# NIST Special Database 19
# Handprinted Forms and Characters
# 2nd Edition

Patrick Grother<br>
Kayee Hanaoka

NIST

National Institute of
Standards and Technology
U.S. Department of Commerce

# NIST Special Database 19
# Handprinted Forms and Characters
# 2nd Edition

Patrick Grother
Kayee Hanaoka
*Information Technology Laboratory*
*Information Access Division*

August 2016

U.S. Department of Commerce
*Penny Pritzker, Secretary*

National Institute of Standards and Technology
*Willie May, Under Secretary of Commerce for Standards and Technology and Director*

# Table of Contents

# NIST Special Database 19
# Handprinted Forms and Characters Database

## 1 Introduction

The updated web released *NIST Special Database 19 (SD19)* consist of 5 zipped files with a total of 3 992 357 Portable Network Graphics (PNG) [1] converted images. The original CD-ROM *Special Database 19* was released in 1995. The original SD 19 contains the full page binary images of 3669 Handwriting Sample Forms (HSFs) and 814 255 segmented handprinted digit and alphabetic characters from those forms. Those segmented characters each occupy 128x128 pixel per raster and are labelled by one of 62 ASCII hexadecimal classes corresponding to “0”- “9”, “A”- “Z” and “a”- “z” [3]. The segmented characters images are included in multiple organizations suited to different recognition applications. The characters are given by writer, by class, by caseless class, and by field origin.

