# *Database Specification*

## Overview

- Over 2200 form-pages of (946) Tunisian town/village names
- All data are digitized using a resolution of 300 dpi (b/w)
- Ground truth information available
  - e.g. sequence of Arabic character shapes
  - baseline/reference line position
- A wide variety of writing styles; 411 different writers
- About 26000 Arabic handwritten Tunisian town/village names
- Approximately 212 000 Arabic characters and ligatures
- Divided into 4 sets (a-d)
- Images and ground truth documentation included

## Set specification

The whole database is divided into 4 disjoint sets for training and testing Arabic OCR systems. We recommend to use the specified sets for comparability with results of other groups.

| SET | Number of words | Number of writer (OK+bad) |
|---|---|---|
| a | 6537 | 88+14=102 |
| b | 6710 | 89+13=102 |
| c | 6477 | 88+15=103 |
| d | 6735 | 90+14=104 |
| SUM | **26459** | 355+56=**411** |

# *Data formats*

## Image format

The form-pages are stored in uncompressed TIFF-file format (file-extension .tif).
All cropped Tunisian town/village names coming as uncompressed TIFF-images and as BMP-images (file-extension .bmp). In the TIFF header section “Image description” label information is stored.

## Ground Truth format

For each cropped Tunisian town/village name you find a truth file (file-extension .tru). The truth-file is an ASCII .txt file including all available ground truth information. One example is shown below.

```
01: COM: IFN/ENIT-database truth (label) file
02: COM: http://www.ifnenit.com
03: COM: IfN, TU-BS
04: COM: di45_019.tif coming from pb377_6.tif
05: X_Y: 498 87
06: BDR: begin data record
07: LBL: ZIP:3032;AW1:مركزد رويش;AW2:maB|raE|keB|zaE|daA|raA|waA|yaB|shE|;QUA:YB1;ADD:P6
08: CHA: 9
09: BLN: 56,42
10: EDR: end of data record
```

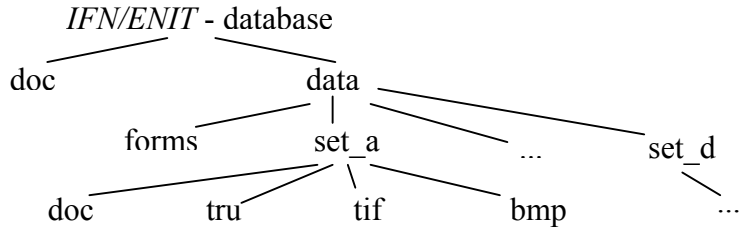
lines 01-04: comments

line 05: image size in pixel (x,y)
line 06: begin data record
line 07: label
- ZIP: Tunisian post code / ZIP code
- AW1: Tunisian town/village name in Arabic windows encoding
- AW2: Arabic character shape sequence of Tunisian town/village name in Latin code (refer Appendix for the lookup table).
- QUA: baseline quality tag (B1=:OK;B2=:bad)
- ADD: number of pieces of Arabic words (PAW's)

line 08: number of characters
line 09: baseline/reference line information Y1,Y2
line 10: end of data record

## *Database Organisation*

The *IFN/ENIT*-database has the following directory structure.




**doc**
In this directory documentation in pdf and/or txt format is available.

**data**
In this directory all available data are stored. Each of the four sets has it's own directory (set_?).

**set_?**
In these directories all data are available in tif and bmp file format. The ground truth information you will find in the directory tru. The subdirectory doc under data/set_? includes a pdf-file for each writer with all words and baselines.

**File name convention**

The following file name convention is used.
SWww_NNN.EXT

- S:=set (a,b,c,d)
- Www:=writerID; W=(e,f,i,j,m,q) ,w=(0..9)
- NNN:=word_number; N=(0..9)

## *Ordering Information*

*IFN/ENIT*-database is made available for non-commercial use. The data is supplied with no guarantee of accuracy or usability. We can't guarantee to maintain the *IFN/ENIT*-database, but would be interested in hearing of any comments or results that you have.
Upon request we make the data available on the Internet for free download. If the database has to be shipped as a CD-Rom production and shipping costs will be charged. In both cases please contact us.

## BUGS

When you are working with the *IFN/ENIT*-database perhaps you will discover bugs concerning the label or the extraction of the data. Please report the bugs you find back to us. So we can improve the quality of the data over the time.

## Reporting results

We kindly invite you to publish reached results with the *IFN/ENIT*-database. To keep recognition results comparable, we suggest reporting results as shown in the following example:

| Database version: ***IFN/ENIT*-database v1.0p1** | | | |
|---|---|---|---|
| Test | Training set(s) | Test set | Recognition result(*) |
| 1 | a,b,c | d | 23.7% |
| 2 | c,b,d | a | 25,4% |
| … | | | |

(*) Percentage of correctly recognised words in the specified test set. We recommend to use the ground truth / label category "ZIP:" as reference for the recognition result. Please use for each test the whole number of 946 different Tunisian town/village names as lexicon, in the case a lexicon is needed.

## Contact

Please feel free to contact contact@ifnenit.com.

## References

Mario Pechwitz, Samia Snoussi Maddouri, Volker Märgner, Noureddine Ellouze, Hamid Amiri; ***IFN/ENIT*-database of handwritten Arabic words**, In Proceedings of CIFED'02, Hammamet, Tunisia, 21.-23.10.2002, p.?

