Automating URDU, ARABIC

8

Other RTL Scripting Documents

Using J-ISIS

Amjad Ali Malik

(ashtamid@gmail.com)

Islamabad, Pakistan

January 2015

Introduction to Urdu

Urdu (وثو (isten)) as a word is Turkish for 'foreign' or 'horde' and as a language, member of the Indo-Aryan group within the Indo-European family of languages. Urdu is spoken by more than 104 million speakers, including those who speak it as a second language, predominantly in Pakistan and India. It is the national language and lingua franca of Pakistan, and an official language of six Indian states. The Urdu language received recognition and patronage under the British Raj when the British replaced the Persian and local official languages of North Indian states with the Urdu and English language in 1837. Urdu is written right-to-left with a version of the Perso-Arabic script since the 12th century. It is normally written in Nastaliq style in an extension of the Persian alphabet, which is itself an extension of the Arabic alphabet. It has 39 characters, each then has two to four different shapes depending upon its position in the word, i.e. initial, medial, final or isolated. This characteristic of having multiple shapes makes it Context Sensitive.

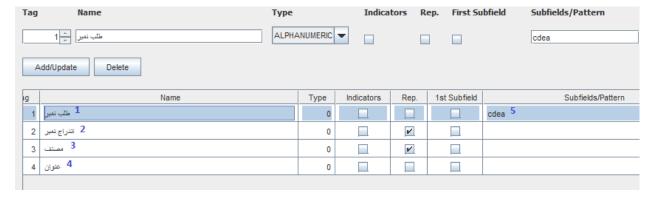
J-ISIS & Urdu

This document expresses the way, how can be a RTL language Database created & configured using <u>J-ISIS</u>. The article will help the <u>RTL language scripting communities</u> to automate their bibliographic library data from creation to deployment in easy sequential steps. <u>J-ISIS New DB</u> creation wizard will be used to create "URDU (RTL)" database with only four fields on Windows 8.1 x64 machine with English (United States) as default & URDU as second language through Control Panel Language option. For field headings URDU & for subfields delimiters English will be used.

Step-1. Database Creation

i. FDT (Field Definition Table)

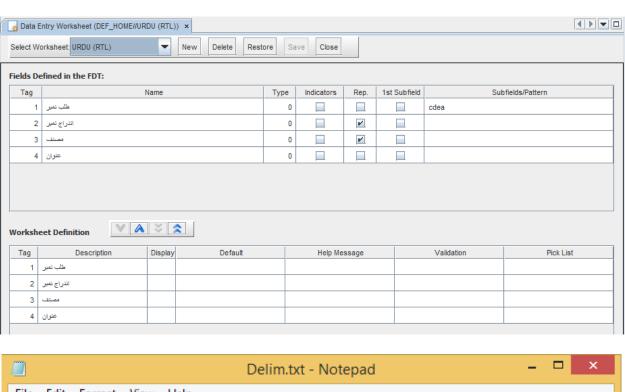
FDT will be created according to following table

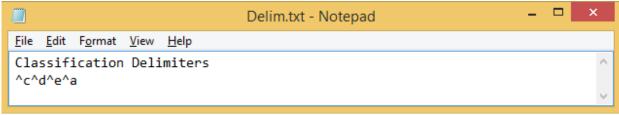


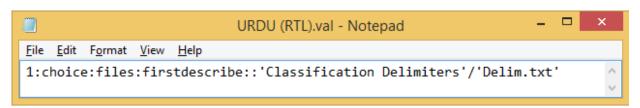
- 1- Classification Number
- 2- Accession Number
- 3- Author
- 4- Title
- cde subfield delimiters will be used to control and properly acclimate mixed notation Classification Numbers. (This will be explained thoroughly while data entry)

ii. Data Entry Worksheet

URDU (RTL) worksheet have been created with fields as defined in FDT. A pick list also designated in "jisis suite 16 September 2014\jisis suite\home example db\URDU (RTL)\jwks" named "URDU (RTL)\jwks" named (RTL).val" referring "delim.txt" to facilitate data entry process in طلب نمبر (classification Number) field by picking subfield delimiters.