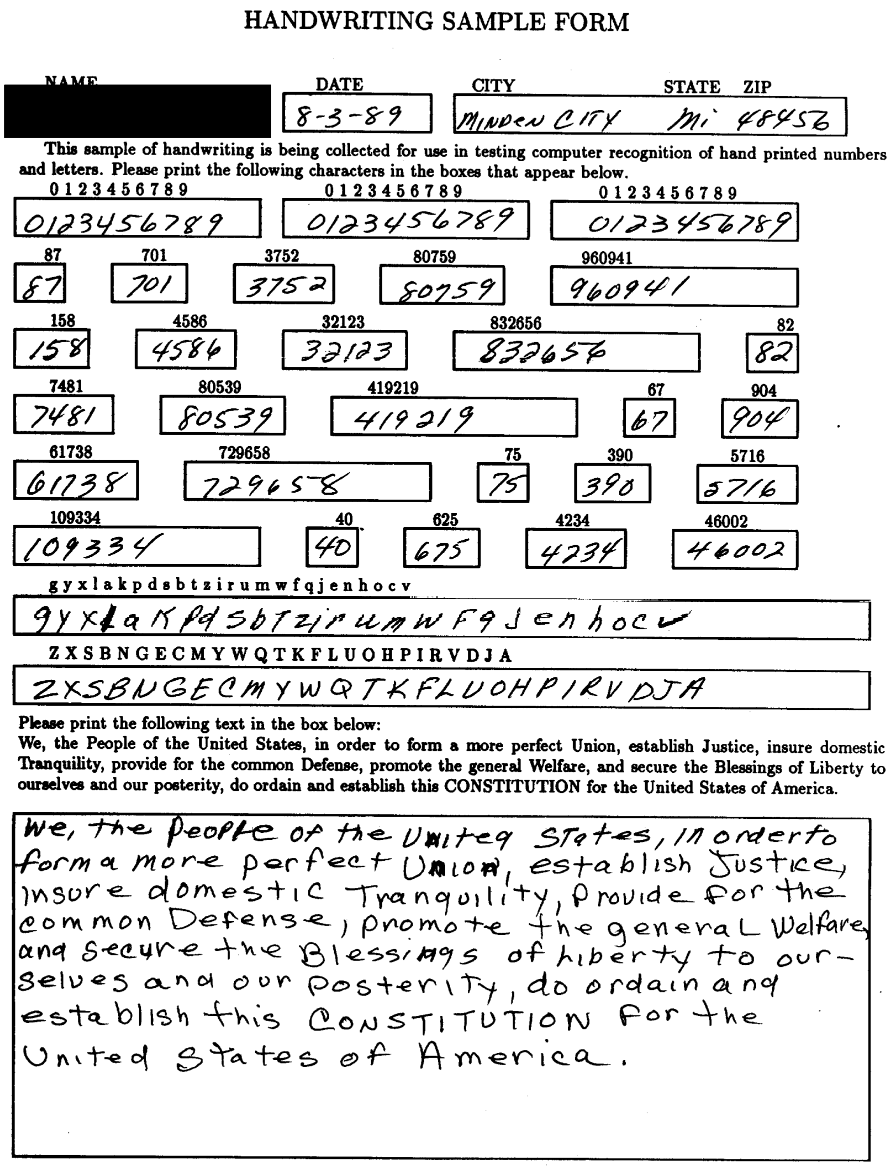
### 1.1 Source Materials

The SD 19 contains eight series of HSF images, denoted by *hsf_{0,1,2,3,4,6,7,8}*. Characters segmented from all field types are included in this database. In 1st Edition, all images are binary and are stored in NIST’s IHEAD format.

The publication statuses of the various writer partitions and field types are given in the Table 1. The partition of *hsf_4* were completed by the Bethesda high school students, and all those of partitions *hsf_{0,1,2,3,6,7,8}* were obtained from Census Bureau employees in Suitland, Maryland.

| partition | writers | writer origin |
|---|---|---|
| hsf_0 | 0000-0499 | Census Field |
| hsf_1 | 0500-0999 | Census Field |
| hsf_2 | 1000-1499 | Census Field |
| hsf_3 | 1500-2099 | Census Field |
| hsf_4 | 2100-2599 | High School |
| hsf_6 | 3100-3599 | Census Field |
| hsf_7 | 3600-4099 | Census Field |
| hsf_8 | 4100-4169 | Census Field |

*Table 1: Break down of the partitions with the writer numbers.*



HANDWRITING SAMPLE FORM

| NAME | DATE | CITY | STATE | ZIP |
|---|---|---|---|---|
| [redacted] | 8-3-89 | MINDEN CITY | Mi | 48456 |

This sample of handwriting is being collected for use in testing computer recognition of hand printed numbers and letters. Please print the following characters in the boxes that appear below.

| | | |
|---|---|---|
| 0 1 2 3 4 5 6 7 8 9 | 0 1 2 3 4 5 6 7 8 9 | 0 1 2 3 4 5 6 7 8 9 |
| 0123456789 | 0123456789 | 0123456789 |

| | | | | |
|---|---|---|---|---|
| 87 | 701 | 3752 | 80759 | 960941 |
| 87 | 701 | 3752 | 80759 | 960941 |
| 158 | 4586 | 32123 | 832656 | 82 |
| 158 | 4586 | 32123 | 832656 | 82 |
| 7481 | 80539 | 419219 | 67 | 904 |
| 7481 | 80539 | 419219 | 67 | 904 |
| 61738 | 729658 | 75 | 390 | 5716 |
| 61738 | 729658 | 75 | 390 | 5716 |
| 109334 | 40 | 625 | 4234 | 46002 |
| 109334 | 40 | 675 | 4234 | 46002 |

g y x l a k p d s b t z i r u m w f q j e n h o c v

gyxlakpdsbtzirumwfqjenhocv

Z X S B N G E C M Y W Q T K F L U O H P I R V D J A

ZXSBNGECMYWQTKFLUOHPIRVDJA

Please print the following text in the box below:
We, the People of the United States, in order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common Defense, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our posterity, do ordain and establish this CONSTITUTION for the United States of America.

We, the People of the Uniteq States, In orderto
form a more perfect Union, establish Justice,
insure domestic Tranquility, provide for the
common Defense, promote the general Welfare,
anq secure the Blessings of Liberty to ourselves
and our posterity, do ordain anq
establish this CONSTITUTION for the
United States of America.

