



**User Manual**

## **Data Collection Tool 1.0.1**



## Table of Contents

1	Introduction .....	2
2	Data collection tool (DCT) .....	2
2.1	Introduction.....	2
2.2	Requirements for running the tool .....	2
2.2.1	Requirements for running the tool on desktop	2
2.3	Configuring DCT to collect data .....	3
2.3.1	Input Text File Format	3
2.4	Steps in data collection .....	5
2.4.1	Providing writer information	5
2.4.2	Setting the reference lines for writing	6
2.4.3	Data capture dialog	6
2.4.4	Verifying the collected data	8
2.5	Troubleshooting.....	8
2.6	Use case walkthrough – Collecting English alphabets.....	9
3	Appendix .....	9
3.1	Steps to install tablet PC runtime .....	9
3.2	Steps to install Indic language support (for DCT) .....	10
3.3	Sample text for uppercase English alphabets.....	11
4	Glossary .....	12



## 1. Introduction

LipiTk (Lipi Toolkit) is a generic toolkit for Online Hand Writing Recognition (HWR). It provides a set of script-independent shape recognizers, tools and building blocks which can be used by different kinds of users for different scripts and shapes.

This document describes the data collection tool for handwriting recognition.

## 2. Data collection tool (DCT)

The data collection tool and its requirements have been discussed in detail under this section.

### 2.1 Introduction

The data collection tool (DCT) provides a user interface to create handwriting datasets by collecting handwriting data samples from different writers. Given the script and the data elements to be collected, writers contribute to the dataset by giving their handwriting samples. The output of the tool is a set of UNIPEN files organized in a directory structure, containing writer profiles, collection procedure details and the digital ink. The name of the executable for the tool is “hw dct”.

### 2.2 Requirements for running the tool

The current version of DCT is supported on

- Tablet PC with Microsoft Windows XP TabletPC Edition
- Desktop PC with Microsoft Windows XP or Microsoft Windows 2000 operating system.

#### 2.2.1 Requirements for running the tool on desktop

In order to run this tool on a desktop, Tablet PC Runtime must be installed. The monitor resolution should be adjusted to a minimum of 1152 x 864 pixels. Refer [Section 3.1](#) for the steps to install Tablet PC Runtime.



## 2.3 Configuring DCT to collect data

The DCT takes text files which contain the data-elements to be collected as an input. The supervisor creates the list of elements that needs to be collected. The input text file should comply with the format given below:

### 2.3.1 Input Text File Format

1. Every input text file starts with “#FONT” and the name of the font to be loaded to display the data-elements. For example #FONT ARIAL
2. All the data elements should be listed in the input text file, with one data element each line.
3. Corresponding to each data element, numeric ID starting with zero and the data-element to be collected (character or word) should be specified.

Typically an input text file looks like this:

**a.** hindichar.txt – Unicode file

[For collecting Hindi characters]

#FONT MANGAL

0 को

1 रौं

2 पा

EOF

**b.** hindiwords.txt – Unicode file

[For collecting Hindi words]

#FONT MANGAL

0 राकि

1 भारत

2 लिपि

3 घाडि

EOF

Note: DCT can read Unicode files and display Unicode characters. Refer [Appendix 3.2](#) for more details on installing Indic scripts support and relevant fonts to display Unicode fonts.

### Configuration details

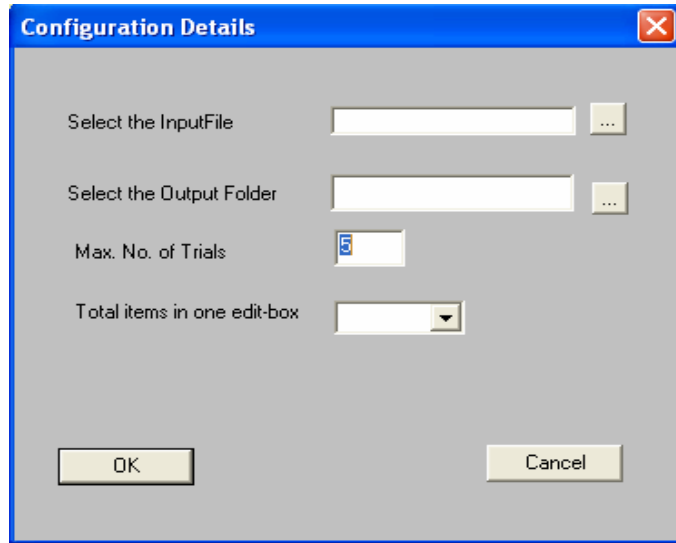
The Supervisor of Data Collection process should undertake the following steps:

1. Click Change Configuration. The configuration gets changed.

A Login dialog box appears.

