pc-linux (i686)  
...  
Server version: 4.0.18-max-log  
Protocol version: 10  
Connection: Localhost via UNIX socket  
Client character set: ujis  
Server character set: ujis  
...  
> SHOW VARIABLES LIKE "%char%";

Variable_name	Value
character_set	ujis



## 4.0.4 MySQL 4.1.x Status Information

• > SHOW VARIABLES LIKE "%char%";

Variable_name	Value
character_set_server	ucs2
character_set_system	utf8
character_set_database	ucs2
character_set_client	utf8
character_set_connection	utf8
character-sets-dir	/usr/local/mysql-max-4.1.1-alpha-pc -linux-i686/share/mysqlCharsets/
character_set_results	utf8



## 4.1.1 Which collation applies ?

- Levels
  - Server
  - Database
  - Table
    - Table
    - Column
  - Connection
- Case 1) Character Set X and Collation Y is defined
  - Use Character set X and Collation Y
- Case 2) Only Character Set X is defined
  - Use Character Set X and its default Collation
- Case 3) Neither the Character Set nor Collation were defined
  - The default set Character Set or from the next higher Level is used



## 4.1.0 Server based Character Settings

Setting the servers default character set

- `>mysqld -default-character-set=ujis`

or (in my.cnf etc.)

- `[mysqld]  
default-character-set=ujis`



## 4.1.2 DB based Character Settings

- `mysql>CREATE DATABASE db_name  
[DEFAULT CHARACTER SET character_set_name  
[COLLATE collation_name]]`
- `mysql>ALTER DATABASE db_name  
[DEFAULT CHARACTER SET character_set_name  
[COLLATE collation_name]]`



## 4.1.3 Table based Character Settings

Note: **Setting will override the default Character Set specified for this database !!**

- `mysql>CREATE TABLE table_name ( column_list )  
[DEFAULT CHARACTER SET character_set_name  
[COLLATE collation_name]]`
- `mysql>ALTER TABLE table_name  
[DEFAULT CHARACTER SET character_set_name]  
[COLLATE collation_name]`



## 4.1.4 Column based Character Settings

Note: Setting will override the default Character Set specified for this  
table !!

- `mysql>CREATE TABLE t1 (f1 CHAR(n) UNICODE);`
- `mysql>CREATE TABLE t1 (f1 CHAR(n)  
CHARACTER SET ucs2);`



# 4.1.5 Connection based Character Settings

Note: **Setting will override the default Character Set specified for this Client connection !!**

- SET NAMES 'character\_set\_name'
- SET CHARACTER SET character\_set\_name



## 4.2 Client side Character Settings

- `>mysql --default-character-set=ujis`  
or (in `my.cnf` etc.)
- `[mysql]`  
`default-character-set=ujis`



## 4.2.1 Driver specific Settings

Jdbc driver with Javas implemented fonts



## 4.2.2 Application Specific Settings

Sample: **DbVisualizer\*\***

```
dbvis.grid.encode=true
```

```
dbvis.grid.fromEncode=EUC_JP
```

```
dbvis.grid.toEncode=EUC_JP
```