*Figure 1: Example HSF Image. This is the file hsf_page/hsf_0/f0002_01.pct. Notice that the first field on this form, the name field, has been intentionally occluded, on some others it remains blank. All fields except those on the first line have been segmented and recognized by NIST.*

### 1.2 About 2nd Edition

The 2nd Edition of Special Database 19 converted all binary images in the 1st Edition dataset to PNG format images. The 2nd Edition SD19 data hierarchies are very similar to the 1st Edition which is describes in *Section 2.* The organization of the PNG images are consistent with the *.mis* files where the original binary images originated. Each *.mis* file (*Section 2*) is converted into multiple PNG images because the *.mis* files consist of more than one segmented digit and/or alphabetic character per binary image. All PNG images are located in the directory that is named after the *.mis* file from which the PNG images were converted. The *.mis* file name is also the front naming of the PNG file followed by an underscore and a number, and the numbering of the files begins with 00000. For example, d0000_14.mis contains 235 characters images in binary format. 235 individual PNG files get converted from d0000_14.mis with filenames from d0000_14_00000.png to d0000_14_00234.png.

## 2 Data Hierarchies

There are five directories in the *data* subtree. The first *hsf_page* contains images of the full page HSF form. The other four directories, *by_**, each have alternative organizations of the segmented character images suited to different recognition applications. The characters are given by writer, by class, by field origin, and finally, and finally by caseless class.

These are the definitions of the file extensions correspond to particular files in 1st Edition.

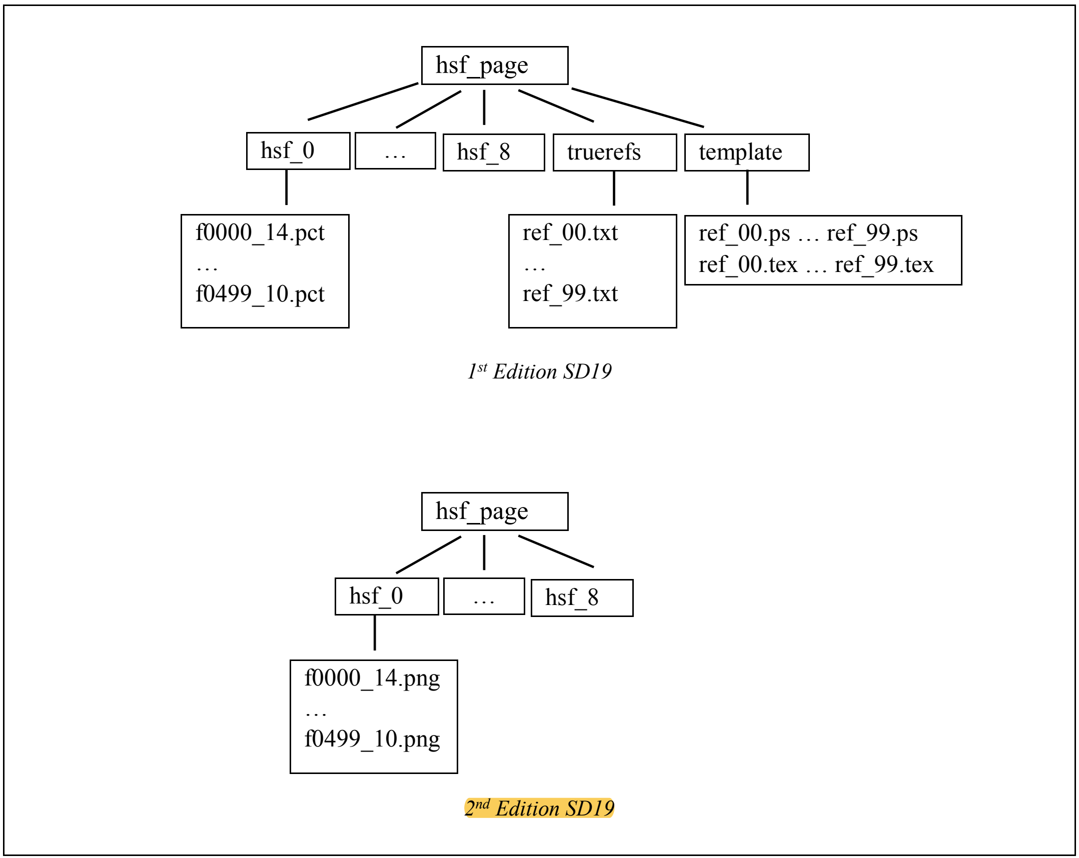
*.mis* – a file containing binary format of multiple isolated character images.