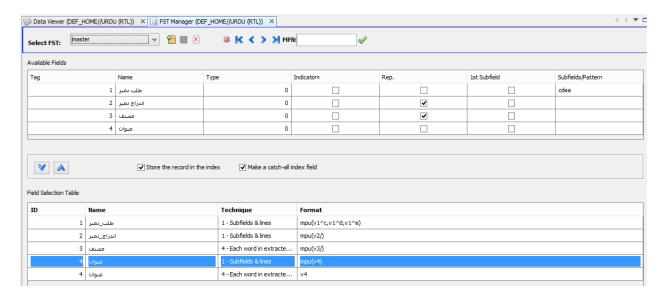






iii. FST (Field Select Table)

URDU (RTL) FST is designed as in subsequent table.



Indexing technique used for Tag ID # 1 (i.e. ללגי ביאע 'Classification Number') is "1 – Subfields & lines" which will index/Invert the field as a line & subfields separately. The format "mpu(v1^c,v1^d,v1^e)" will control the inverted data in the patron exercised. This format will authorize the J-ISIS to index the mixed notation classification numbers in structure as they were inserted in the data field using sequence of subfields. If we use only mpu(v1) or v1 here the data will not be indexed in required shape and formation due to the character layout problem. This problem happens due to the universal rule that numerals are always written from LTR, and system confuses to opt which direction when we use combination of both numerals & RTL alphabets to produce mixed notation classification numbers. For example 929 י (549.1) will be entered in data entry worksheet as ^c929 י (549.1), & will be indexed according to their occurrence position according to FST.

Step-2. Data Entry, Searching & Displaying

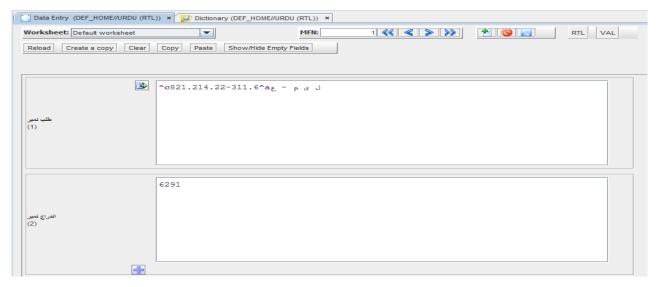
When we enter plain data in DB, it returns no problem while data insertion as well as in indexing/searching. Problem arises when data consists a combination of numerals and alphabets. Combination of numerals & alphabets are commonly used in classification numbers. Here in this step only one sub-field will be used for classification numbers.

i. Simple Data (Without RTL Button Pressed)

When we enter simple data in DB, it returns desired output while data insertion as well as in indexing/searching. Problem arose when data consists a combination of numerals and RTL alphabets. Combination of numerals & alphabets are commonly used in classification numbers. In Tag # 1 that will be used to store classification numbers for the documents both single & multiple sub-fielded techniques will be exercised to accumulate simple as well as mixed notation classification numbers.

Data Entry (Without RTL Button Pressed)

Data entry is as simple as it was in WinISIS. All the entries are done using RTL (URDU) keyboard.

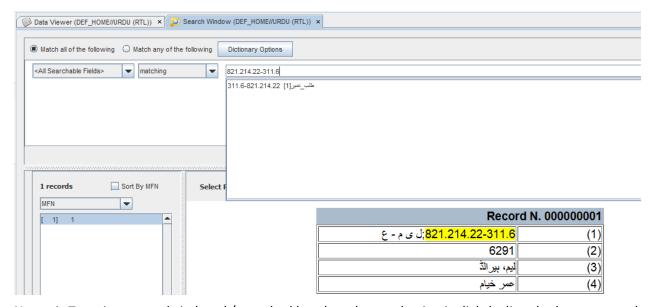


Terms indexed in Dictionary (Without RTL Button Pressed)

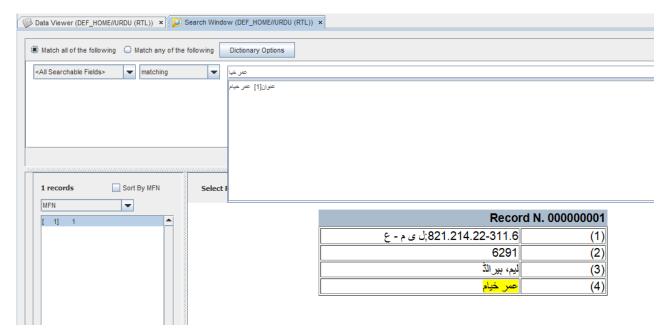
		Index Content	
iTerm	Field	Term	Freq
0	اندراج_نمبر	6291	1
1	طلب_نمبر	821.214.22-311.6	1
2	عنوان	خيام	1
3	عنوان	عمر	1
4	عنوان	عمر خيام	1
5	مصنف	ليم	1
6	المستف	بيرالأ	1