2. Enter “123” as the password.

Configuration Details dialog box appears.



3. In the Configuration Details dialog box provide the following information:
  - a) Select the input text file which contains the set of elements to be collected. I.e. (hindichar.txt, EnglishAlphabets.txt...)
  - b) Select the folder in which all the collected output data should be stored.  
(\$LIPI\_ROOT/projects/<projectname>/data/raw (or) any specific folder where all the samples should be stored.)
  - c) Enter maximum number of trials per user. The samples will get collected as many times as specified here.
  - d) Enter the total number of items to be displayed in a single edit box.

The next window appears.

4. This window displays “EnglishAlphabets.txt” configuration file when the value is set to 3.



5. Make changes if required then click Ok.

All the configuration details will get saved in a text file called configuration.txt in the current working directory. The Writer Information dialog box appears again.

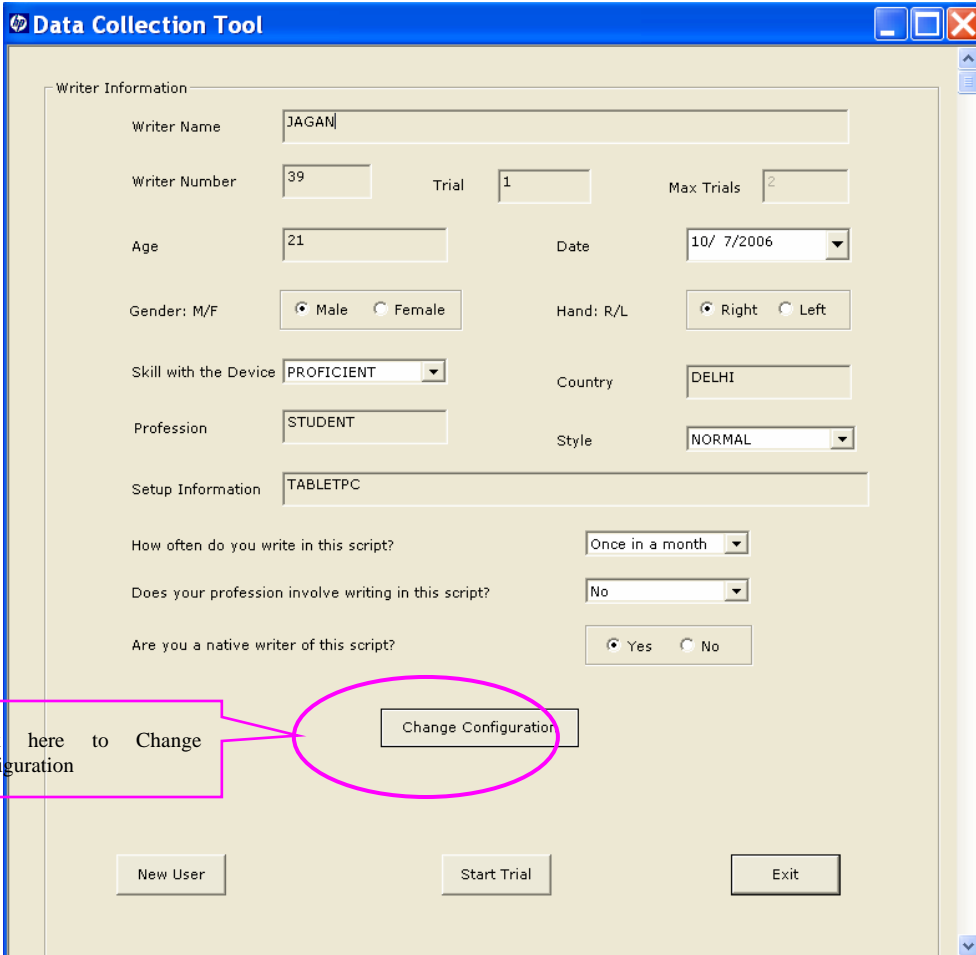
## 2.4 Steps in data collection

The data collection procedure involves the following steps:

### 2.4.1 Providing writer information

Run the 'hwDCT.exe'. Writer Information dialog box appears. This box enables the writer to enter his personal information like Name, Age, Profession and information about his familiarity with the device and the script.

A snapshot of the Writer Information dialog box is shown below.



The screenshot shows the 'Data Collection Tool' window with the 'Writer Information' dialog box. The dialog box contains the following fields and controls:

- Writer Name: JAGAN
- Writer Number: 39
- Trial: 1
- Max Trials: 2
- Age: 21
- Date: 10/ 7/2006
- Gender: M/F (Male selected)
- Hand: R/L (Right selected)
- Skill with the Device: PROFICIENT
- Country: DELHI
- Profession: STUDENT
- Style: NORMAL
- Setup Information: TABLETPC
- How often do you write in this script? (Once in a month)
- Does your profession involve writing in this script? (No)
- Are you a native writer of this script? (Yes selected)
- Buttons: New User, Start Trial, Exit, and Change Configuration (highlighted with a pink callout).

The Writer Information dialog box essentially collects the following information from the user:

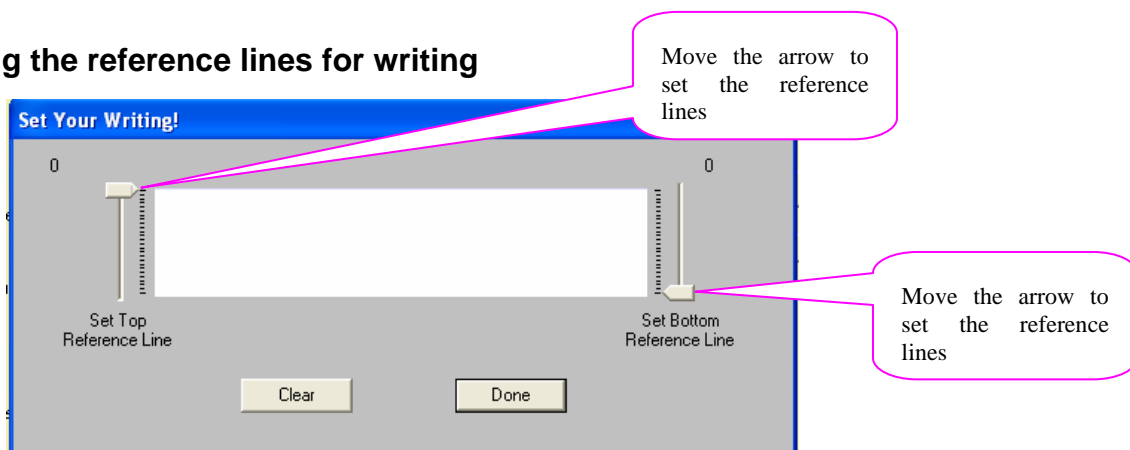
- Writer name (mandatory field)
- Writer number (automatically generated)
- Writer trial number denoted by trial (automatically generated)

- Writer’s age (mandatory field)
- Writer’s gender (mandatory field)
- Writer’s hand: If the writer is a left handed or write handed writer. (mandatory field)
- Writer’s skill: If the writer is familiar with the usage of the device using which the data is collected. (mandatory field)
- Writer’s country (mandatory field)
- Writer’s profession (mandatory field)
- Writer’s style: (Normal/Cursive/Mixed/Printed) (mandatory field)
- The setup information, wherein the writer gives the device name and specifications (mandatory field)
- The frequency of using the particular script? (mandatory field)
- If the writer’s profession involves any amount of writing in the particular script in which he is asked to give his handwriting samples? (mandatory field)
- Is the writer a native writer of the script? (mandatory field)

NOTE : To reset the writer information in all the edit boxes, click “New User” button at the bottom left of the Writer Information dialog box.

