Different Ways of Displaying Non-Latin Character Sets in Blaise

Leif Bochis, Statistics Denmark

Abstract

Blaise offers excellent support for multi-language interview instruments allowing the interviewer to select interview language on the fly. While this is easily managed with European languages using characters supported by the standard ANSI character set, there are a number of problems to address when using non-Latin character sets - for example, Arabic script.

This paper addresses certain problems concerning preparation of documents in non-Latin scripts and how to incorporate these as field texts in Blaise instruments - including problems of converting texts from word processing format into the simple text file format supported by Blaise.

As new versions of Blaise offer still more new facilities in general, also new ways to make non-Latin characters displayable in Blaise instruments is introduced - through separate display software, as field texts using special character sets and fonts or as images.

The aim of this paper is to describe a range of possible solutions and to discuss advantages and drawbacks of each solution in existing and (close) future versions of Blaise.

This paper elaborates further on a paper presented at the 6th Blaise Users' Conference in Kinsale 2000 and discusses advantages and drawbacks of the different solutions exemplified by Farsi text, based on work with incorporating Farsi field texts for Immigrant Surveys carried out by Statistics Denmark 1998-2001.

What is it all about?

Statistics Denmark has carried out a number of Immigrant surveys over the last 3 years concerning living conditions, integration on the labour market, language skills et cetera. Among them a major survey in 1998-99, which was repeated – slightly modified – in 2000-2001.

The questions concerning language skills implied that the respondents should be interviewed in Danish if possible, otherwise in their own language if applicable, or in English as a third alternative. The instrument should therefore be able to provide question texts in a range of different languages.[i]

The surveys were designed as multi-language surveys and made use of the language facility of Blaise in the implementation in order to enable the interviewers to change language on the fly. The questionnaire was for the purpose translated from Danish into the languages English, Polish, Serbo-Croatian, Turkish, Somali, Arabic, Farsi, Urdu and Vietnamese.

Six of the languages are written with (a variant of) the Latin alphabet, which caused no trouble in the implementation, the latter four languages, however, are written with either Arabic script (Arabic, Farsi, Urdu) or a widened version of Latin (Vietnamese). The efforts made to get the texts incorporated in the Blaise instrument were concentrated on the Farsi version.

What is the Problem?

The solutions should be developed to work as CATI applications and run in the Danish (i.e. standard Western) version of Windows NT – through the period in which Blaise versions from Blaise III over Blaise 4.1 to Blaise 4.3 were upgraded.

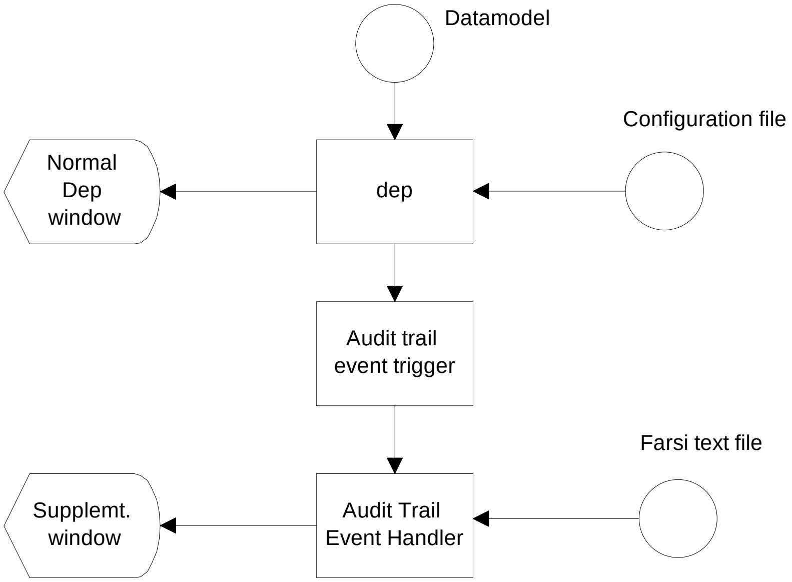
With the Dos version (Blaise III) the options were generally limited by the operating system. Some local versions of Dos supported an Ascii codepage including a local script – also with right to left orientation – which made it possible to switch between, for example, Latin and Arabic or Latin and Hebrew, but in practice it would hardly have been possible to develop Blaise-instruments using more than one non-Latin alphabet. At least, while we were running Windows NT as operating system, we experienced that it was difficult to apply additional codepages for Dos applications.