Figure 1Terms are accurately indexed i.e. Numerals from LTR & Alphabets from RTL

Searching in Terms Dictionary (Without RTL Button Pressed)



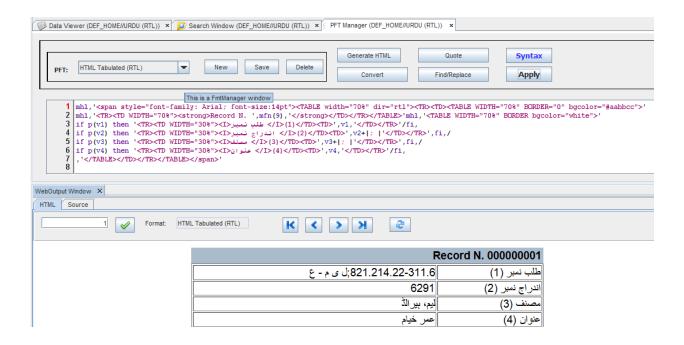
Numeric Term is accurately indexed / searched but drop down selection is slightly disturbed as compared to previous slide but according to RTL rule it is correct.



Urdu Term is accurately indexed / searched & drop down selection is absolutely correct.

PFT for Searched Term (Without RTL Button Pressed)

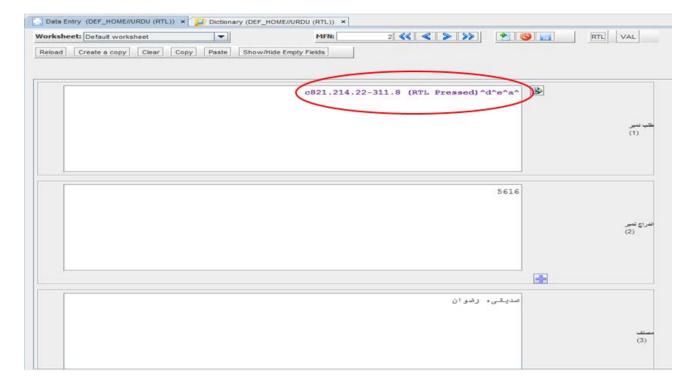
PFT Generated through PFT Manager & slightly modified for RTL Display



ii. Simple Data (with RTL button Pressed)

Data Entry (with RTL button Pressed)

Analyze the delimiters' position when data was entered in "^c". Delimiters got disturbed by their position.



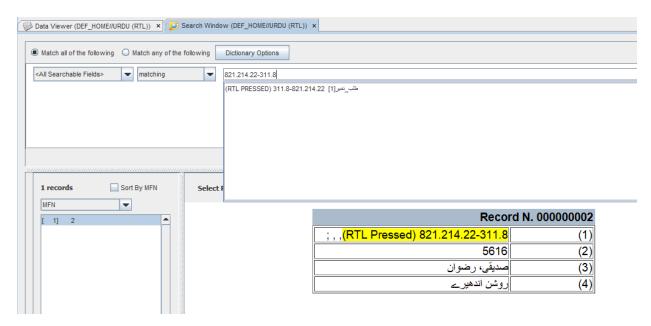
Terms in Dictionary (with RTL button Pressed)

Indexing is quite accurate although data entry GUI is disturbed

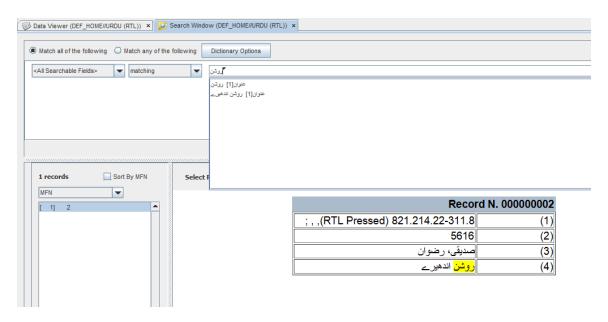


Searching in Terms Dictionary (with RTL button Pressed)

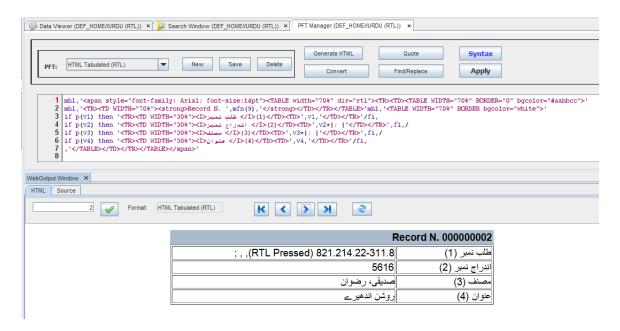
Term is accurately indexed / searched but drop down selection & Display is minor disturbed. According to RTL drop down suggestion is true.



