TYPO3 Core - Feature #80398

Make default charset and collation for new tables configurable

2017-03-22 15:58 - Marco von Arx

Status:	Closed	**Start date:**	2017-03-22
Priority:	Should have	**Due date:**	
Assignee:		**% Done:**	100%
Category:	Database API (Doctrine DBAL)	**Estimated time:**	0.00 hour
Target version:	9 LTS		
PHP Version:	7.0	**Complexity:**	
Tags:	charset utf8mb4	**Sprint Focus:**	

Description

to be able to store 4 byte unicode characters we need to set database to utf8mb4. since typo3 8 there is a configuration parameter for that but it seems that it is not taken into account.

LocalConfiguration.php

```
    'DB' => [
        'Connections' => [
            'Default' => [
                'charset' => 'utf8mb4',
                'dbname' => '--dbname--',
                'driver' => 'mysqli',
                'host' => '127.0.0.1',
                'password' => '--mypassword--',
                'port' => 3306,
                'user' => '--myuser--',
            ],
        ],
    ],
```

create table statements do have a fallback but do not read from configuration

```
    private function buildTableOptions(array $options)
    {
        if (isset($options['table_options'])) {
            return $options['table_options'];
        }

        $tableOptions = array();

        // Charset
        if ( ! isset($options['charset'])) {
            $options['charset'] = 'utf8';
        }
        ....
    }
```

DatabaseConnection class also does not read charset configuration either it takes utf8 as a default.

```
        $connection = \Doctrine\DBAL\DriverManager::getConnection([
            'driver' => 'mysqli',
            'wrapperClass' => Connection::class,
            'host' => $host,
            'port' => (int)$this->databasePort,
            'unix_socket' => $this->databaseSocket,
            'user' => $this->databaseUsername,
            'password' => $this->databaseUserPassword,
            'charset' => $this->connectionCharset,
```

```
        ]);
```

it was stated that it would be fixed in CMS 8
https://forge.typo3.org/issues/71454

is this on roadmap? before LTS?

Related issues:

Related to TYPO3 Core - Feature #80659: Set Charset to utf8mb4	**Closed**	**2017-04-03**
Related to TYPO3 Core - Bug #82551: Upgrade Wizard Deadlock	**Closed**	**2017-09-25**
Related to TYPO3 Core - Bug #82080: Indexes too large for some tables with ut...	**Closed**	**2017-08-11**
Related to TYPO3 Core - Feature #71454: Allow setting Connection Charset	**Closed**	**2015-11-10**
Related to TYPO3 Core - Bug #86793: Renamed columns are not correctly detecte...	**Closed**	**2018-10-30**
Has duplicate TYPO3 Core - Bug #85524: Charset for DB Connections in LocalCon...	**Closed**	**2018-07-09**

Associated revisions

Revision ed806ef5 - 2018-09-11 17:30 - Lienhart Woitok

[FEATURE] Use utf8mb4 on mysql for new instances

If installing a new TYPO3 instance on mysql, utf8mb4 is now used as
default charset for the database connection and as default collation.

Upgraders may change LocalConfiguration to use utf8mb4, too. They
however need to take care of changing their collations and setting
according table detaults on their own.

A reports status check verifies there is no mixed collation.

Resolves: #80398
Resolves: #82080
Resolves: #82551
Releases: master
Change-Id: I6bf464a22c6ed74631bf5aacff9c2cfe670077da
Reviewed-on: https://review.typo3.org/56440
Reviewed-by: Christian Kuhn <lolli@schwarzbu.ch>
Tested-by: Christian Kuhn <lolli@schwarzbu.ch>
Tested-by: TYPO3com <no-reply@typo3.com>
Reviewed-by: Lienhart Woitok <lienhart.woitok@netlogix.de>
Tested-by: Lienhart Woitok <lienhart.woitok@netlogix.de>
Reviewed-by: Georg Großberger <garfieldius67@gmail.com>
Reviewed-by: Jigal van Hemert <jigal.van.hemert@typo3.org>
Tested-by: Jigal van Hemert <jigal.van.hemert@typo3.org>

History

#1 - 2017-03-23 14:11 - Marco von Arx

the issue is not DatabaseConnection class. charset is read properly from configuration there

it seems that
TYPO3\CMS\Core\Database\Schema\ConnectionMigrator
or TYPO3\CMS\Core\Database\Schema\SchemaMigrator
does not read that configuration parameter

I was able to work arround by adding the following
in TYPO3\CMS\Core\Database\Schema\ConnectionMigrator line 1211

```
$tableOptions = $table->getOptions();
            $connectionParams = $connection->getParams();
            if (isset($connectionParams['charset'])) {
                $tableOptions['charset'] = $connectionParams['charset'];
            }
            if (isset($connectionParams['collate'])) {
                $tableOptions['collate'] = $connectionParams['collate'];
            }
```

#2 - 2017-03-23 14:20 - Morton Jonuschat

- Status changed from New to Needs Feedback

Hi!

I think you are mixing two concepts here. Also I think the buildTableOptions() code example is from Doctrine, which is a 3rd-Party Library and has no idea about TYPO3 configuration

1. Connection Charset

This defines the character set the client will use to send SQL statements to the server. It also specifies the character set that the server should use for sending results back to the client. (For example, it indicates what character set to use for column values if you use a SELECT statement.)

2. Storage character set

This defines in which way the Database stores data on disk/in memory. This is controlled by Server/Database/Table/Column options, not the Connection Charset.

If I understand your report correctly you are looking for a way to tell TYPO3 to override the UTF8 default character set (and collation?) for created tables?