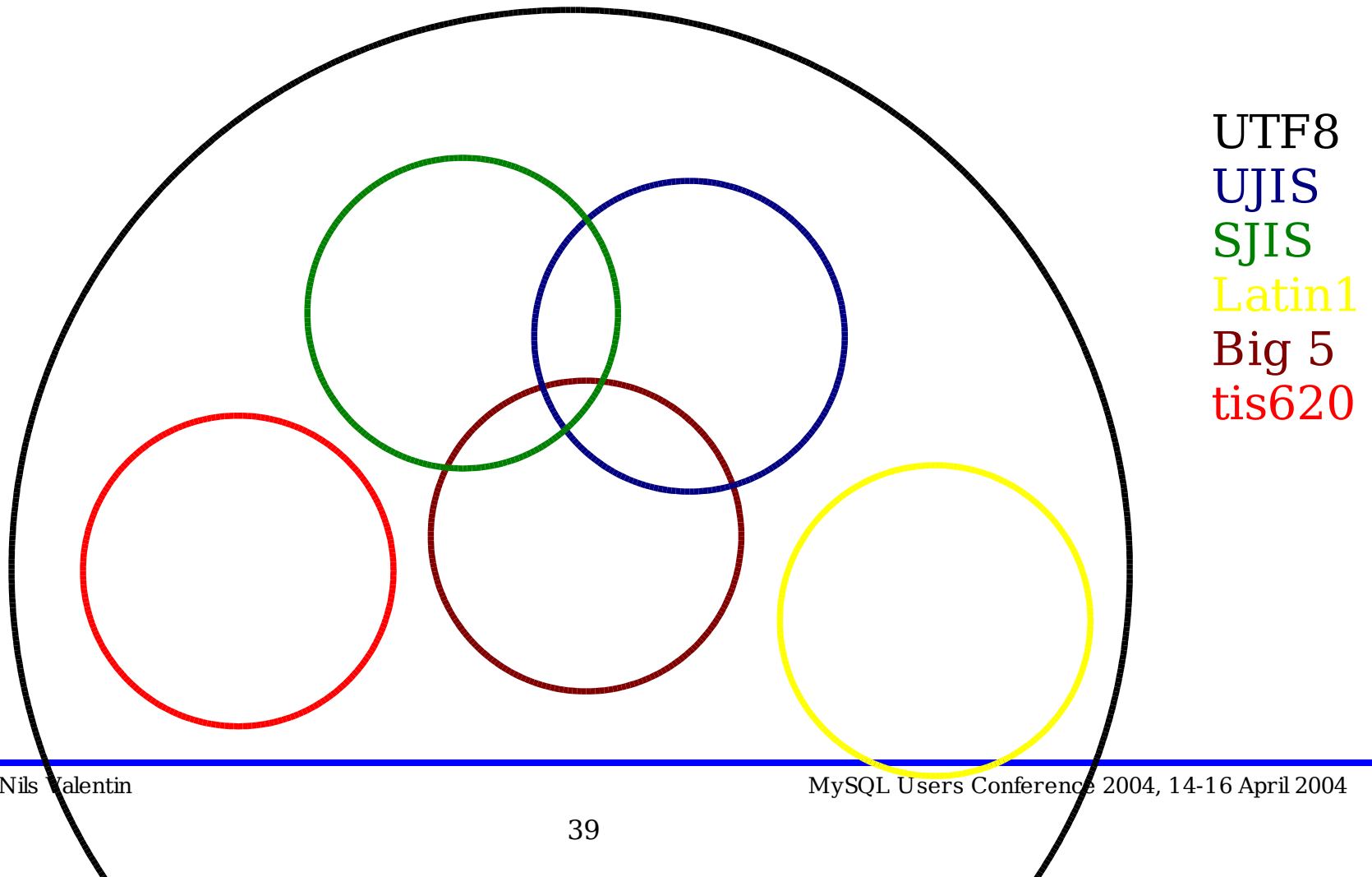
---

\*\*Settings can be added in `~/DbVisualizer-ver-xx/.dbvis.lax`



## 4.3.0 Troubleshooting notes

Note: Most fonts only support a subset of Unicode characters





## 4.3.1 Troubleshooting with a Command-line Interface (1/2)

- 1) Try selecting the hex value of the columns to see if values are being stored and transmitted properly:

```
SELECT HEX(Japanese) FROM table;  
SHOW FULL COLUMNS FROM table;
```

- 2) Check the code points :

```
SELECT HEX(convert(Japanese using ucs2)) FROM table;
```

Note: hex (UTF8 value) != Unicode point

## 4.3.1 Unicode charts (www.unicode.org)

### Hiragana

### Codechart 3040

	304	305	306	307	308	309
0		ぐ 3050	だ 3060	ば 3070	む 3080	る 3090
1	あ 3041	け 3051	ち 3061	ぱ 3071	め 3081	ゑ 3091
2	あ 3042	げ 3052	ぢ 3062	ひ 3072	も 3082	を 3092
3	い 3043	こ 3053	っ 3063	び 3073	や 3083	ん 3093
4	い 3044	ご 3054	っ 3064	び 3074	や 3084	う 3094



## 4.3.2 Troubleshooting with a Command-line Interface (2/2)

3) If you're testing in xterm with the mysql client be sure that it is set up for Unicode and is using a proper font too.