Term is accurately indexed / searched & drop down selection is absolutely correct



PFT for Searched Term (With RTL Button Pressed)

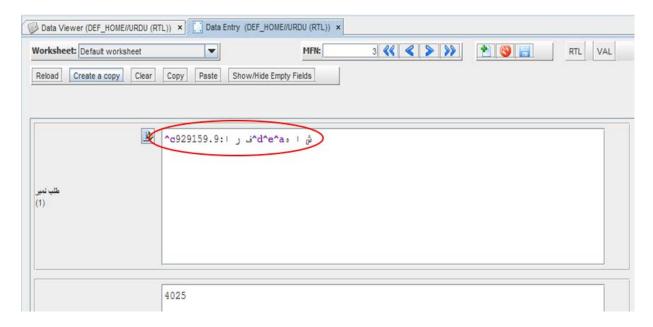


iii. Complex Data (Without RTL Button Pressed)

Complex data entry is being performed using both single & multiple subfields for classification number with mix notation (combination of numerals & characters). This entry will be done without using RTL button.

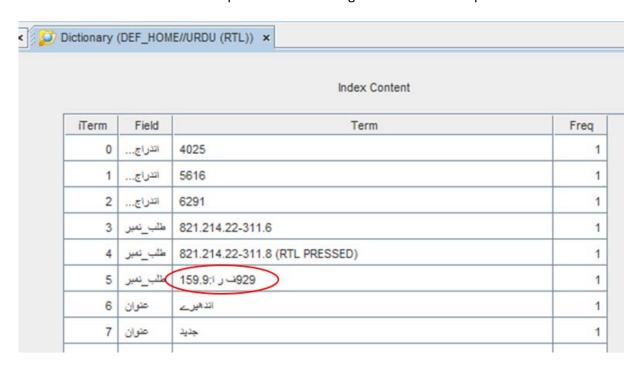
Data entry (using single subfield delimiter)

Data become disturbed when we use mixed notation classification with in one sub-field. Classification number "929" will be entered. Notice the sequence of the classification number from Left To Right (LTR) that is absolutely correct in MS Word.

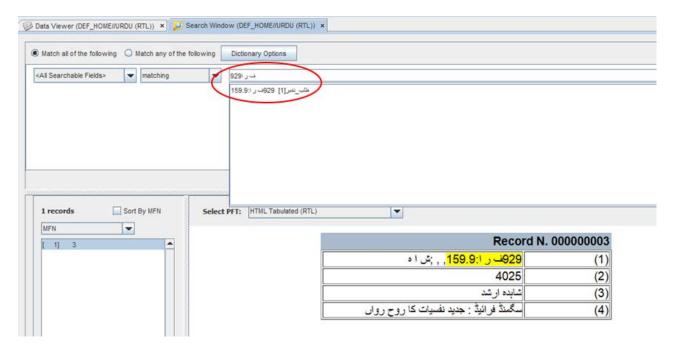


Term in Dictionary (using single subfield delimiter)

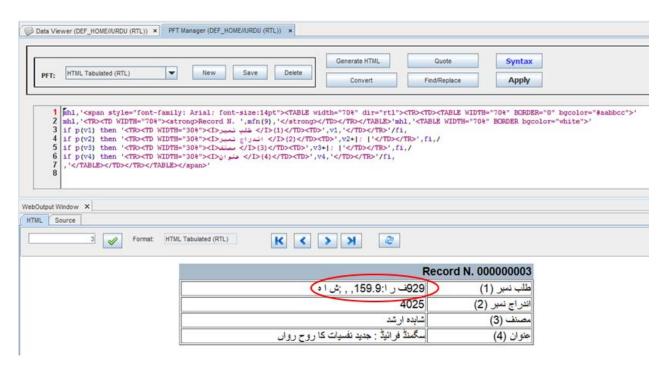
Term indexed is correct from search point of view starting from 929 but its sequence is RTL.