When Blaise entered the Windows world many more features were added allowing, for example, the use of more than one font – The early versions, however, were based on the OEM character set in order to keep compatibility between the Dos and Windows versions. To accomplish, this all data and metadata were stored in OEM format and automatically converted to ANSI when displayed in the Windows version, causing that the use of special fonts to display foreign alphabets were practically impossible.

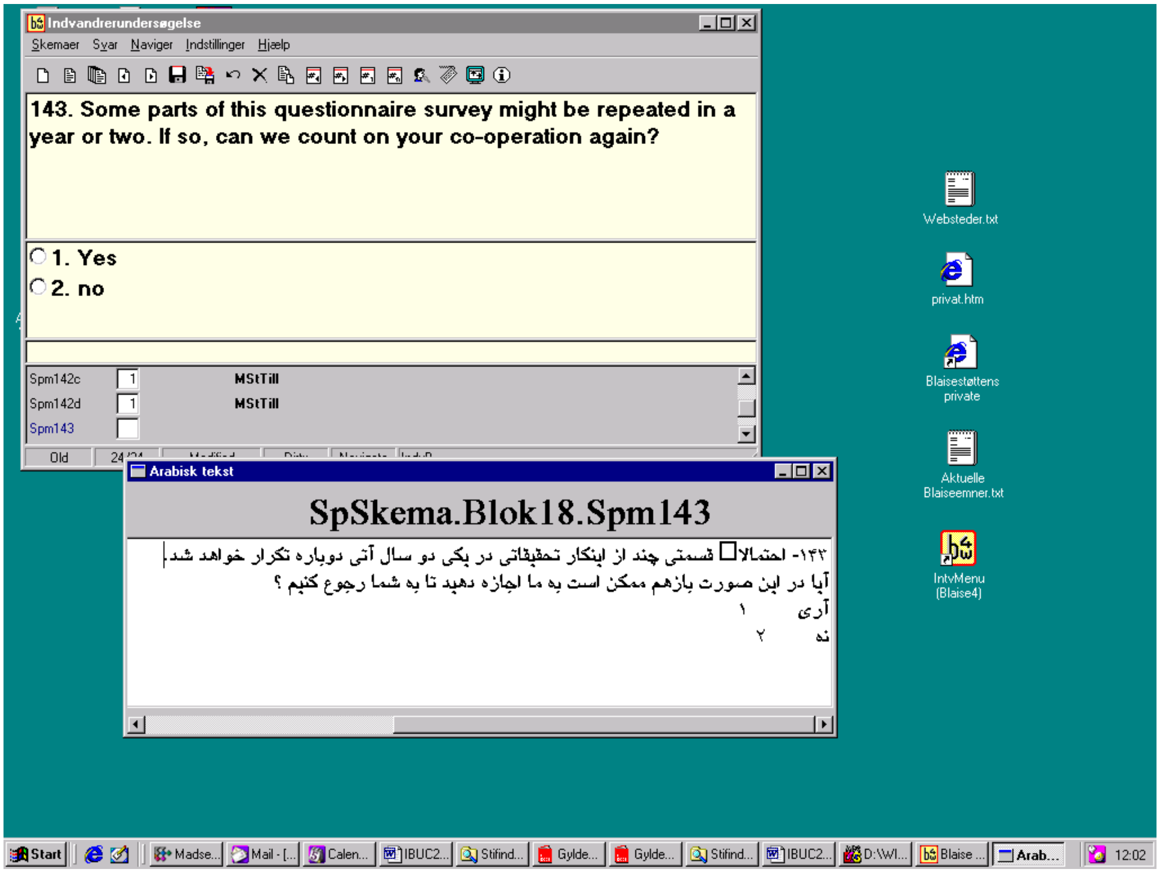
The use of Windows, however, made calls to external routines and programs far easier, and the introduction of the audit trail mechanism made it possible to let different kinds of displaying take place in separate systems but still controlled from Blaise.

