No.68/Digitalizing Kerala's Past/1045/KCHR/2012 07/12/2012

QUOTATION NOTICE

Sub: Digitalization of the Kerala District Gazetteers 10 Volumes –inviting quotations–reg

Sealed Quotations are invited for digitalizing the Kerala District Gazetteer 10 volumes. The proposal and the details of each volumes and the terms and conditions are as follows;

Digitization of Kerala District Gazetteers

PROJECT SPECIFICATIONS

1. Scope

The project proposes the digitization of the ten volume information compendium on Kerala, The Kerala District Gazetteers (KDG) which was last printed about three decades back. The project will be named and referred as GAZETTE and will have the following stated objectives.

1. Digitize the KDG volumes for its historical value.
2. Disseminate the digitised contents, tapping the internet resources.
3. Provide for public participation by means of commenting.
4. Attempt, updation of content through experts' coordination.

This document details the specifications for digitization along with other project details.

2. Digitised deliverables

The digitization project will produce

1. PAGE IMAGES of the complete ten volumes of the District Gazetteers to preserve the historical value of the last printed edition.
2. eTexT of all the volumes - The data of the last printed KDGvolumes shall be converted to electronic text using OCR or manual entry, or a mix of both or any other optimal methods which will meet the objective.

2.1 Page images

1. ARCHIVAL IMAGES – master scans of each volume - all pages of each volume, inclusive of covers (outside front and back) if they carry information.
2. ACCESS IMAGES – Derivative images from master scans - for viewing in various devices.

2.1.1 Archival Images

1. Resolution: 400 ppi
2. Type of image: Colour (RGB)
3. Bit-depth: 24-bit
4. Colour management: Embedded ICC Colour Profile
5. Compression: Lossless
6. File format: TIF, uncompressed

2.1.2 Access Images – I

1. Resolution: 400 ppi
2. Type of image: Colour (RGB)
3. Bit-depth: 24-bit
4. Colour management: Embedded ICC Colour Profile
5. Compression: Lossless
6. File format: JPEG

2.1.3 Access Images – II

1. Resolution: 200 ppi
2. Type of image: Colour (RGB)
3. Bit-depth: 24-bit
4. Colour management: Embedded ICC Colour Profile
5. Compression: Lossless
6. File format: JPEG

2.1.4 Access Images – III

1. Resolution: 200 ppi
2. Type of image: Grayscale
3. Bit-depth: 8-bit
4. Colour management: Embedded ICC Colour Profile
5. Compression: Lossless
6. File format: JPEG

2.1.5 Access Images – IV

1. File format: Djvu. Details.
2. Each volume should be a bundled multipage DjVu document with metadata pertaining to the volume embedded.
3. Resolution shall be 400 ppi.
4. Pages can be in bitonal mode, provided consistent and comfortable reading quality is achievable across all volumes.

2.2 Specifications for image capture, clean up and delivery

1. Images should be of correct orientation, without any tilt or skew.
2. Care should be taken to get the best possible flattened surface of the page before scanning.
3. Images are to be cropped, keeping the page numbers intact.
4. While cropping, care should be taken to include printed material outside the main body of text.
5. A reasonable margin shall be maintained. Margin shall be uniform across a single volume.
6. All images of a volume shall be of the same dimensions and the same shall hold good for all 'sets' of image deliverables.
7. No image processing (sharpening, tonal/colour correction etc.) shall be done on archival images.
8. Access Images shall be enhanced sufficient enough to improve readability.
9. One image per page shall be the output. i.e. each numbered page shall have one corresponding image.
10. Unnumbered pages shall also follow the same rule as above - one image per page - but shall be numbered as fm-1, fm-2 or bm-1, bm-2 etc. dpending on whether the content is front matter or back matter of a particular volume.
11. Page Image file names and directory structure shall remain the same across all 'sets' of deliverable images.
12. File or directory names shall not contain spaces or special characters.
13. Fully hyperlinked index page (in HTML format) shall be provided for each set of access image deliverables.
14. Access images may very well be generated by automated processes from a cleaned and corrected intermediate master.
15. Image deliverables shall be viewed, checked and corrected for presence of uniform quality. Image deliverables shall be delivered on a compact disk or DVD in confirmance with delivery schedules.

2.3 eTexT

eTexT refers to digitised textual content as strings of characters in UTF-8 encoding and its presentations in various well accepted standard formats as listed below:

2.3.1 eTexT Formats

1. NLM 3.0 XML for archival and future usage
2. PDF/X-1A for print
3. PDF/A-1B for web
4. HTML5 for web
5. MediaWiki format, for collaborative editing
6. ePub for e-book readers and small devices

Digital resources can be generated in a multitude of formats depending on the requirement and life span of the content. If it is meant for one of print generation like a two page report or a memo or a personal letter, any binary word processor format will suffice, because, the longevity expected is a few days since re-using the document is not needed. It is when the document needs to be preserved and needs to be re-used by future generation with then invented technologies and devices, the documents need to be in platform neutral and application neutral formats. XML, item 1 above is one such format, a brief introduction can be found here while an exhaustive one is available at this Wikipedia entry.

PDF/X-1A is an ISO standard that defines a subset of the popular PDF standard to ensure universal printing of the document in any printing device. The standard imposes certain restrictions and looks for certain properties in the document so that it does not create any printing errors like non-compliant colorspace (printer needs CMYK colorspace while the document might have RGB colorspace), fonts may not be subset, thickness of lines does not have the required 0.4pt or more, so on and so forth. Hence, an X-1A compliant PDF is necessary if it is intended for universal printing.