Searching in Terms Dictionary (using single subfield delimiter)



PFT & Display (using single subfield delimiter)

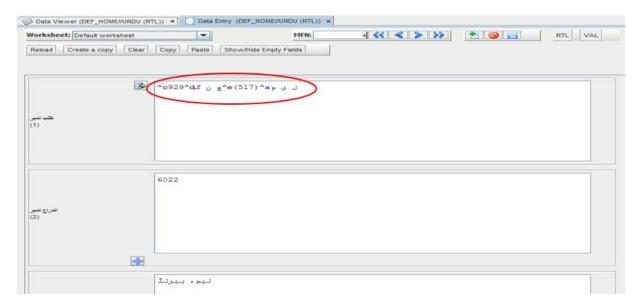


iv. Complex -- Controlled Data Entry & Display (Without RTL Button Pressed)

In this step, Mixed Notation Data as discussed earlier which is normally used in Classification Numbers will be automated using multiple subfields. Use of subfields enables the data entry as well as J-ISIS managing individuals to control the inserted data to their local library requirements.

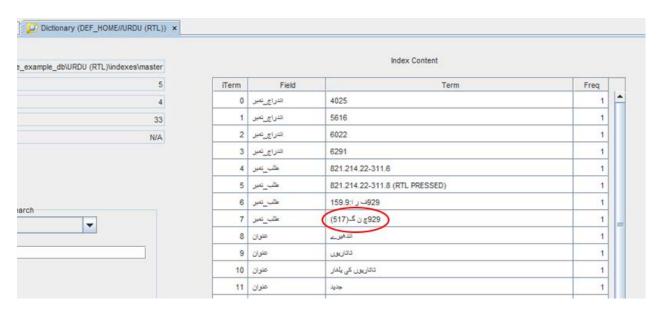
Data entry (using multiple subfield delimiters)

Here the data is being inserted using multiple subfield delimiters

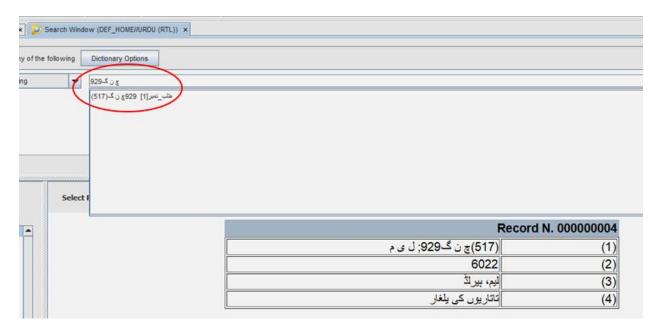


Terms in Dictionary (using multiple subfield delimiters)

Term's indexing is absolutely correct from Right To Left (RTL)

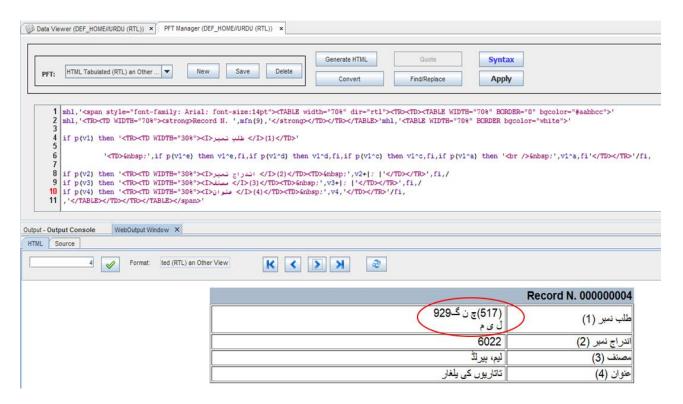


Searching in Terms Dictionary (using multiple subfield delimiters)



PFT & Display (using multiple subfield delimiters)

PFT is modified to display the data to meet local library requirements. Display of Tag#1, first line is from LTR & second line is from RTL.