*.pct* – a file containing a full page HSF IHEAD formatted image file.

*.cls* – a file containing the checked classes of the images held in the accompanying *.mis* file.

### 2.1 *hsf_page* – Full HSF Page Images

The 2nd Edition SD19 *hsf_page* contains all PNG images of the HSF forms in the *hsf_{0,1,2,3,4,6,7,8}* directories. They were converted from the *.pct* format of the HSF forms (Figure 1) in 1st Edition. The file hierarchies for 1st and 2nd Edition are very similar (Figure 2) except the *truerefs* directory holds the text reference files, and the *template* directory contains postscript (*.ps*) and LaTeX files (*.tex*) for the unfilled HSF forms are not included in the 2nd Edition. User can download the 1st Edition zip file to access that data.



*1st Edition SD19*

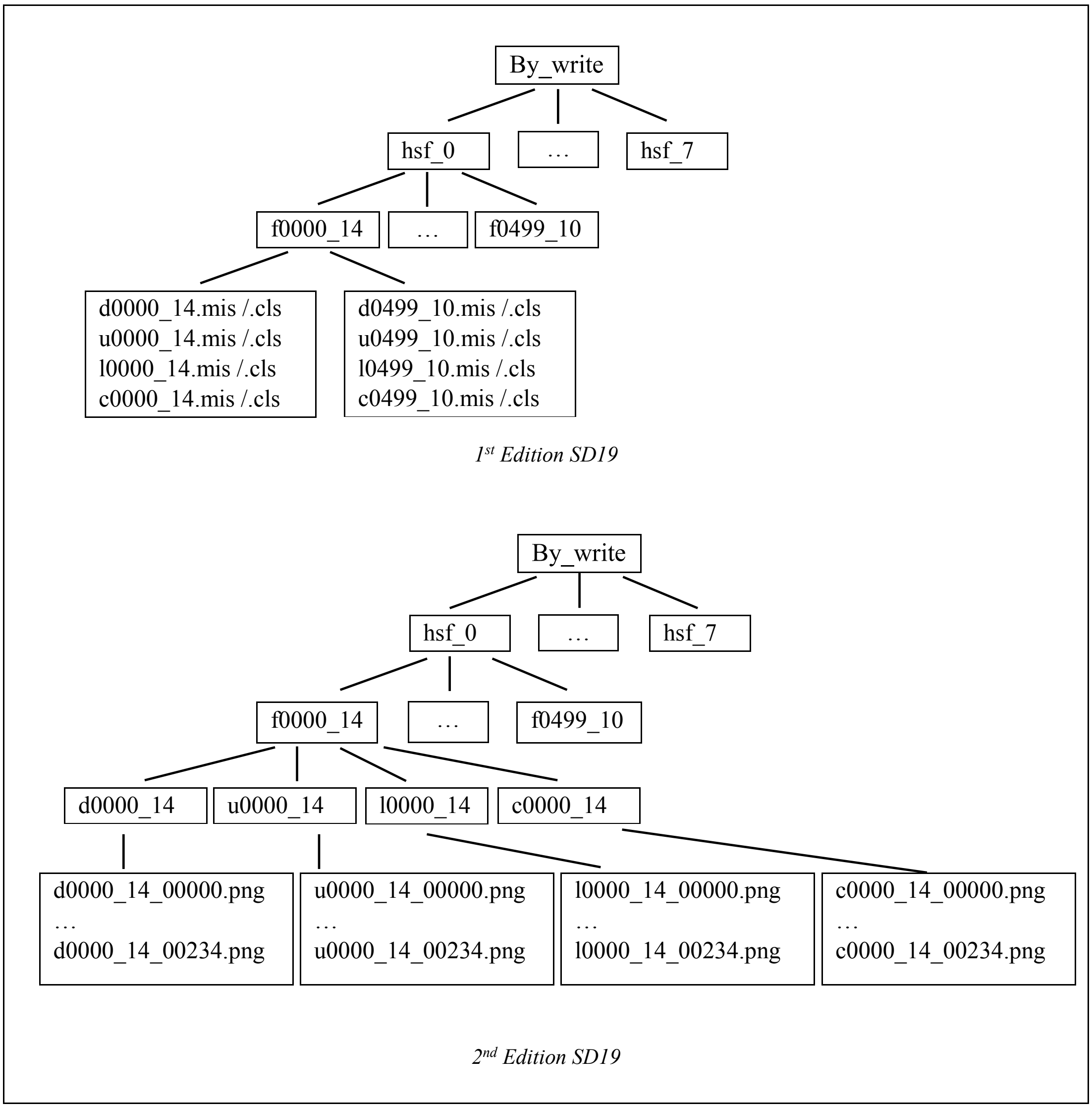
*2nd Edition SD19*

*Figure 2: 1st and 2nd Edition SD 19 file hierarchies for hsf_page*

## 2.2 *by_write* – By Author

*by_write* contains the segmented characters organized by *hsf_?* partition then by writer. This organization is generally not particularly useful for OCR studies since the image files contain multiple classes. The files are, however, the primary output of the segmentation and checking process, and the other hierarchies that follow were derived from it.

Each writer directory contains files for each field type; digit, upper, lower and, constitution alphas. The 2nd Edition SD 19 contains all PNG images of individual characters that were converted from the *.mis* files in 1st Edition. Each *.mis* file in 1st Edition contains multiple characters images in binary format, for example, the u0000_14.mis file in f0000_14 directory, contains all uppercase alphabetic character images in binary format written by writer f0000_14. The file hierarchies for 1st and 2nd Edition are given below (Figure 3).



*1st Edition SD19*

*2nd Edition SD19*

*Figure 3: 1st and 2nd Edition SD 19 file hierarchies for by_write partition*

## 2.3 *by_field* – By Field Type

*by_field* contains characters organized by *hsf_?* then partitioned by field type, and finally by class. Writer information is discarded though the files' entries are included by concatenation of the writer characters from the *by_write* tree. The digit directory contained images from all digit fields from the form. The upper and lower directories contained images from the uppercase field and the lowercase field. The const directory contained all images of the characters from the constitution box.(Figure 1) All images in the digit/upper/lower/const directories are partitioned by the ASCII hexadecimal classes. The file hierarchies in 1st and 2nd Edition are very similar. The 2nd Edition SD 19 contains all PNG images of individual characters that were converted from the *.mis* files in 1st Edition. Each *.mis* file in 1st Edition contains multiple images in binary format of the same digit or alphabetic characters by field type. The file hierarchies for 1st Edition and 2nd Edition SD19 are given below. (Figure 4)