Enter the information in the edit boxes and click Start Trial.

### 2.4.2 Setting the reference lines for writing

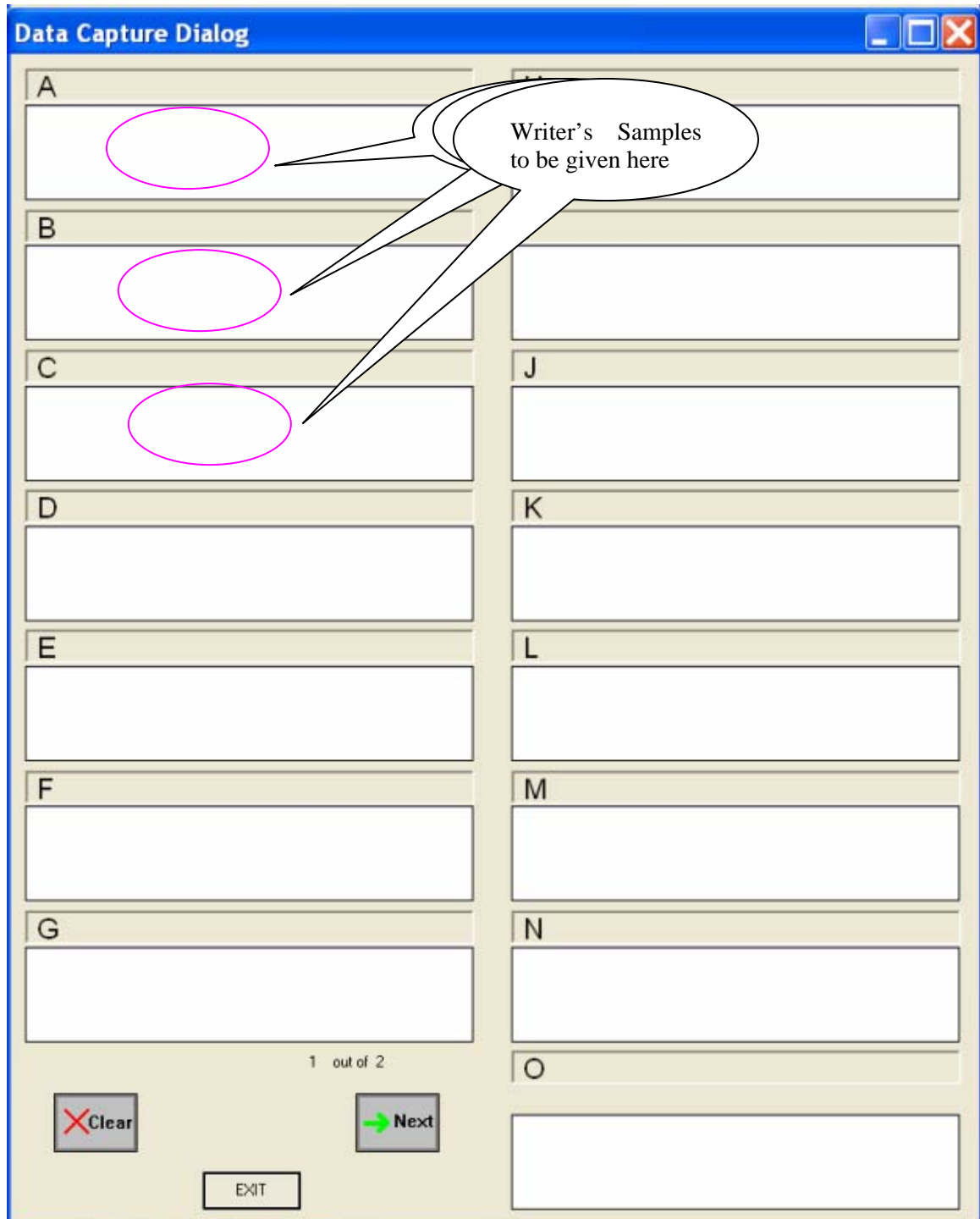


The Set Writing dialog box allows the user to set the grid reference lines using the arrows provided on both the sides. The text should be written between these grid reference lines. The Clear button erases the information written previously. It can also be used to check the new trial again. Click “Done” once the reference lines have been set.