Print Sort Examples

i. Sorting by Accession Numbers



Don't Confuse with the last two accession numbers. It is because that the field type for accession numbers is 'Alphanumeric' and it is being handled as alphabet, when we add leading alphabet followed by Zeros as "B00..' it will be sorted accurately.

ii. Sorting by Author



iii. Sorting by Title

D:\jisis_s	uite 15 December 2014\jisis_suite\\ 🔎 🔻 (Ø URDU (RTL) ×)	- □
س ا د ـ ا	(62) س ا د929		انورالسادات كي آپ بيتي	انورالسادات	684
ل ی م	(517) ج ن گ 929		تاتاریوں کی بلغار	ليم، بيراڭ	6022
	821.214.22-311.8 (RTL Pressed)		روسَن اندھیرے	صديقي، رضوان	5616
آس ا ہ	ف ر ا:929159.9		سگمنڈ فرائیڈ : جدید نفسیات کا روح رواں	شابده ارشد	4025
ع ق ی - ع	891.433: م ف ت92		على يور كا مفتى	احمد عقیل روبی	5385
ل ی م - ع	821.214.22-311.6		عمر خيام	ليم، بيرالة	6291
غزا-ل	297(091)		لا تَقرَّقُو	غزالہ اظہر	6217
خ ل ی - م	32: اب و 92		مولانا ابوالكلام آزاد: شخصیت اور كارنامے	انجم، خليق	1892
ق ا ض - ج	(512) ج ن گ929		چنگیز خان	فاضمى نورالله	688
				,	

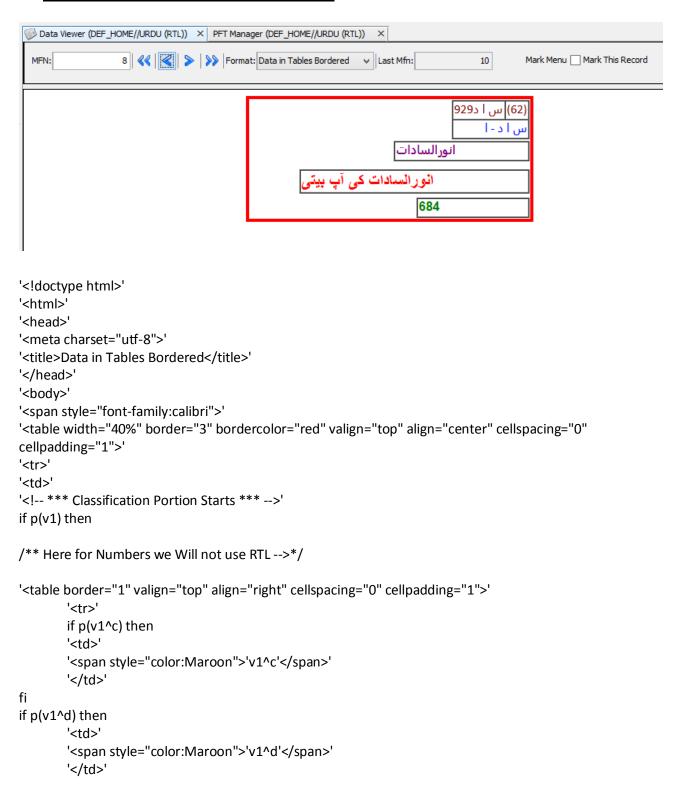
PFT Used for Print Sort Output

```
'<!doctype html>'
'<html>'
'<head>'
'<meta charset="utf-8">'
'<Title> Print Sort List </Title>'
'</head>'
'<body>'
''
''
//** Author Mark */
'<div align="right" dir="RTL">'
if p(v1^a) then v1^a, fi
'</div>'
//** Classification Number */
'<div align="right">'
''
if p(v1) then
''
     if p(v1^c) then
     ''
          v1^c
     ''
if p(v1^d) then
     ''
     v1^d
```

```
''
if p(v1^e) then
      ''
      v1^e
      ''
fi
      ''
fi
''
'</div>'
//** Title */
'<div align="right">'
if p(v4) then v4, fi
'</div>'
//** Author */
'<div align="right">'
if p(v3) then v3+|; |,fi
'</div>'
//** Accession Number */
'<div align="right">'
if p(v2) then v2+|; |, fi
'</div>'
''
''
'</body>'
'</html>'
```

