Terms of Reference – Optical Character Recognition, Sinhala

1. *Background:*

The ICT Agency of Sri Lanka (ICTA) has been promoting the use of ICT in Sinhala and Tamil, and has been addressing issues relating to standard fonts and keyboards in Sinhala and Tamil.

The objective is to ensure that the benefits of ICT should be taken to the majority of the population in Sri Lanka which comprises people who prefer to use ICT in Sinhala or Tamil, if they have a choice.

Earlier, applications used their own fonts. No standard was used in the industry. This gave rise to private, non-standard solutions and to a large number of proprietary codes for fonts. Therefore documents produced using one application could be accessed and used only through that application. When accessing a Sinhala website, various legacy fonts had to be downloaded. Otherwise the websites were displayed as indecipherable jargon. This was a major problem when a person tried to use a document created by another, which had been produced using a different font. The font had to be sent to the recipient together with a Sinhala document, unless the recipient already had the font. This made the use of Sinhala email impractical, and slowed the use of Sinhala on the web. Specific applications such as word processing, did not integrate with other applications, and functions such as sorting, were not standardized among applications. There was no way in which Sinhala content could be developed for the Internet. It was not possible to search, or to sort.

But now it is possible to type in Sinhala and Tamil, exchange information in Sinhala and Tamil using computers and browse the web in Sinhala and Tamil. New avenues are now open in the use of ICT for most people in Sri Lanka.

The way of getting out of the disorder caused by the use of numerous non-standard solutions was standardization. The only available international standard for a language character set is Unicode (*Uni*versal *En*coding). The Unicode standard includes all the world's languages. At present with the implementation of projects under ICTA's Local Languages Initiative (LLI), and by stakeholders, enablers for local language computing are in place.

Other enablers for Sinhala and Tamil include Optical Character Recognition, Text to Speech systems etc. (OCR, Speech, and Printing etc). The present project envisages developing an optical character recognition system for Sinhala, whilst ensuring adherence to standards.

2. *Objective(s) of the Assignment:*

Develop Optical Character Recognition system for Sinhala, thus facilitating the development of digital content.

3. *Scope of Services, Tasks to be carried out and Expected Deliverables:*

Develop a Sinhala Optical Character Recognition system with the following features:

a. Should be able to detect Sinhala characters in images, including jpg and png.

b. Should be able to detect Sinhala characters in PDF documents.

c. Should be able to detect Sinhala characters in printed books with a minimum of 80% accuracy and independent of the font.

d. Should be able to recognize Sinhala characters in conformance to Level 3 Fonts defined in the *Sri Lanka Sinhala Character Code for Information Interchange, SLS 1134; Part 2 : Requirements and Methods of Test.*

e. Should be able to give suggestions for correcting spelling to the user, which the user could select if necessary.

f. Should have a minimum level of 80% accuracy.

g. The software should be provided as a service and should have REST API, which should have clean documentation and provided as a SDK.

4. ***Deliverables:***

Optical Character Recognition (OCR) software with ability to detect Sinhala characters accurately in images.
OCR software with ability to detect Sinhala characters accurately in pdf documents and in printed books.
Final OCR software for Sinhala with a feature for suggesting spelling alternatives, (with source code); and available as a service. Comprehensive user manual and technical documentation.

5. ***Qualification Requirements for the Consultant:***

Key Expert	**Minimum qualifications**	**Minimum Experience**
Team Leader.	A bachelor's degree in computer science or IT from a recognized University.	10 years experience in working in the area of Unicode compliant local language computing. Proficient in OCR technology. Proficient in the Sinhala language.
Key staff	A bachelor's degree from a recognized University.	Five years experience in working in the area of Unicode compliant local language computing. Experience in OCR development. Proficient in the Sinhala language.
Key staff	A bachelor's from a recognized University.	Three years experience in working in the area of Unicode local language IT projects. Experience in OCR development. Proficient in the Sinhala language.

6. Final outputs:

- The optical character recognition system for Sinhala is to be delivered with source code and technical documentation in two DVDs.
- The comprehensive user manual should be delivered as both a softcopy and as a printout.

7. Client's Input and Counterpart Personnel:

Data, services and facilities to be made available to the Consultant by the Client:

The client will make available the Sri Lanka Standard Sinhala Character Code for Information Interchange, SLS 1134, Requirements and Methods of Test.

The client will set up a review team with the necessary expertise to review the project while it is being developed and ensure quality assurance. The client will set up project review meetings as necessary.

8. Composition of review committee and review procedure to monitor consultants work:

The review team will comprise members with qualifications and expertise in:

- Experience in work relating to Unicode compliant local languages and ICT.
- Sinhala language expertise.