PDF/A-1B identifies a profile for electronic documents that ensures that the documents can be reproduced exactly the same way in years to come. A key element to this reproducibility is the requirement for PDF/A-1B documents to be 100% self-contained. All of the information necessary for displaying the document in the same manner every time is embedded in the file. This includes, but is not limited to, all content (text, raster images and vector graphics), fonts, and color information. A PDF/A-1B document is not permitted to be reliant on information from external sources (e.g. font programs and hyperlinks). Usage of standards based metadata likeDublin Core Metadata, no javascript for ensuring security, device-independent colorspace, no encryption are some of the mandatory properties.

HTML5 is a markup language for structuring and presenting content in the World Wide Web and is the core technology of the Internet. It is the fifth revision of the specification which supersedes all other previous versions with lot more advanced features.

MediaWiki is a free web based wiki software application developed by Wikimedia Foundation for its projects including Wikipedia. MediaWiki provides an intuitive and easy to understand interface with which anybody can easily write, edit or manage long chunks of textual content collaboratively without constrained by national boundaries with facilities for moderating and controlling the persons involved in the work. MediaWiki which powers the

Wikipedia, is also useful to render our documents in any browser in an elegantly appealing fashion.

ePUB (short for electronic publication) is a free and open e-book standard by the International Digital Publishing Forum (IDPF). Files have the extension .epub. ePUB is designed for reflowable content, meaning that an ePUB reader can optimize text for a particular display device. With the advent of smart phones with different display capabilities and tablets, ePub format assumes much importance for any content that is intended for publishing.

2.4 Specifications for eTexT

1. The eTexT deliverables shall have UTF-8 encoding.
2. The digitised text shall be proof read and corrected to match the originals.
3. Differences in contents in various eTexT versions will not be permissible.
4. eTexT deiliverables shall confirm to standards (as mentioned in 2.3.1) and reports of confirmance shall be submitted along with the deliverables.
5. The digitising agency shall have the freedom of transliterating the vernacular terms in the eTexT versions.
6. Anomalies/errors found in the original texts shall not be corrected except in consultation with KCHR.
7. eTexT deliverables shall have the provision of invoking display of original page numbers, if requested by KCHR during the process of digitisation.
8. MediaWiki format should follow the guidelines provided at the MediaWiki site for formatting.
9. Pictures forming part of KDG volumes
    (a) shall appear in its original order in print deliverable - Section 2.3.1 - Item 2. Sepcifications shall be as specified in 2.1.2.
    (b) shall be hyperlinked in electronic deliverables, if cited in original text (applicable to all items under Section 2.3.1 except Item 2)
    (c) shall be shown as a thumbnail image, around 150-200 pixels on the longer side, close to the place of citation in HTML/MediaWiki deliverables (Section 2.3.1 - Items 4, 5), with hyperlinks to higher resolution versions.
    (d) for non-print deliverables, image specifications shall be as specified in Section 2.1.3
    (e) the pictures shall not be retouched
    (f) maps and other folded material shall be handled in the same manner as pictures
10. Source files used to create the eTexT shall form supplementary deliverables.

11. Design and layout of the eTexT deliverables shall be finalised in consultation with KCHR.

3 Quoting, costing and other notes

1. Dimension of a typical page of KDG will be approximately 6 in x 9 in.
2. Volumes of KDG can be seen prior to quoting, after fixing an appointment with KCHR librarian.
3. An item-wise cost per page or group of pages shall be submitted.
4. Quoting for part of listed deliverables will not permissible.
5. From awarding of work/contract the output (delivery) shall be at the rate of one volume per month from the end of second month onwards.
6. A detailed note on the workflow to be adopted has to be submitted for eTexT deliverables, including softwares to be used (with version details) as well as Operating System.
7. In case of ties in quotes, decisions will be in favour of agencies relying on Free and Open Source Software (FOSS) for full or substantial part of workflow.
8. On submission of the quote, material necessary to create a sample of all listed deliverables will be issued.

I. Details of the Gazetteers

Sl	Title	Total Pages	Size & print nature
1	Thiruvananthapuram District Gazetteer	834	1/8 dummy, B/W
2	Quilon District Gazetteer	753	1/8 dummy, B/W
3	Aleppy District Gazetteer	684	1/8 dummy, B/W
4	Kottayam District Gazetteer	640	1/8 dummy, B/W
5	Ernakulam District Gazetteer	926	1/8 dummy, B/W
6	Thrichur District Gazetteer	684	1/8 dummy, B/W
7	Malappuram District Gazetteer	926	1/8 dummy, B/W
8	Palghat District Gazetteer	692	1/8 dummy, B/W
9	Kozhikode District Gazetteer	861	1/8 dummy, B/W
10	Cannanore District Gazetteer	791	1/8 dummy, B/W

All these district gazetteers are available at the KCHR Library.

Quotation rates per volume and other terms and conditions may please be forwarded to KCHR office within 30 days from the date of this notice.

Director
K.C.H.R

Copy to,

1. KCHR website- www.keralahistory.ac.in
2. PRD
3. Notice Board
4. Notice board of the Directorate of Archives, Kerala Bhasha institute/ MPCC/ Lalitha kala Academy/ Sangeetha Nataka Academy/ Bharath Bhavan