More PFTs for Customized Output

i. Tabulated Colorful Output with Borders



```
fi
if p(v1^e) then
     ''
     '<span style="color:Maroon">'v1^e'</span>'
     ''
fi
     ''
if p(v1^a) then
     ''
/** Here for Alphabets we Will use RTL -->*/
     ''
     '<span style="color:blue">'v1^a'</span>'
     ''
     ''
''
'<!-- *** Classification Portion ends *** -->'/
' '
'<!-- *** Author Portion Starts *** -->'/
if p(v3) then
''
          ''
                '<span style="font-family:calibri; color:purple; font-size:12pt; font-
weight:bold">'x25,v3+|; |'</span>'
          ''
''
'<!-- *** Author Portion ends *** -->'/
' '
'<!-- *** Title Portion Starts *** -->'/
if p(v4) then
''
          ''
                '<span style="font-family:arial; color:#FF0000; font-size:16pt; font-
weight:bold">'x20,v4'</span>'
          ''
''
'<!-- *** Title Portion ends *** -->'/
' '
'<!-- *** Accession Number Portion Starts *** -->'/
if p(v2) then
''
          ''
```

```
'<span style="font-family:arial; color:Green; font-size:12pt; font-weight:bold">'x25,v2+|; |''

''

''
fi
'<!-- *** Accession Number Portion ends *** -->'/
''
''
''
''
''
''
'</span>'
'</body>'
'</html>'
```

ii. Tabulated Colorful Output without Borders



```
'<|doctype html>'
'<html>'
'<html>'
'<head>'
'<meta charset="utf-8">'
'<title>Tables Without Border</title>'
'</head>'
'<body>'
'<span style="font-family:calibri">'
''
'' '<tt>'<tt>'<tt>'<'t-' *** Classification Portion Starts *** -->'/
if p(v1) then

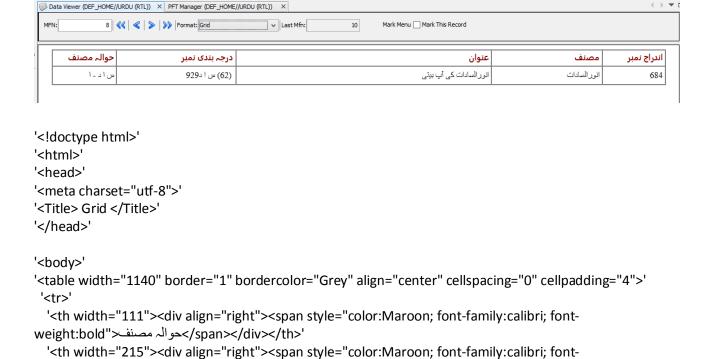
/** Here for Numbers we Will not use RTL -->*/
```

```
''
     ''
     if p(v1^c) then
     ''
     '<span style="color:Maroon">'v1^c'</span>'
if p(v1^d) then
     ''
     '<span style="color:Maroon">'v1^d'</span>'
     ''
if p(v1^e) then
     ''
     '<span style="color:Maroon">'v1^e'</span>'
     ''
fi
     ''
if p(v1^a) then
     ''
     /** Here for Alphabets we Will use RTL -->*/
     ''
     '<span style="color:blue">'v1^a'</span>'
     ''
     ''
''
'<!-- *** Classification Portion ends *** -->'/
' '
'<!-- *** Author Portion Starts *** -->'/
if p(v3) then
''
          ''
               '<span style="font-family:calibri; color:purple; font-size:12pt; font-
weight:bold">'x25,v3+|; |'</span>'
          ''
''
'<!-- *** Author Portion ends *** -->'/
' '
'<!-- *** Title Portion Starts *** -->'/
if p(v4) then
''
          ''
```

```
'<span style="font-family:arial; color:#FF0000; font-size:16pt; font-
weight:bold">'x20,v4'</span>'
           ''
''
'<!-- *** Title Portion ends *** -->'/
' '
'<!-- *** Accession Number Portion Starts *** -->'/
if p(v2) then
''
           ''
                 '<span style="font-family:arial; color:Green; font-size:12pt; font-
weight:bold">'x25,v2+|; |''
           ''
''
'<!-- *** Accession Number Portion ends *** -->'/
''
''
''
'</span>'
'</body>'
'</html>'
```

iii. Tabulated Output with Headings

/span></div>



```
'</span></div>
 '<div align="right"><span style="color:Maroon; font-family:calibri; font-
/span></div>
 '<div align="right"><span style="color:Maroon; font-family:calibri; font-
"span></div>اندراج نمبر<"weight:bold
''
''
//** Author Mark */
'<div align="right" dir="RTL">'
if p(v1^a) then v1^a, fi
'</div>'
//** Classification Number */
'<div align="right">'
''
if p(v1) then
''
      if p(v1^c) then
      ''
           v1^c
      ''
fi
if p(v1^d) then
      ''
      v1^d
      ''
fi
if p(v1^e) then
      ''
      v1^e
      ''
fi
      ''
fi
''
'</div>'
//** Title */
'<div align="right">'
if p(v4) then v4, fi
'</div>'
//** Author */
'<div align="right">'
if p(v3) then v3+|; |,fi
'</div>'
```