4) Check wich locale your system is set to.

Unix/Linux/BSD: > locale

Windows: Cont.rol Panel/locale settings

5) Problem: Database contents shows as gibberish on konsole, also the contents is shown correct in a web environment.

This occurs f.e when saving a html-file in sjis encoding and viewing it from a Unix/Linux/BSD konsole or terminal with the system set to UJIS or EUC-JP coding by default.



## 4.4 Using a 4.0.x mysql Client + 4.1.x Server

Symptom / Error message:

```
>bin/mysqladmin: connect to server at 'localhost' failed
```

Problem : 'Client does not support authentication protocol requested by server; consider upgrading MySQL client'

Cause: character-set = utf8 in [client] section of my.cnf

Solution:

- upgrade the client version (or)
- disable the related client setting in my.cnf



## 4.5 Considerations for Data Conversions

- Does the destination Character Set support xx-byte characters ?
- Does the destination Character Set support the Characters I am converting ?
- How much “real” space is required for which data ?



## 4.5.1 How much space for which data ?

- One Byte :  
Basic Latin letters, digits and punctuation
- Two Bytes:  
Special Latin letters (with tilde, macron, acute, grave and other accents), Cyrillic Greek, Armenian, Hebrew, Arabic, etc.
- Three Bytes:  
Korean, Chinese and Japanese ideographs
- Four Bytes (no support yet on Client side)



## 4.5.2 Data conversion example

- `CONVERT (<EXPR> USING <character set>)`

Example:

- `INSERT INTO ujis (ujiscolumn)  
SELECT CONVERT (sjisfield USING ujis) from sjistable;`



## 4.6 Upgrading from Version 4.0 to 4.1

- Run the script  
“mysql\_fix\_privilege\_tables”.
- This will upgrade the User's access  
database (named “mysql”)



## 4.7 Operating System Transfer

- Check for locale settings of your OS
- Check for installed Fonts
  - f.e Mincho, Gothic
- Check for drivers you are using
- Check for “LINES TERMINATED BY”



## 5.0 Summary & Future

- New quality for Character Set related tasks
- Increased detail level
- Increase of overall DB system efficiency
- Client side Character Set conversion (in later 4.0 Release)



# 6.0 Questions and Answers

And now ... its your turn ...

;-)

Any Questions ?



# Appendix A) Useful Weblinks

- MySQL Documentation Chapter 11 Character Set Support  
<http://dev.mysql.com/doc/mysql/en/Charset.html>
- <http://www.unicode.org>
- <http://www.cogsci.ed.ac.uk/~richard/>



# Appendix B) Good Literature

- **Paul Dubois books**
  - MySQL Certification Study Guide, ISBN: 0-672-32632-9
  - MySQL Administrator's Guide, ISBN: 0-672-32634-5 (upcoming)
- **MySQL Press**
  - MySQL Tutorial, ISBN: 0-672-32584-5
  - MySQL Language Reference ISBN: 0-672-32633-7 (upcoming)

No, ...actually I mean it.  
This stuff IS really good.

;-)



# Hope to **see you at the UC-2005**

- Thank you for Listening

**;-)**



# License

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/1.0/> or send a letter to

Creative Commons  
559 Nathan Abbott Way  
Stanford  
California 94305  
USA.

Note: Links to external sources are provided “as is” in respect to the owners of the contents referring to. Please contact the owner of those contents, should you require a permission to use it in your publication.



# What is not yet covered here ?

- Coercibility Rules
- Result String Conversions
- `cvt_file.pl`
  - `f.e` convert a tab-delimited file format to a colon-delimited file format