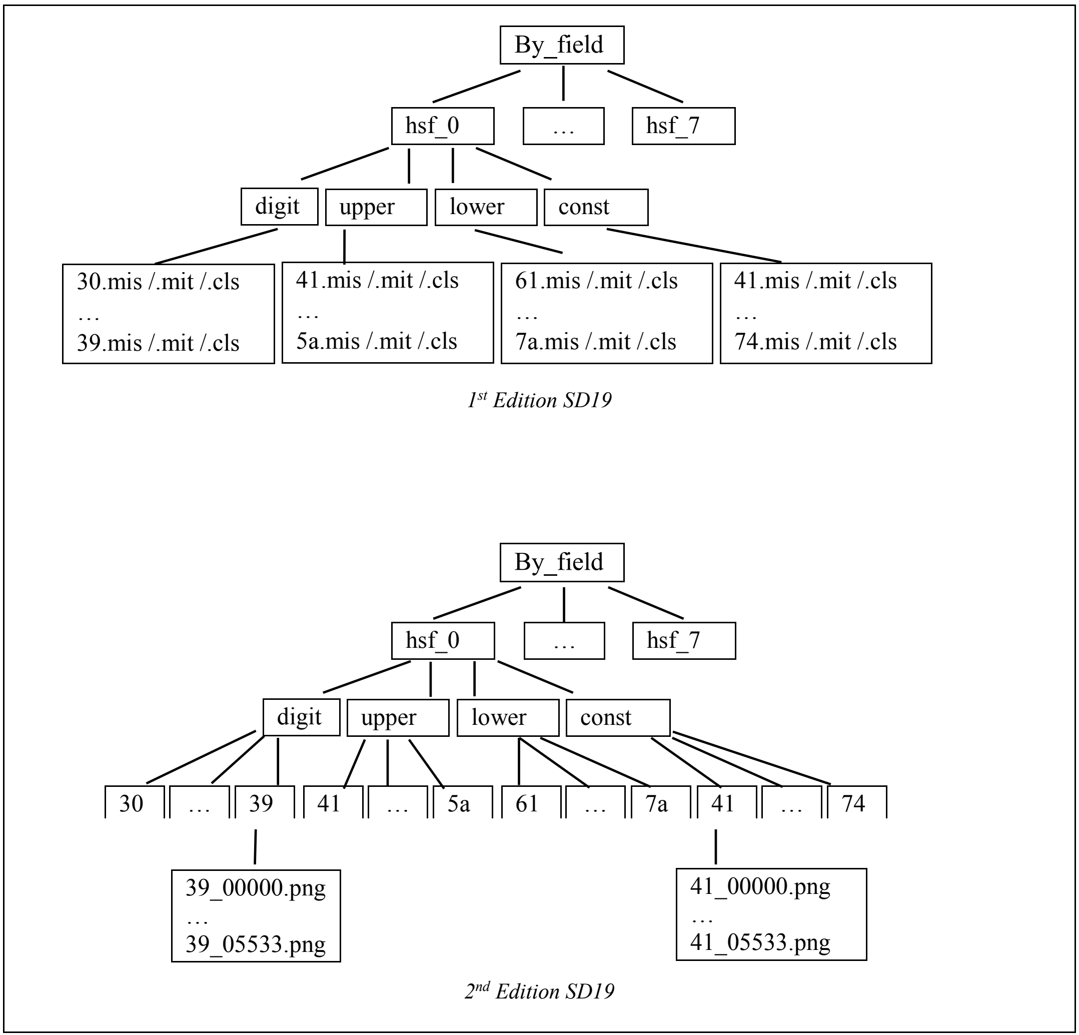