'<div align="right"><span style="color:Maroon; font-family:calibri; font-

```
//** Accession Number */
'<div align="right">'
if p(v2) then v2+|; |, fi
'</div>'
''
'blue>'
'</body>'
'</html>'
```

iv. Tabulated Output for Generating Lists



```
'<!doctype html>'
'<html>'
'<head>'
'<meta charset="utf-8">'
'<Title> Print Sort List </Title>'
'</head>'
'<body>'
''
''
//** Author Mark */
'<div align="right" dir="RTL">'
if p(v1^a) then v1^a, fi
'</div>'
//** Classification Number */
'<div align="right">'
''
if p(v1) then
''
     if p(v1^c) then
     ''
          v1^c
     ''
if p(v1^d) then
     ''
     v1^d
```

```
''
if p(v1^e) then
      ''
      v1^e
      ''
fi
      ''
fi
''
'</div>'
//** Title */
'<div align="right">'
if p(v4) then v4, fi
'</div>'
//** Author */
'<div align="right">'
if p(v3) then v3+|; |,fi
'</div>'
//** Accession Number */
'<div align="right">'
if p(v2) then v2+|; |, fi
'</div>'
''
''
'</body>'
'</html>'
```

v. Generating Stock List

Stock List covers status of all Accession Numbers automated in the DB. Multiple occurrences of Accession # are breakup to become on the separate line or in the Row & Title is replicated with every occurrence of Accession.

```
(ATL)
                                                                                       ☆ ★ ☆
                                       مولانا ابوالكلام أزاد: شخصيت اور كارنامـــ
                                                                                      1892
                                         مكمنة فرائية : جديد نفسيات كا روح روال
                                                                                      4025
                                                            على يور كا مفتى
                                                                                      5385
                                                             روشن اندھیرے
                                                                                      5616
                                                            تاتاريوں كى بلغار
                                                                                      6022
                                                                   لا تَقرَّفُو
                                                                                      6217
                                                                  عمر خيام
                                                                                      6291
                                                      انورالسادات کی آپ بیتی
                                                                                        684
                                                                 جِنگيز خان
                                                                                        688
                                                   طب اسلامي كا انسائيكلوپيڻيا
                                                                                   B000001
                                                   طب اسلامي كا انسائيكلوييثيا
                                                                                   B000002
                                                   طب اسلامي كا انسائيكلوپيڻيا
                                                                                   B000003
                                                   طب اسلامي كا انسائيكلوپيڻيا
                                                                                   B000004
                                                   طب اسلامي كا انسائيكلوپيڻيا
                                                                                   B000005
```

```
'<!doctype html>'
'<html>'
'<head>'
'<meta charset="utf-8">'
'<Title> Stock List </Title>'
'</head>'
'<body>'
      ''
      ''
            if p(v2) then,
            '<!-- Loop is being used to break more than one occurrences of Accession # field-->'/
            e1:=1, e2:=nocc(v2), while e1<=e2
            ''
            '<span style="font-family:calibri">'
            if p(v2[e1]) then, ,v2[e1],'</span><span style="font-
family:calibri">',v4, fi,'<br />'
            '<!-- Loop is being incremented to jump over to next Accession#-->'/
            e1:=e1+1,
            '</span>'
            ''
```

```
fi,
''
''
'</body>'
'</html>'
```

References:-

Information about the Urdu

- 1. http://www.britannica.com/EBchecked/topic/619612/Urdu-language
- 2. http://en.wikipedia.org/wiki/Urdu
- 3. http://www.bbc.co.uk/voices/multilingual/urdu.shtml

Information about the Urdu script

- 1. http://en.wikipedia.org/wiki/Urdu alphabet
- 2. http://www.bbc.co.uk/languages/other/urdu/guide/alphabet.shtml
- 3. http://www.omniglot.com/writing/urdu.htm

Information about Right-to-Left

1. http://en.wikipedia.org/wiki/Right-to-left

Information about J-ISIS & Downloads

- 1. https://kenai.com/projects/j-isis
- 2. https://kenai.com/projects/j-isis/downloads