Audit Trail Solution

The Blaise 4 Windows Audit Trail mechanism triggers the relevant events during the interview such as the AuditTrailEnterField event, which is triggered each time a field is entered and passes the relevant information (Field Name, for example) to the Audit Trail Event Handler.



For the 1998-survey a special version of the Audit Trail Event Handler was developed in order to manage the communication between the Blaise instrument and a supplementary display window. This Audit Trail Event Handler then looks up the proper question text and displays it through the Supplementary Window Control.



143. Some parts of this questionnaire survey might be repeated in a year or two. If so, can we count on your co-operation again?

○ 1. Yes
○ 2. no

Spm142c	1	MStTill
Spm142d	1	MStTill
Spm143		

Old 24/24

Arabisk tekst

SpSkema.Blok18.Spm143

۱۴۳- احتمالا□ قسمتی چند از اینکار تحقیقاتی در یکی دو سال آتی دوباره تکرار خواهد شد.
آیا در این صورت بازهم ممکن است به ما اجازه دهید تا به شما رجوع کنیم ؟
آری ۱
نه ۲

This Audit Trail Event Handler was written through a slight modification of the Audit Trail example, delivered with the Blaise system and is described in detail in ref. 1. This system was made ready in the last phase of the 1998-survey and worked reasonably well in testing, though it wasn't used in production.

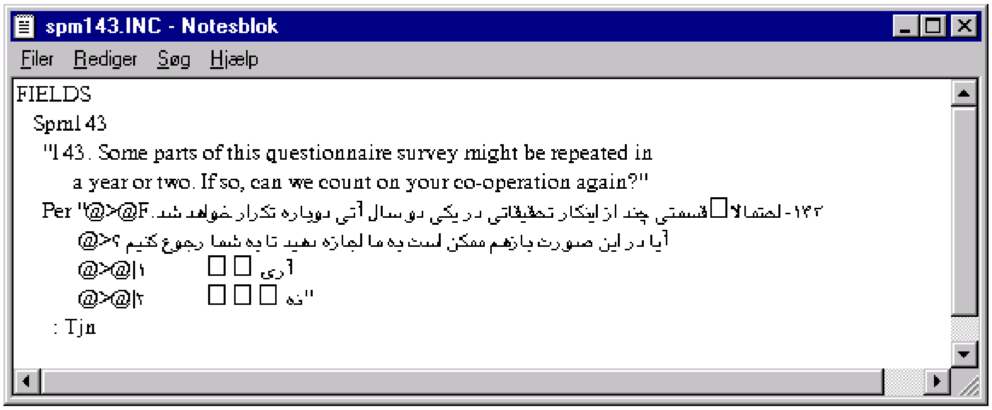
Blaise 4 Ansi-solution

With version 4.3 Blaise became ANSI-based, and oem-ansi conversions were no longer needed. Thus, it became possible to use the built-in facility to define special fonts to use in (parts of) field texts – for example, to represent other scripts than Latin – provided they could be represented by a single-byte character set.

Certain considerations, however, should to be addressed when, for example, Arabic script needs to be displayed in Blaise Field texts.