### 2.4.3 Data capture dialog

The data capture dialog box collects the handwriting samples. It has ink picture controls for capturing the ink. Each ink picture control can contain up to a maximum of 3 elements. This can be set by the Supervisor. (Refer section 4.3, Step 2 for details)

- 1 The cursor changes to “dot cursor” when you float on the ink picture controls.



- 2 Write the character/word elements in the ink picture control with the help of the input device.





- 3 Use Clear to delete the input incase it is wrong or distorted due to human/device errors.  
The Clear option removes all the ink in the corresponding ink picture control and allows user to rewrite the data.
- 4 Click “Next” to bring up the next set of data-elements.
- 5 After all the data-elements have been collected, the application goes back to the Writer Information dialog box and the trial ends at this stage.
- 6 The trial number gets incremented and is displayed in the trial edit box of the Writer Information dialog box.
- 7 Start the next trial by clicking Start Trial.  
  
NOTE: The next trial needs to be undertaken if more than one trial per user is configured.
- 8 At the end of all the trials, exit the Data Collection process by clicking Exit.  
  
NOTE: If the user exits during the trial, then all the data will be lost and the user will have to restart the trial.

#### **2.4.4 Verifying the collected data**

- All the collected handwriting samples can be found in the output folder specified while configuring the tool.
- Corresponding to every user a folder is created with the name “user<usr num>”. This folder contains text files corresponding to each element of every trial.
- Every text file is named using the user number, trial number and the element number in the collection list. (DATAROOT/usr<usr num>/<data element ID>t<trial ID>.txt ...)  
For example: D:\data\usr0\000t01.txt
- Every text file containing the collected data is stored in UNIPEN format with the user information at the beginning of the text file followed by the pen traces captured between a pen-up and pen-down for all the strokes that make up the data-element.

## **2.5 Troubleshooting**

- In a Desktop PC the mouse pointer is not changed to a dot or the mouse pointer does not write.  
  
To run the DCT on a Desktop PC, the OCX and ink Library with the Tablet PC SDK should be installed in the system.
- Clicking the “Change configuration” button results in the loss of all the writer information entered information in the edit boxes.



Note that the configuration details should be changed only by the supervisor. The process of data collection starts by changing the configuration details which is done by the supervisor. So the writer information should be entered only after the configuration changes have been done.

- In a Desktop PC the full dialog box of the application can not be seen.  
The monitor resolution should be increased to a minimum of 1152x864 pixels.

## 2.6 Use case walkthrough – Collecting English alphabets

- 1 Create an Input text file similar to the one shown in Figure titled EnglishAlphabets.txt.  
[ Refer [Section 3.3](#) ]
- 2 Run hwdct.exe.
- 3 Change the configuration details as follows:
  - a) Set the path of the input file to the created text file
  - b) Set the output folder
  - c) Set the desired number of trials and the number of items in edit control
- 4 Click New User and fill in all the fields of the Writer Information dialog box.
- 5 Click Start Trial.
- 6 Provide data samples in the ink picture controls.
- 7 Click Clear to undo your mistakes.
- 8 Click Next to move to the next set of symbols.
- 9 Click Start Trial to move to the next trial.
- 10 If all trials are completed click Exit to close the application.