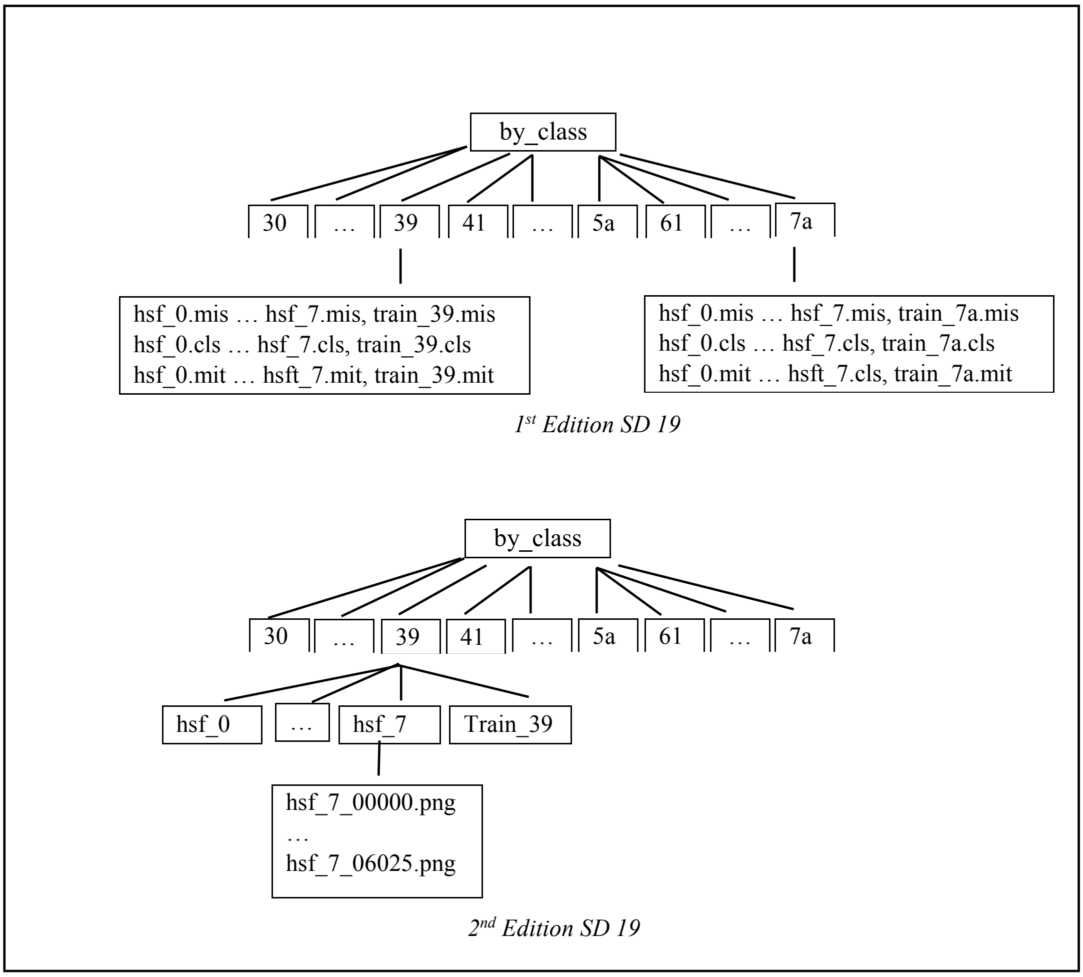


