

Technical Note:

Mondrian and non-ASCII characters

Revision History:

	Version	Author	Details
1	1.0	<u>Laurentiu</u>	Created this document

Table of Contents

Servers Used	3
Essential Configuration	3
Schema definition	3
Non – ASCII characters in the member names.	4

Servers Used

Pentaho BI Server 5.0

Mondrian 3.6.1

Essential Configuration

In order to have Mondrian work with Unicode characters it is important to declare the correct character encoding for the XMLA servlet. In web.xml check for the introduction of the XMLA servlet and verify there is an init parameters as follows:

```
<init-param>
  <param-name>CharacterEncoding</param-name>
  <param-value>UTF-8</param-value>
</init-param>
```

For example, Pentaho BI Server comes out of the box with:

```
<servlet>
  <servlet-name>Xmla</servlet-name>
  <servlet-class>org.pentaho.platform.web.servlet.PentahoXmlaServlet</servlet-class>
  <init-param>
    <param-name>DataSourcesConfig</param-name>
    <param-value>${pentaho.solutionpath}${pentaho.olap.xmladatasources}</param-
value>
  </init-param>
  <init-param>
    <param-name>CharacterEncoding</param-name>
    <param-value>UTF-8</param-value>
  </init-param>
</servlet>
```

Schema definition

Special attention should be paid to the correct definition of the schema files. The captions for Cubes, Dimensions, Hierarchies, Levels, and Measures are stored in the markup as attributes. This should obey the Unicode Technical Report (check here for reference <http://www.w3.org/TR/unicode-xml/#Notation>)

The following is a correct definition for a dimension, containing a non-ASCII character (Schön):

```
<Dimension foreignKey="CUSTOMERNUMBER" name="Sch&#246;n">
```

Observe the correct definition of the dimension name: (name="Schön" as opposite to Schön).

Another example, this time for Far Eastern characters:

```
<Dimension primaryKey="PRODUCTCODE" name="Product &#x50AB;&#x50AC;&#x50AD;">
```

Leads to a dimension named Product 僕僙儔.

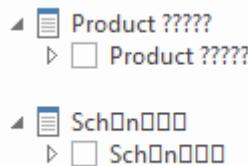
It is also important to define the correct encoding in the beginning of the file:

```
<?xml version="1.0" encoding="UTF-8"?>
```

Observe the encoding declared as UTF-8 here.

It is important to note that this file should be saved as an ASCII file. This means the UTF-8 Byte Order Mark in the beginning of the file (the byte sequence 0xEF, 0xBB, 0xBF) **should not be present!** Trying to load such a file, although well-formed leads to errors in the Pentaho BI server

Failing to observe these rules leads to unreadable captions as those in the picture below:



The picture below shows the intended outcome:



Non – ASCII characters in the member names.

This chapter applies to member names stored in relational databases. It exemplifies with the SteelWheels sample data from Pentaho BI Server 5.0 and MySQL.

It is important to observe the correct encoding for the columns involved in the member description. This is to be set to **utf8_unicode_ci**.

In the test environment used for this paper the encoding was wrong (latin1_general_cs) and was corrected. The following is the corrected definition for products table of SteelWheel

The screenshot shows the ArQuery interface with the URL localhost/sampleddata/products. The top navigation bar includes 'Browse', 'Structure', 'SQL', 'Search', 'Insert', 'Export', 'Import', and 'Operations'. Below this is a table structure for the 'products' table:

#	Column	Type	Collation	Attributes	Null	Default	Extra	Action
<input checked="" type="checkbox"/>	1 PRODUCTCODE	varchar(50)	latin1_general_cs		No			<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾
<input type="checkbox"/>	2 PRODUCTNAME	varchar(70)	utf8_unicode_ci		No			<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾
<input type="checkbox"/>	3 PRODUCTLINE	varchar(50)	latin1_general_cs		No			<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾
<input type="checkbox"/>	4 PRODUCTSCALE	varchar(10)	latin1_general_cs		No			<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾
<input type="checkbox"/>	5 PRODUCTVENDOR	varchar(50)	latin1_general_cs		No			<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾
<input type="checkbox"/>	6 PRODUCTDESCRIPTION	mediumtext	latin1_general_cs		No	None		<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾
<input type="checkbox"/>	7 QUANTITYINSTOCK	smallint(6)			No	0		<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾
<input type="checkbox"/>	8 BUYPRICE	decimal(17,0)			No	0		<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾
<input type="checkbox"/>	9 MSRP	decimal(17,0)			No	0		<input type="button" value="Change"/> <input type="button" value="Drop"/> More ▾

At the bottom of the interface are buttons for 'Check All / Uncheck All With selected' and several operation buttons: 'Delete', 'Change', 'Drop', 'Rename', 'Update', 'Insert', and 'Index'.

Observe the PRODUCTNAME column with the correct encoding.

Having this encoding set an arbitrary name was changed using the following statement:

```
UPDATE `products` SET `PRODUCTNAME`='The MayflowerA 僕僙儍僄僀' WHERE PRODUCTCODE='S700_1938'
```

The results are mixed. Some queries return the proper result. For example, when trying to filter on row labels one gets a view as follows:

The screenshot shows a 'PivotTable Fields' dialog box and a SharePoint ribbon interface. The dialog box has a 'Select field:' dropdown set to 'Line'. It contains options like 'Sort A to Z', 'Sort Z to A', 'More Sort Options...', 'Clear Filter From "Line"', 'Label Filters', 'Value Filters', and a 'Search Line' input field containing a tree view of items. The tree view includes items like 'Autoart Studio Design', 'Carousel DieCast Legends', 'Min Lin Diecast', 'Red Start Diecast', 'Studio M Art Models', 'The MayflowerA 僕僙儍僄僀', 'Unimax Art Galleries', 'Welly Diecast Productions', and 'Trains'. At the bottom are 'OK' and 'Cancel' buttons.