## Appendix

### Label legend / lookup table – *Arabic* label to *Latin* label & statistic of occurrence in the database

| ***Arabic*** label | ***Latin*** label | quantity |
|---|---|---|
| 0_A | 0A | 341 |
| 1_A | 1A | 280 |
| 2_A | 2A | 383 |
| 6_A | 6A | 314 |
| 7_A | 7A | 354 |
| 8_A | 8A | 285 |
| 9_A | 9A | 347 |
| ء_A | hhA | 520 |
| آ_A | amA | 544 |
| أ_A | aeA | 1660 |
| إ_A | ahA | 631 |
| أ_E_ل_B | aeElaB | 360 |
| إ_E_ل_B | ahElaB | 122 |
| ئ_M | alM | 354 |
| ا_A | aaA | 20256 |
| ا_E | aaE | 13308 |
| ا_E_ل_B | aaElaB | 799 |
| ا_E_ل_M | aaElaM | 1076 |
| ب_A | baA | 331 |
| ب_B | baB | 5637 |
| ب_E | baE | 344 |
| ب_M | baM | 3406 |
| ة_A | teA | 2184 |
| ت_A | taA | 358 |
| ت_B | taB | 2324 |
| ة_E | teE | 7261 |
| ت_E | taE | 357 |
| ت_M | taM | 1045 |
| ث_A | thA | 338 |
| ث_B | thB | 353 |
| ث_M | thM | 327 |
| ج_A | jaA | 505 |
| ج_B | jaB | 981 |

| | | |
|---|---|---|
| ج_E | jaE | 346 |
| ج_M | jaM | 1219 |
| ج_Mل_B | jaMlaB | 539 |
| ح_A | haA | 314 |
| ح_B | haB | 2483 |
| ح_E | haE | 295 |
| ح_M | haM | 1806 |
| ح_Mل_B | haMlaB | 365 |
| ح_Mم_Mل_B | haMmaMlaB | 64 |
| ح_Mن_B | haMnaB | 100 |
| خ_A | khA | 341 |
| خ_B | khB | 863 |
| خ_E | khE | 321 |
| خ_M | khM | 425 |
| خ_Mل_B | khMlaB | 310 |
| د_A | daA | 2717 |
| د_E | daE | 4886 |
| ذ_A | dhA | 353 |
| ذ_E | dhE | 703 |
| ر_A | raA | 6255 |
| ر_E | raE | 9370 |
| ز_A | zaA | 1765 |
| ز_E | zaE | 2719 |
| س_A | seA | 873 |
| س_B | seB | 4218 |
| س_E | seE | 811 |
| س_M | seM | 1111 |
| ش_A | shA | 475 |
| ش_B | shB | 1616 |
| ش_E | shE | 351 |
| ش_M | shM | 1277 |
| ص_A | saA | 355 |
| ص_B | saB | 956 |
| ص_E | saE | 357 |
| ص_M | saM | 1096 |
| ض_A | deA | 351 |
| ض_B | deB | 731 |
| ض_E | deE | 343 |
| ض_M | deM | 328 |
| ط_A | toA | 343 |
| ط_B | toB | 359 |
| ط_E | toE | 350 |
| ط_M | toM | 1259 |
| ظ_B | zaB | 338 |

| | | |
|---|---|---|
| ظ_M | zaM | 690 |
| ع_A | ayA | 915 |
| ع_B | ayB | 1650 |
| ع_E | ayE | 347 |
| ع_M | ayM | 1990 |
| غ_B | ghB | 326 |
| غ_M | ghM | 600 |
| ف_A | faA | 319 |
| ف_B | faB | 898 |
| ف_E | faE | 316 |
| ف_M | faM | 1647 |
| ق_A | kaA | 397 |
| ق_B | kaB | 2608 |
| ق_E | kaE | 348 |
| ق_M | kaM | 1307 |
| ك_B | keB | 1221 |
| ك_E | keE | 335 |
| ك_M | keM | 980 |
| ل_A | laA | 1485 |
| ل_B | laB | 14345 |
| ل_E | laE | 1056 |
| ل_M | laM | 2594 |
| م_A | maA | 890 |
| م_B | maB | 3885 |
| م_E | maE | 536 |
| م_M | maM | 4629 |
| م_Mل_B | maMlaB | 457 |
| ن_A | naA | 1267 |
| ن_B | naB | 3723 |
| ن_E | naE | 2119 |
| ن_M | naM | 2913 |
| ه_A | heA | 696 |
| ه_B | heB | 1924 |
| ه_E | heE | 351 |
| ه_M | heM | 347 |
| و_A | waA | 3511 |
| و_E | waE | 6529 |
| ى_A | eeA | 322 |
| ي_A | yaA | 2932 |
| ي_B | yaB | 4385 |
| ى_E | eeE | 350 |
| ي_E | yaE | 2167 |
| ي_M | yaM | 7760 |

**Note:**

- "llL" is often added and means ligature "chadda"
- The character shape indicators (A,B,M,E) are sometimes supplemented with a "1" or a "2", like "baA1". In this case there is a point error detected. Take care: this feature is not consistent over the whole database ☹. We recommend ignoring these kinds of label supplements.