*Figure 4: 1st and 2nd Edition SD 19 file hierarchies for by_field partition.*

## 2.4 *by_class* – By Hexadecimal Class

*by_class* contains images organized by class, then by database. Both writer and field information are discarded: there is no distinction between an "e" from the constitution box of writer 0000 and one from the lower case field of writer 4044. In the directory structures that follow the second layer directories have labels which are the hexadecimal ASCII representations of the textual class labels.

The *train_30* files contains the "0"s of all writers of partitions *hsf_{0,1,2,3,6,7}*. The *train_??* files comprise the suggested training set for OCR studies. The *hsf_4* is likewise earmarked as a standard testing results reporting set. Note that the class files are redundant in this tree, since they contain only one unique hexadecimal class string, and the class has already been indicated in the parent directory name. The 2nd Edition SD 19 contains all PNG images of individual characters that were converted from the *.mis* files in 1st Edition. Each *.mis* file in 1st Edition contains multiples images in binary format of the same digit or alphabetic characters by partition. The file hierarchies for 1st Edition and 2nd Edition SD29 are given below. (Figure 5)


by_class
30
…
39
41
…
5a
61
…
7a
hsf_0.mis … hsf_7.mis, train_39.mis
hsf_0.cls … hsf_7.cls, train_39.cls
hsf_0.mit … hsft_7.mit, train_39.mit
hsf_0.mis … hsf_7.mis, train_7a.mis
hsf_0.cls … hsf_7.cls, train_7a.cls
hsf_0.mit … hsft_7.cls, train_7a.mit

*1st Edition SD 19*

by_class
30
…
39
41
…
5a
61
…
7a
hsf_0
…
hsf_7
Train_39
hsf_7_00000.png
…
hsf_7_06025.png

*2nd Edition SD 19*

*Figure 5: 1st and 2nd Edition SD 19 file hierarchies for by_class partition.*

## 2.5 *by_merge* – By Merged Class

The class abundancies show up to an order of magnitude disparity between classes. This situation may be ameliorated for certain applications by folding the upper and lower case letters of some classes into one another; for instance, an upper case “W” is largely equivalent for recognition purposes to its lower case analogue “w”. Indeed, it could be argued that a classifier could equally be trained to recognize classes of different appearance, “A” and “a” for example, on the basis that, although examples of the two classes may form separate clusters in a representative feature space, some classifiers will still perform well. For this hierarchy the upper and lower case examples of the following thirteen classes have been merged:

C I J K L M O P S U V W X Y Z

The resulting tree hierarchy contains exact replicas of the files of the unmerged classes from the *by_class* tree, and the merged classes labelled by the hexadecimal codes of the upper and lower case labels delimited by a period. The final number of classes is 37. The 2nd Edition SD 19 contains all PNG images of individual characters that were converted from the *.mis* files in 1st Edition. Each *.mis* file in 1st Edition contains multiples characters images in binary format. The file hierarchies for 1st Edition and 2nd Edition SD29 are given below.(Figure 6)

