UTF-8: Experiences & Issues in a Helma-MySQL Setup

Michael Platzer

Character Encodings

„A character encoding is a code that pairs a set of natural language characters (..) with a set of something else, such as numbers or electrical pulses.” [http://en.wikipedia.org/wiki/Character_encoding]

- ASCII: 7bit (i.e. 128chars)
- ISO-8859-1 (a.k.a. Latin-1): 8bit;
  covers English, German, French (except for œ), Spanish, Portuguese, Scandinavian Languages,..
  Does *not* include Eastern-European Languages (ISO-8859-2), resp. Turkish (ISO-8859-9)
- UTF-8: Variable-Length-Encoding, 1-4Byte
- UTF-16: 2Byte-Encoding, used in Windows
- UTF-32: Fixed-Length 4Byte Encoding
- Punycode: Used for DNS-Entries

Helma (the source)

- CharacterSet of the OS
  - Windows: Cp1252 (~ISO-8859-1)
  - Linux: “locale”

- CharacterSet of Java

```
set JAVA_OPTIONS = -Dfile.encoding=ISO-8859-1
```

- CharacterSet of Skins

```
app.properties
skinCharset = UTF-8
```

- CharacterSet of Message-Files

  java.util.Properties is *always* ISO-8859-1 !

Helma (the response)

- CharacterSet of Helma-Application

```
app.properties
charset = UTF-8
```

- CharacterSet of the Response (HTTP)

```
main.hac
res.charset = UTF-8
```

- CharacterSet of the Response (HTML)

```
<meta http-equiv=“Content-Type”
content=“text/html; charset=utf-8”/>
```

- Open Issues:
  - format(.) also encodes Characters
  - HopObject.href(.)

Helma/MySQL (the data)

- Requires MySQL 4.1+

- MySQL CharacterSets and Collations

```
mysql> show character set;
mysql> show collation like ‘latin%’;
```

- MySQL: Determine current settings

```
mysql> show variables like ‘%char%’;
mysql> show table status from db_twoday;
Mysql> show full columns from AV_POLL;
```

  Note: Just the columns actually store the character-encoding-information. All other settings are just used as defaults.

- CharacterSet of the DB-Connection

```
db.properties
twoday.url=jdbc:mysql://localhost/db_twoday?useUnicode=true&characterEncoding=utf8
```

MySQL Issues

- Conversion of CharacterSets
  1. full database dump
  2. replace ‘latin1’ with ‘utf8’
  3. insert dump

- Client uses different CharacterSet than Column

```
mysql> select SITE_ALIAS from AV_SITE where SITE_ALIAS='michi';
ERROR 1267 (HY000): Illegal mix of collations
   (latin1_swedish_ci,IMPLICIT) and (utf8_general_ci,COERCIBLE)
   for operation '=‘

Lösung:
mysql> set names ‘utf8’;
```