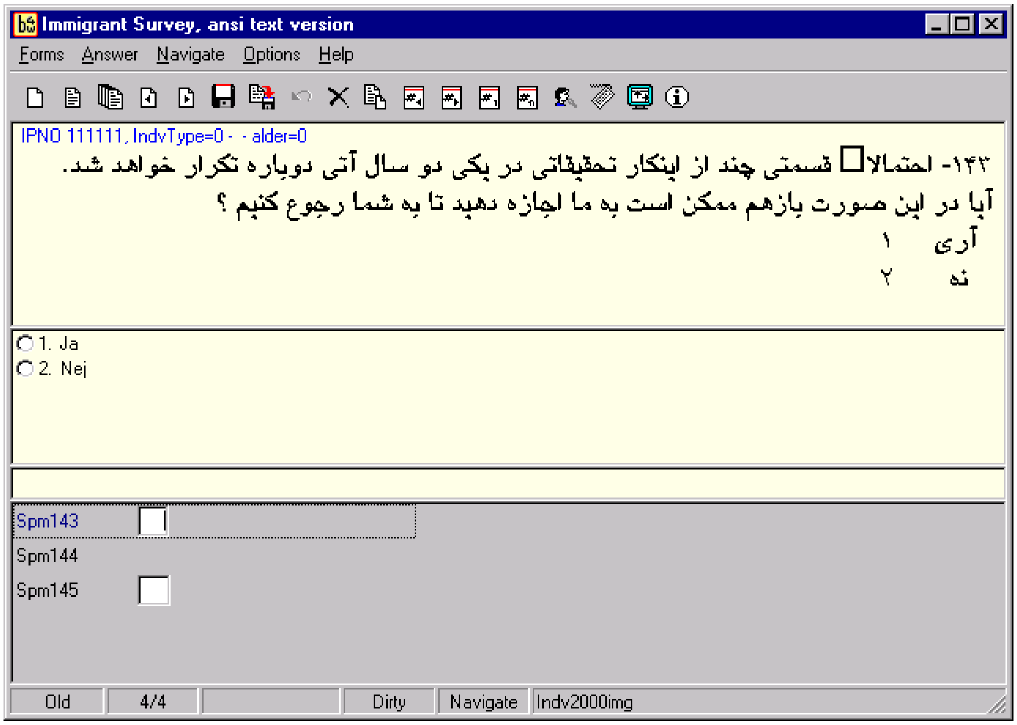
Arabic is written from right to left (hereafter referenced *orientation*), is aligned to the right (*alignment*) and the shape of the letters is dependent of where the letter is positioned in the word (*presentation form*).

These properties are not handled by the Blaise Editor, and consequently the translators will need a word processing tool to work out the translation and afterwards there will be a need to convert the outcome into a plain text format that may be inserted into the Blaise datamodel source.



```
FIELDS
  Spm143
   "143. Some parts of this questionnaire survey might be repeated in
      a year or two. If so, can we count on your co-operation again?"
   Per "@>@F۱۴۳- احتمالا□قسمتي چند از اينکار تحقيقاتي در يکي دو سال آتي دوباره تکرار خواهد شد.
      @>آيا در اين صورت بازهم ممکن است به ما اجازه دهيد تا به شما رجوع کنيم ؟
      @>@|۱        □□ آری
      @>@|۲        □□□ نه"
    : Tjn
```

@F here denotes the use of Farsi (Persian) Font, while **@>**, **@|** and hard spaces take care of alignment and spacing for better readability.


Immigrant Survey, ansi text version
Forms Answer Navigate Options Help

IPNO 111111, IndvType=0 - - alder=0

۱۴۳- احتمالا□ قسمتی چند از اینکار تحقیقاتی در یکی دو سال آتی دوباره تکرار خواهد شد.
آیا در این صورت بازهم ممکن است به ما اجازه دهید تا به شما رجوع کنیم ؟
آری ۱
نه ۲

○ 1. Ja
○ 2. Nej

Spm143 □
Spm144
Spm145 □

Old	4/4		Dirty	Navigate	Indv2000img

As the Blaise field text display cannot by itself handle orientation, alignment, presentation forms and line shifts for Arabic text, it is important that all these properties are generated during the conversion process.