## 3 Appendix

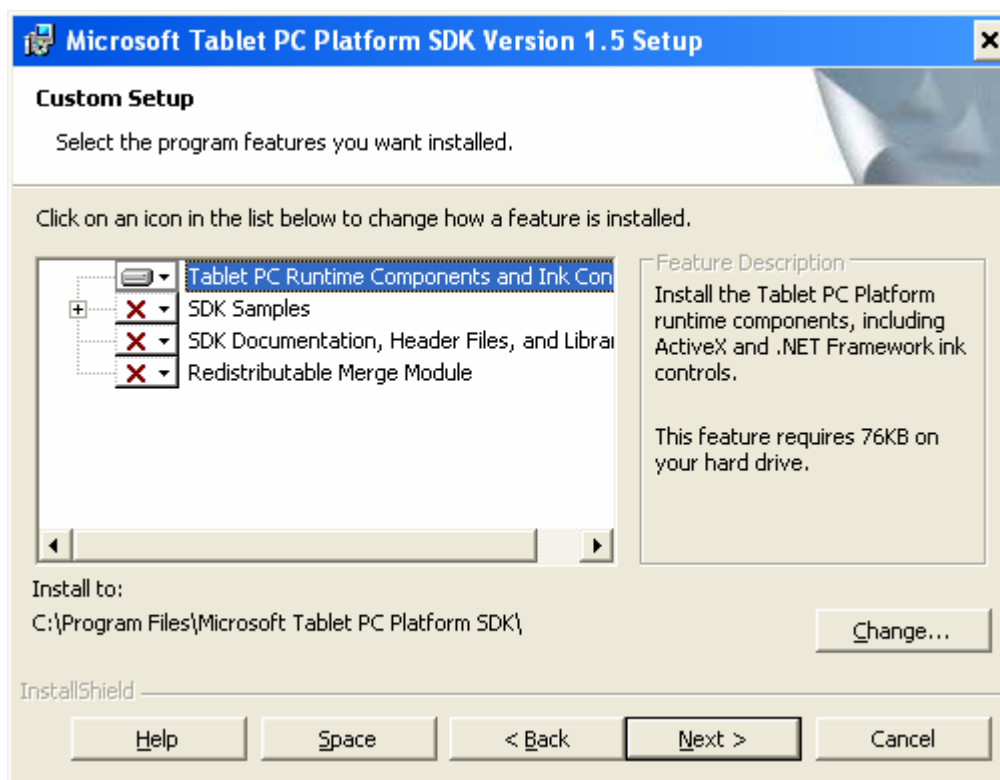
### 3.1 Steps to install tablet PC runtime

Microsoft Tablet PC SDK can be downloaded from the following link:

<http://www.microsoft.com/downloads/details.aspx?FamilyId=B46D4B83-A821-40BC-AA85-C9EE3D6E9699&displaylang=en>

To install the Microsoft Tablet PC Runtime:

- 1 Run setup.exe after downloading.
- 2 Continue to follow the instructions given in the setup dialogs.
- 3 From the Setup Type dialog box, choose Custom option and continue by clicking on Next.
- 4 Choose Tablet PC Runtime Components and Ink Controls option. Deselect other options as shown below.



- 5 Click Next and continue.

The installation is successfully done.

### 3.2 Steps to install Indic language support (for DCT)

- 1 Click Start. From the menu which appears, select Settings. A drop down list appears. From that drop down list select Control Panel.
- 2 Select Regional Options.
- 3 From the general tab of Regional Options dialog box, check Indic under Language Settings for the system list.
- 4 From Input Locales tab add required languages using the Add button to the Installed Input Locales list.



- 5 Click Ok. Prompt for restart will appear. Restart the machine.

The installation is successfully done.

### **3.3 Sample text for uppercase English alphabets**

Sample text file for collecting data for Upper Case English Alphabets

Contents of EnglishAlphabets.txt

#FONT ARIAL

0 A

1 B

2 C

3 D

4 E

5 F

6 G

7 H

8 I

9 J

10 K

11 L

12 M

13 N

14 O

15 P

16 Q

17 R

18 S

19 T

20 U

21 V

22 W

23 X

24 Y

25 Z

EOF



Lipi Toolkit

DATA COLLECTION TOOL 1.0.1

USER MANUAL



**HP Labs, India**

---

## 4 Glossary

LipiTk	<b>Lipi Toolkit</b>
DCT	<b>Data Collection Tool</b> – A tool for collection of handwriting data shipped with LipiTk 1.0
HWR	<b>Hand-Writing Recognition</b>
Stroke	The sequence of pen points between two consecutive pen events, pen down and pen up (can we find a standard definition somewhere ??)
UNIPEN 1.0	A standard format from the International Unipen Foundation ( <a href="http://www.unipen.org">www.unipen.org</a> ) to store on-line handwriting data (as digital ink) and its annotations.