Document No: - Character Encoding: 01

Version: 1.0
November, 2009

Character Encoding Standard

For

Indian Languages

Government of India
Department of Information Technology
Ministry of Communications and Information Technology
New Delhi – 110 003

CONTENT

1.0 SCOPE 3

1.1 Purpose 3

1.2 Issues and Challenges 3

2.0 TARGET AUDIENCE 4

3.0 TYPE & ENFORCEMENT CATEGORY OF STANDARD 4

4.0 DEFINITIONS AND ACRONYMS 4

5.0 RECOMMENDED ENCODING STANDARD 4

5.2 Unicode as Character Encoding Standard 4

6.0 PROCEDURE TO BE FOLLOWED FOR IMPLEMENTATION / ENFORCEMENT 5

Annexure-I Definitions and Acronyms 5

9.0 LIST OF EXPERT COMMITTEE MEMBERS 6

1.0 SCOPE

The Government of India has launched the National e-Governance Plan with the intent to support the growth of e-Governance within the country. The Plan envisages creation of right environment to implement G2G, G2B, G2E and G2C services.

The lack of availability of information in the locally understandable language is the main reason for the slow progress in the Information and Communication Technology (ICT) sector. In today's age, access to ICT plays a major role in the overall development of a country, it has become a challenge to bridge the digital divide caused by the language barrier.

Standardisation is one of the baselines to be followed in localisation. Standardisation means to follow certain universally accepted standards, so that the developers from any part of the globe could interact through the application. Standardisation becomes applicable in almost everything specific to the language – for instance, a standard glossary of terms for translation, a standard keyboard layout for input system, a standard collation sequence order for sorting, a standard font etc.

The Department of Information Technology, Government of India has constituted an "Expert Committee" on "Technology Standards for Indian Languages" to address specific areas of the Localisation issues. The draft report prepared on the basis of Expert Committee meetings and comments received from chairperson and members of the committee.

1.1 Purpose

Character Encoding standard for all constitutionally recognized Indian Languages should be such that it facilitates global data interchange.

1.2 Issues and Challenges

ISCII is the National Standard and Unicode is the global character encoding standard. The average data packet size is less for representation of any Indian languages using ISCII. However, being limited code space, coexistence of multiple languages within the same code page is not possible in ISCII. The migration from ISCII to Unicode has become imperative due to the following reasons:

- All major operating systems, browsers, editors, word processors and other applications & tools are supporting Unicode.
- It is possible to use Indian languages and scripts in the Unicode environment, which will resolve the compatibility issue.
- The documents created using Unicode may be searched very easily on the web.
- As Unicode is widely recognized all over the world and also supporting Indian languages, it will ease Localization applications including e-Governance application for all the constitutionally recognized Indian languages.

- Since Indian languages are also used in the other part of the world, it is possible to have Global data exchange.

2.0 Target Audience

e-Government Services providers in all constitutionally recognized Indian Languages from Government to Citizens.

3.0 Type & Enforcement Category of Standard Document

Type of Standard Specification - Enforcement Category: Mandatory

4.0 Definitions and Acronyms

Refer Annexure – I

5.0 Recommended Encoding Standard

5.1 Background

TDIL Programme, Department of Information Technology, Government of India is the voting member of the Unicode Consortium since year 2000 to ensure adequate representation of Indian scripts in the Standard. DIT is participating in the Unicode Technical Committee meetings regularly and taking up the issues with Unicode Consortium with respect to representation of Indian languages/ scripts in the standard.

5.2 Unicode as Character Encoding Standard

Unicode shall be the storage-encoding standard for all constitutionally recognised Indian Languages including English and other global languages as follows:

Specification Area	**Standard Name**	**Owner**	**Nature of the Standard**	**Nature of Recommendations**	**Indian Liaison**
Character Encoding for Indian Languages	Unicode 5.1.0 and its future up-gradation as reported by Unicode consortium from time to time.	Unicode Consortium, Inc.	Matured	Mandatory	TDIL Programme HCC Division, MC&IT, Dept. of IT New Delhi 110001

6.0 Procedure to be followed for Implementation / Enforcement

Unicode encoding is available at OS level; instructions should be issued to users how to enable language in Operating system. Training Programmes may be conducted to train different classes of users.

Dissemination of facility to work in Indian Languages by enabling Unicode at Operating System level would help uniformity and avoid duplications.

7.0 Annexures

Annexure-I Definitions and Acronyms

Character: Character is the smallest component of any written language that has semantic value.

ISCII: Indian Script Code for Information Interchange (ISCII - IS 13194:1991) is the character-encoding standard approved by Bureau of Indian Standards (BIS). ISCII is an 8 bit-encoding scheme, catering to 128 code spaces for representation of Indian languages. Nine Indian scripts are included in ASCII standard to represent 10 Indian languages i.e. Assamese, Bengali, Devnagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil and Telugu.

Unicode: Unicode is a 16-bit character-encoding standard. All the major written scripts of the world are included in the Unicode Standard. The first version of Unicode was published in year 1991.

Unicode vis-à-vis ISO10646

Unicode is a 16 bit Character Encoding standard. All the major written scripts are included in the Unicode Standard. Indian scripts are also included in the standard. There are 22 constitutionally recognised Indian Languages, written in 12 different scripts. ISO/IEC 10646 is the character-encoding scheme evolved by the International Organisation for Standardisation (ISO) in 1990.

In 1991, the ISO Working Group responsible for ISO/IEC 10646 (JTC 1/SC 2/WG 2) and the Unicode Consortium decided to create universal standard for coding multilingual text. Since then, the ISO 10646 Working Group (SC 2/WG 2) and the Unicode Consortium are working together closely to extend the standard and to keep their respective versions synchronised.

In addition to the code tables as per ISO/IEC 10646, the Unicode Standard also provides an extensive set of character specifications, character data, algorithms and substantial technical material, which is useful for developers and implementers.

8.0 References

(i) http://www.unicode.org/versions/Unicode5.1.0/

(ii) http://unicode.org/

9.0 List of Expert Committee Members

S. No.	Name & Designation	Organisation
1.	Prof. Rajeev Sangal, Director & **Chairman**	IIIT, Hyderabad
2.	Smt. Swaran Lata, Director & HoD	TDIL Programme, HCC Division, DIT, MC&IT
3.	Shri Mahesh Kulkarni, Coordinator, GIST Group	CDAC, Pune GIST Group
4.	Dr. Mukul K. Sinha, Managing Director	Expert Software Consultants Ltd.
5.	Shri Kewal Krishan, Technical Director, **Nodal Officer**	NIC