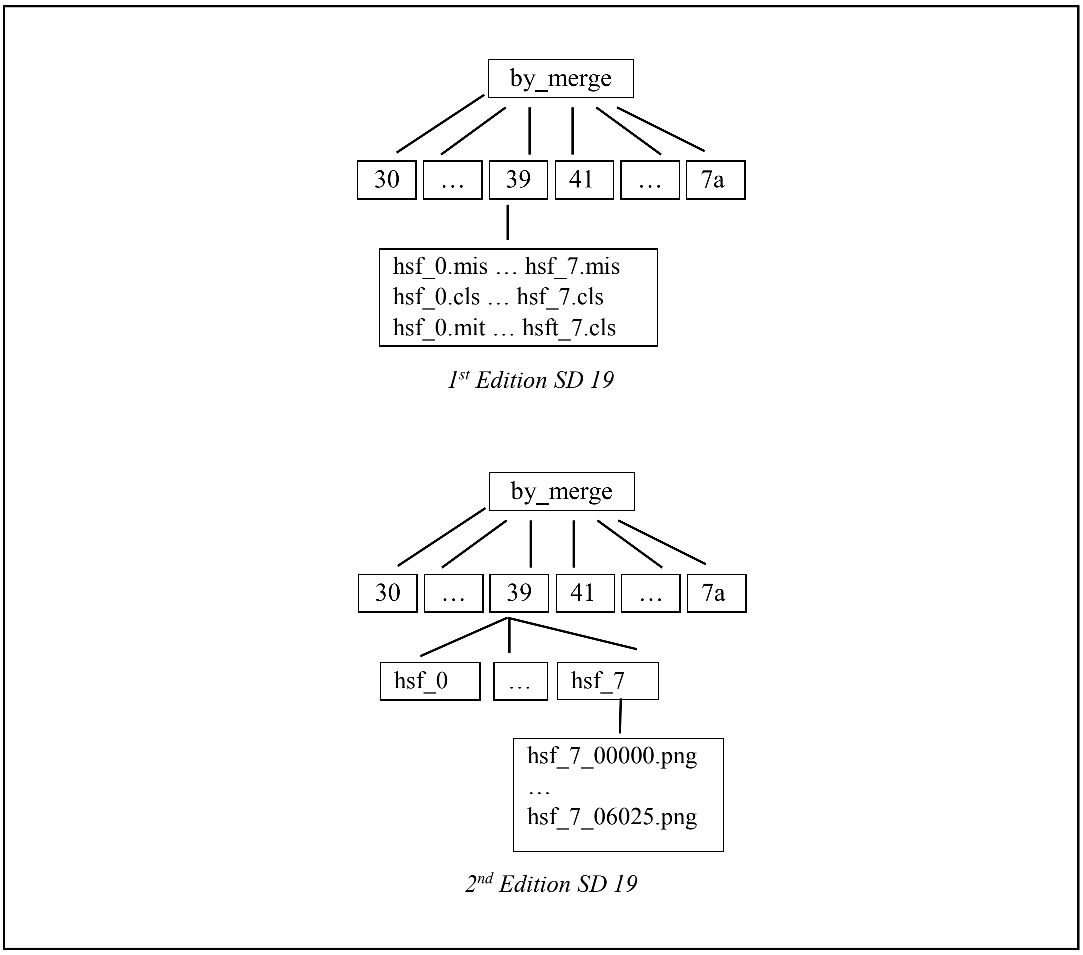


*Figure 6: 1st and 2nd Edition SD 19 file hierarchies for by_merge partition.*

# 3 Reference

[1] “Information technology – Computer graphics and image processing – Portable Network Graphics (PNG): Functional specification”, International Organization for Standardization/International Electrotechnical Commission, ISO/IEC 15948:2004

[2] Patrick J. Grother. NIST Special Database 19 – Handprinted Forms and Characters Databas 1st Edition User’s Guide, National Institute of Standards and Technology, March 1995

[3] “ASCII Table and Description”, www.asciitable.com