The SharePoint ribbon interface shows the same list of items under 'Choose fields to add to report:' in the 'Fields' section of the ribbon. The ribbon tabs include 'File', 'Home', 'Insert', 'Page Layout', 'List', 'Library', 'View', 'Format', 'SharePoint', and 'Help'.

The member with far eastern characters appears here.

Technical note: Using the XMLA connector with SharePoint 2013

However, when drilling in the pivot table to the member, the member does not appear:

Further analysis shows that a select like the following:

```
SELECT {AddCalculatedMembers({[Ispettore_D].Ispettore_H].[Ispettore].Members})} DIMENSION
PROPERTIES MEMBER_TYPE ON COLUMNS FROM [Vendite] CELL PROPERTIES CELL_ORDINAL
```

Correctly returns the member.

However, a select like the following:

```
SELECT NON EMPTY
Hierarchize(AddCalculatedMembers({DrilldownLevel({[Ispettore_D].Ispettore_H].[All Ispettori]}))) DIMENSION
PROPERTIES PARENT_UNIQUE_NAME ON COLUMNS FROM [Vendite] CELL PROPERTIES VALUE
```

Will not return a member.

This is interpreted as a bug in Mondrian.

There is however a workaround as follows.

The schema used till now had no captionColumn for the level where this member belongs:

```
<Level name="Product" table="PRODUCTS" column="PRODUCTNAME" type="String"
uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
```

In such a case Mondrian uses the name of the member for the display caption. As found before this is affected by a bug.

However, the table can be adjusted to accommodate the caption too. The figure below shows the adjusted table:

#	Column	Type	Collation	Attributes	Null	Default	Extra	Action
1	PRODUCTCODE	varchar(50)	latin1_general_ci		No			Change Drop More ▾
2	PRODUCTNAME	varchar(70)	utf8_unicode_ci		No			Change Drop More ▾
3	PRODUCTCAPTION	varchar(70)	utf8_unicode_ci		No	None		Change Drop More ▾
4	PRODUCTLINE	varchar(50)	latin1_general_ci		NO			Change Drop More ▾
5	PRODUCTSCALE	varchar(10)	latin1_general_ci		No			Change Drop More ▾
6	PRODUCTVENDOR	varchar(50)	latin1_general_ci		No			Change Drop More ▾
7	PRODUCTDESCRIPTION	mediumtext	latin1_general_ci		No	None		Change Drop More ▾
8	QUANTITYINSTOCK	smallint(6)			No	0		Change Drop More ▾
9	BUYPRICE	decimal(17,0)			No	0		Change Drop More ▾
10	MSRP	decimal(17,0)			No	0		Change Drop More ▾

And the entry corresponding to the member can be adjusted as follows:

PRODUCTCODE	PRODUCTNAME	PRODUCTCAPTION	PRODUCTLINE	PRODUCTSCALE	PRODUCTVENDOR	PRODUCTDESCRIPTION
S700_1938	The Mayflower	The MayflowerA 傑佛 僑佛	Ships	1:700	Studio M Art Models	Measures 31 1/2 inches Long x 25 1/2 inches High x...

In here, the product name remains an ASCII string, but the caption has Far Eastern characters.

Correspondingly, the schema should be adjusted as follows:

```
<Level name="Product" table="PRODUCTS" column="PRODUCTNAME" type="String"
uniqueMembers="true" levelType="Regular" hideMemberIf="Never"
captionColumn="PRODUCTCAPTION">
```

Observe the new attribute `captionColumn="PRODUCTCAPTION"` to the end. With this schema the result is correct in both cases (check pictures on the next page).

	A
1	Row Labels
2	+ Classic Cars *
3	+ Motorcycles *
4	+ Planes *
5	+ Ships *
6	+ Autoart Studio Design *
7	+ Carousel DieCast Legends *
8	+ Min Lin Diecast *
9	+ Red Start Diecast *
10	+ Studio M Art Models *
11	The MayflowerA 傑崔儕僄
12	+ Unimax Art Galleries *
13	+ Welly Diecast Productions *
14	+ Trains *
15	+ Trucks and Buses *
16	+ Vintage Cars *
17	Grand Total *
18	

1	Row Labels
Select field:	
Vendor	
A	Sort A to Z
Z	Sort Z to A
More Sort Options...	
Clear Filter From "Vendor"	
Label Filters	
Value Filters	
Search Line	
<input checked="" type="checkbox"/> Motorcycles <input checked="" type="checkbox"/> Planes <input checked="" type="checkbox"/> Ships <input checked="" type="checkbox"/> + Autoart Studio Design <input checked="" type="checkbox"/> + Carousel DieCast Legends <input checked="" type="checkbox"/> + Min Lin Diecast <input checked="" type="checkbox"/> + Red Start Diecast <input checked="" type="checkbox"/> + Studio M Art Models <input checked="" type="checkbox"/> + The MayflowerA 傑崔儕僄	
<input type="button" value="OK"/> <input type="button" value="Cancel"/>	

24 | 25