#3 - 2017-03-23 14:28 - Marco von Arx

Hi Morton

we need to store 4 Byte Unicode characters like emoji 'http://apps.timwhitlock.info/emoji/tables/unicode'
the default utf8 does only allow storing 3 byte unicode characters. most of emoji characters cannot be stored into utf8.

thats why i need the connection to be utf8mb4 and the database to create tables with charset utf8mb4 and collate utf8mb4_unicode_ci

the first part does indeed work. Typo3 does connect with charset utf8mb4 if i set it in LocalConfiguration.php

but how can I ensure that tables are created with correct charset and collate during setup?

#4 - 2017-04-08 07:05 - Morton Jonuschat

- Status changed from Needs Feedback to New

- Priority changed from Must have to Should have

- Target version set to Candidate for Major Version

This is a new feature which could be implemented for TYPO3 9.0. Doing it using the connectionParameters is not the preferred way as the connection and the tablespace are two different things.
Also this needs to be supported across multiple database connections and database engines.

#5 - 2017-04-08 07:05 - Morton Jonuschat

- Tracker changed from Bug to Feature

- Subject changed from connection charset ignored to Make default charset and collation for new tables configurable

#6 - 2017-05-04 08:29 - Marco von Arx

Symfony has separate config parameter for table schemes http://symfony.com/doc/current/doctrine.html

```
doctrine:
    dbal:
        charset: utf8mb4
        default_table_options:
            charset: utf8mb4
            collate: utf8mb4_unicode_ci
```

as a suggestion

```
'DB' => [
        'Connections' => [
            'Default' => [
                'charset' => 'utf8mb4',
                'dbname' => '--dbname--',
                'driver' => 'mysqli',
                'host' => '127.0.0.1',
                'password' => '--mypassword--',
                'port' => 3306,
                'user' => '--myuser--',
                'tableoptions' => [
                     'charset' => 'utf8mb4',
                     'collate' => 'utf8mb4_unicode_ci'
                ]
            ],
        ],
```

```
    ],
```

#7 - 2018-02-16 15:01 - Tymoteusz Motylewski

please keep in mind that utf8mb4 uses 4 bytes per char, while "standard" utf8 collation uses 3 bytes per char, which means that indices created might exceed maximum key length limit in mysql.
E.g. by default the key size is 767 which is lower than varchar(255) in utf8, but exceeded with varchar(255) with utf8mb4 (255*4 = 1020)

#8 - 2018-03-02 13:45 - Tymoteusz Motylewski

- Related to Bug #82551: Upgrade Wizard Deadlock added

#9 - 2018-03-02 13:45 - Tymoteusz Motylewski

- Related to Bug #82080: Indexes too large for some tables with utf8mb4 added

#10 - 2018-03-02 13:51 - Tymoteusz Motylewski

FYI, MySQL 8 will come with utf8mb4 as default charset

#11 - 2018-03-23 14:06 - Gerrit Code Review

- Status changed from New to Under Review

Patch set 1 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#12 - 2018-03-23 14:40 - Lienhart Woitok

I have pushed a change to gerrit which implements the config suggestion by Marco von Arx. I'm not entirely sure I found all relevant places to change, but in my tests this worked for the database analyzer in the install tool. Newly created tables are generated with utf8mb4.

#13 - 2018-07-12 10:12 - David Henninger

- Has duplicate Bug #85524: Charset for DB Connections in LocalConfiguration.php ignored added

#14 - 2018-07-12 11:00 - Riccardo De Contardi

- Related to Feature #71454: Allow setting Connection Charset added

#15 - 2018-08-31 15:54 - Gerrit Code Review

Patch set 2 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#16 - 2018-08-31 15:58 - Gerrit Code Review

Patch set 3 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#17 - 2018-09-06 15:50 - Tymoteusz Motylewski

- Target version changed from Candidate for Major Version to 9 LTS

#18 - 2018-09-09 12:54 - Gerrit Code Review

Patch set 4 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#19 - 2018-09-09 13:07 - Gerrit Code Review

Patch set 5 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#20 - 2018-09-10 12:44 - Gerrit Code Review

Patch set 6 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#21 - 2018-09-10 14:51 - Gerrit Code Review

Patch set 7 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#22 - 2018-09-10 22:58 - Gerrit Code Review

Patch set 8 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#23 - 2018-09-10 23:07 - Gerrit Code Review

Patch set 9 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at https://review.typo3.org/56440

#24 - 2018-09-11 18:00 - Lienhart Woitok

- Status changed from Under Review to Resolved

- % Done changed from 0 to 100

Applied in changeset ed806ef550a63d9034bf4edba8b38b92b1fd71ed.

#25 - 2018-09-12 13:14 - Lienhart Woitok

- File typo3-utf8mb4-0.png added

- File typo3-utf8mb4-1.png added

- File typo3-utf8mb4-2.png added

As requested by Tymoteusz Motylewski some demonstration screenshots of utf8mb4 support in content (using the introduction package). For the first screenshot with normal utf8 (utf8mb3) I added the heart again to demonstrate the failed content, it wasn't there after saving as it couldn't be written to the database.

#26 - 2018-10-02 10:26 - Benni Mack

- Status changed from Resolved to Closed

#27 - 2018-10-30 00:23 - Helmut Hummel

- Related to Bug #86793: Renamed columns are not correctly detected by database schema diff added

Files

typo3-utf8mb4-0.png	104 KB	2018-09-12	Lienhart Woitok
typo3-utf8mb4-1.png	233 KB	2018-09-12	Lienhart Woitok
typo3-utf8mb4-2.png	454 KB	2018-09-12	Lienhart Woitok