All this, however, implies that it is very important to state a set of requirements concerning the format of the delivered translations in order to achieve that the format is actually either possible to apply directly in the Blaise source, or that it is possible to convert for that purpose.

Sparse attempts based on the delivered Farsi text from the 1998-survey showed that this should be possible to carry out in practice in the next survey. So in 1999 we were quite optimistic and thought that "next time we are ready!"

Discourse: Why We Were Not Ready!

What complicated this approach, however, is the nature of a constantly changing world. What worked fine as a conversion program two years ago didn't work in 2001 because there was no program to produce the same kind of output. Though the same person who did the 1998-translation should also do the 2000-translation of the questionnaire, the Farsi Word processor that was used in 1998 was not available. Instead, Microsoft Word 2000 was used to produce a Farsi version of the revised questionnaire – primarily on the basis of a plain text file that was used in the Audit Trail solution for the 1998-version.

Word 2000 is actually capable of processing Farsi text, i.e. Word 'knows' what part of a certain document that is in Farsi, English or whatever of a wide range of supported languages with their respective scripts. The letters are stored in the Unicode double byte character set and displaying as well as printing is controlled by Word concerning orientation, default alignment and actual presentation form.

The translator therefore decided to use Word to write his translation and passed the resulting Word document on for further processing.

In order to pass it over to a plain text format that could be inserted in Blaise field texts, the Word document was converted to a Unicode text file, which in turn could be converted into a plain text file with a single byte character set (with respect to orientation, presentation forms etc.) that might be displayed with a True Type font that could be used in Blaise field texts.

Unfortunately, the delivered Word document for a greater part consisted of the translation from the 1998-version, which was copied into Word from a plain text file and – supplied with the proper font and alignment – looked rather good in display and print. Word, however, was not 'aware' that it actually was Farsi text but 'thought' it was Danish with respect to Unicode-representation, orientation etc. The resulting document turned out to be quite a mix of 'Farsi' and 'quasi-Farsi' text.

Though we were ready to receive a Farsi text and had selected a suitable font for the purpose, we didn't manage to get the received text to fit into the text format expected.

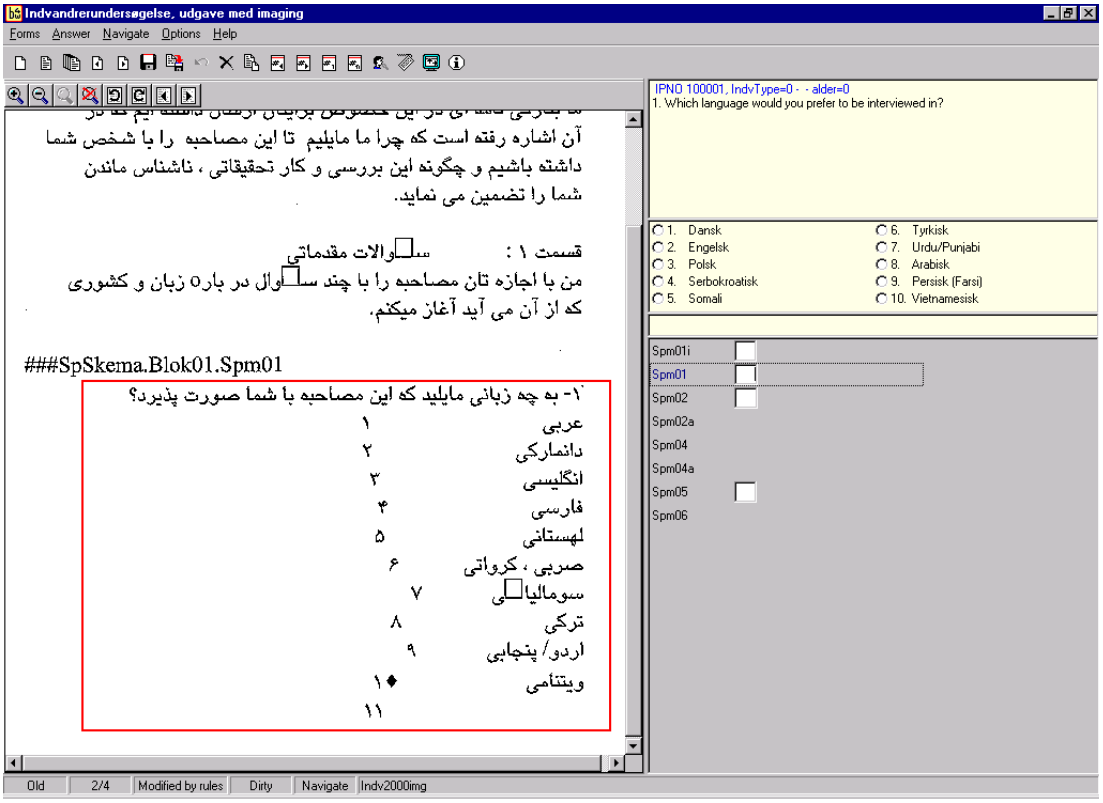
Imaging

With Blaise 4.5 and BCP a quite new way to get across the conversion and editing problems were introduced – the imaging feature. As explained in the documentation this feature is originally intended for displaying scanned paper forms in a data editing process in order to ease data editing - and probably reduce the amount of paper and burden of paper handling.

This feature makes possible a general way to show graphics on the screen and at the same time relates parts of the graphics to specific fields of a datamodel. In this way it is possible to show field texts (eventually combined with graphics) in any script of the world - provided only that it may be presented on plain paper. And – most important of all – it can be done

- Without the need to bother how the script is represented on the screen
- Without writing or purchasing third party software to handle the presentation
- Without the need to specify any file format in which the translators should deliver, and
- Without needing to convert from one file format to another.

You might even make a template that defines how much size should be used for each field (because a long field text in Farsi probably also would be a long field text in Chinese, Arabic or any other language) – and thus make a general solution for all language versions involved in the survey.


Indvandrerundersøgelse, udgave med imaging
Forms Answer Navigate Options Help
Old
2/4
Modified by rules
Dirty
Navigate
Indv2000img

آن اشاره رفته است که چرا ما مايليم تا اين مصاحبه را با شخص شما
داشته باشيم و چگونه اين بررسي و كار تحقيقاتي ، ناشناس ماندن
شما را تضمين مي نمايد.

قسمت ١ :   س□والات مقدماتي
من با اجازه تان مصاحبه را با چند س□وال در بارo زبان و كشوري
كه از آن مي آيد آغاز ميكنم.

###SpSkema.Blok01.Spm01

١- به چه زباني مايليد كه اين مصاحبه با شما صورت پذيرد؟

عربي	١
دانماركي	٢
انگليسي	٣
فارسي	۴
لهستاني	۵
صربي ، كرواتي	۶
سومالیا□ی	٧
تركي	٨
اردو/ پنجابي	٩
ويتنامي	١◆
	١١

IPNO 100001, IndvType=0 - - alder=0
1. Which language would you prefer to be interviewed in?

○ 1. Dansk
○ 2. Engelsk
○ 3. Polsk
○ 4. Serbokroatisk
○ 5. Somali
○ 6. Tyrkisk
○ 7. Urdu/Punjabi
○ 8. Arabisk
○ 9. Persisk (Farsi)
○ 10. Vietnamesisk

Spm01i □
Spm01 □
Spm02 □
Spm02a
Spm04
Spm04a
Spm05 □
Spm06

It is not possible, however, to change to a quite different language (for example, from Farsi to Chinese) right on the fly and continue interviewing, as the graphics files are read by the program upon each entry of a new form. You will need to redefine the respondents preferred language, make an appointment and save the form. When the appointment comes forward next time, the images will then show the Chinese version – but probably another interviewer is needed also in this situation.

Take an example:

- The interviewer calls the respondent, and finds out that though an immigrant from Iran the respondent appears to be Chinese and wants to be interviewed in Chinese.

- The interviewer fills in the answer that the preferred language is Chinese.

- The program then fills in the proper values – names of files with the scanned texts – in the Imaging Management fields.

- The interviewer makes an appointment, and transfers to a Chinese-speaking interviewer.

Other solutions

An alternative graphical representation should be mentioned. It is possible – and has been since the first Windows version of Blaise – to display graphics also as bitmaps in the field pane of a Blaise instrument through the use of a multi-media language. This requires the preparation of individual bitmaps for each question and with a large questionnaire that could be quite a tedious job. For that reason, we haven't followed that trail in our search for possible solutions. On the other hand, there is a number of advantages of the solution. Because the name of the bitmap files may be defined in run time, it is possible to change the non-Latin-based language right on the fly. This also allows the preparation of slightly different versions of the questions, although that implies an increase of the preparatory work to be done.

Conclusions

With the newer versions of Blaise, still more features are added allowing new ways to work out solutions to problems concerning how to provide question texts in non-Latin scripts.

First, the Audit trail mechanism in Blaise 4.0 – 4.2 allowing events during interviewing to control other programs, and thus make it possible to control displaying from Blaise that Blaise cannot display by itself. The solution suffers from the fact that text substitution in field texts cannot be done unless you are willing to develop tailored Audit Trail Event Handlers for each instrument that should be deployed. On the other hand, it allows the use of third-party software and standards like Unicode, which might ease the preparatory work.

Second, the ANSI-based versions 4.3 – 4.4 with the possibility to define special fonts for other scripts (without the need to worry about oem-ansi conversions) and then make it possible to display any single byte character set, that can be presented through a True Type font.

And third, the imaging feature of Blaise 4.5 and BCP that makes it possible to display the text as graphics alongside the ordinary field text.

The Imaging solution also suffers from the fact that text substitution in field texts is not possible – on the other hand, the solution supports that the ordinary field text is displayed alongside and here supplies the interviewer with any additional information.

If text-substitution in field texts is needed there is no realistic alternative to the built-in language facility – this facility, however, is only well suited for languages written with the Latin alphabet (and is probably too problematic to apply in practice for a multi-language instrument).

Both the Audit Trail solution and the inclusion of non-Latin alphabets in the Blaise field texts suffer from the fact that the conversion from a given script into a suitable simple text file format – and eventually later editing – might be quite a cumbersome project, though the Audit Trail solution does open up more possibilities, for example, the non-Latin text does not have to fit into a single byte character set.

The imaging feature of Blaise 4.5 / BCP is without competition the fastest and easiest way to incorporate non-Latin scripts in a Blaise questionnaire as almost all concerns about conversion and presentation can be avoided.

In the future – as Internationalization emerges – coming versions of Windows will probably support the use of non-Latin scripts better – for example, through the consistent use of Unicode and support of controls able to display Unicode-characters and take care of matters like orientation, alignment, presentations forms etc. And Blaise eventually might support Unicode controls for field text displaying.

As long as this is not the current state-of-art, other solutions must be applied. Blaise still offers a variety of ways to do this.

References

Leif Bochis: *Audit Trails or How to Display Arabic Text in Blaise*, in: Proceedings of the 6th International Blaise Users' Conference, Kinsale 2000, available at www.blaiseusers.org

Blaise documentation, *Audit Trail*, in: Blaise 4.1 Developers' Guide, Chapter 5.6 Audit Trail, pp. 282-293, *Showing Form Images in the DEP*, in: On line documentation for Blaise 4.5

[i] Please note that it has only been the ambition to provide question texts in non-Latin scripts. It is not the intention of this paper to discuss the possibility of changing the whole user interface, including typing in answers to questions in other languages than what is default in Statistics Denmark.