Digital revival of the text of the Bhāratamañjarī and Rāmāyaṇamañjarī

TANUJA P. AJOTIKAR
Assistant Professor, The Sanskrit Library
Co-director Sanskrit Dept. India Programs
Maharishi Vedic University
Vlodrop, The Netherlands
tanuja@sanskritlibrary.org

PETER M. SCHARF
President, The Sanskrit Library
Adjunct Professor, IIIT Hyderabad
scharf@sanskritlibrary.org

Abstract

This paper outlines the meticulous, multi-stage workflow developed for creating scholarly Text Encoding Initiative (TEI) digital editions of the Sanskrit texts *Rāmāyaṇamañjarī* and *Bhāratamañjarī*. The transition from print to digital for these texts presented distinct challenges that necessitated a critical approach beyond simple character conversion. For the *Rāmāyaṇamañjarī*, which was available only in the form of images, the process began with Optical Character Recognition (OCR), leading to the first-ever systematically marked-up digital character data. This data was corrected and sandhi-analyzed manually. The *Bhāratamañjarī*, though digitally available, required extensive text corrections and verification against the printed source to produce a highly accurate dataset. We also included the identification of metrical patterns. The resulting TEI digital texts from our project not only preserve textual heritage but actively enhance it, creating a robust foundation for future digital humanities scholarship and broad public access.

1 Introduction

The migration of Sanskrit literary heritage from traditional print form to digital environments presents unique and complex challenges. These challenges include the difficulty of ensuring OCR accuracy, the complexity of standardizing metadata, and the need to reconcile textual variations present in printed editions. The Sanskrit Library has developed a systematic method to make this migration possible. The steps are as follows:

1. Acquire images in JPEG, PNG, or PDF form.
2. Obtain the digital character data of the images of the chosen edition through high-quality Optical Character Recognition (OCR) using available software.
3. Correct OCR errors through meticulous human verification.
4. Transcode to the Sanskrit Library Phonetic ASCII encoding (SLP1) using the Sanskrit Library's TranscodeFile Java program locally.
5. Transform the raw character data into structured XML markup in accordance with the Text Encoding Initiative (TEI) guidelines. This stage involves:
   (a) using regular expressions and replacement expressions to identify metrical patterns and transform the corrected data into structured TEI markup (semi-automated by TEITAgger), and
   (b) preserving the original spacing between verse quarters in an explicit `space` element.
6. Analyze Sandhi in the marked-up data.
7. Annotate morphology and syntactico-semantic relations using The Sanskrit Library tagsets in accordance with TEI guidelines.

Here we are going to present our work of transforming two rare editions to digital form, namely the *Rāmāyaṇamañjarī* (*RM.*) and *Bhāratamañjarī* (*BhM.*), both composed by Kṣemendra. Kṣemendra is a well-known Śaiva scholar who flourished in 11th CE in Kashmir. Among his many literary works, he composed three summaries: the *Rāmāyaṇamañjarī*, alias *Rāmakathāsāra*, the *Bhāratamañjarī*, and the *Bṛhatkathāmañjarī*. The *Rāmāyaṇamañjarī*, a summary of the great epic *Rāmāyaṇa*, was composed in 6,198 verses and published in 1903 by Nirṇayasāgara Press (Śāstrī and Paraba 1903). The *Bhāratamañjarī*, a summary of the great epic *Mahābhārata*, consists of 10,666 verses and was first published in 1898 by Nirṇayasāgara Press (Śarmā and Paraba 1898). An edition of the *Ādiparvan* and *Sabhāparvan* was published in 1918 by The Standard Publishing Company (Bhandare 1918). Finally, the *Bṛhatkathāmañjarī* is a summary of the *Kathāsaritsāgara*, which contains 7,787 verses published in 1901 by by

Nirṇayasāgara Press (Paṇḍit and Paraba 1901). In the next sections, we discuss the challenges we faced at each step in the transformation of the *RM.* and *BhM.*

2 OCR

2.1 OCR of the *RM.* and proofreading

A PDF image file of the *RM.* was available at `archive.org`. The OCR of the same was obtained using the OCR service developed by The Sanskrit Research Institute in Auroville available at `https://ocr.sanskritdictionary.com` as recommended by Ravi Kiran Sarvadevabhatla at IIIT Hyderabad and one of the options mentioned by Tyler Neill (Neill 2024). This website simply states, "you can paste, drop, or upload your image," but unfortunately does not specify the required image format. Since it did not recognize files in PDF format for upload, we used only PNG images. The output is generated in a plain text file.

This plain text output does not retain the formatting of the original page. The printed edition of the *RM.* preserves the metrical form of the text by printing each half verse on a new line. This formatting is completely distorted in the output: original text lines are incorrectly merged or broken up. The page numbers printed on the original page, either in the upper left or upper right corner, are not consistently placed on a new line; they sometimes get mixed with the text. Superscript footnote numbers printed above a Devanāgarī character are either omitted from the output or erroneously interpreted as an ascending vowel sign (*mātrā*). Hence, each of the footnote numbers in the text had to be restored manually both in-line and in the footnote section at the bottom of the page where the note had to be separated from the main text. The omission of visarga and anusvāra is quite common. Consonant clusters with more than two consonants are not processed correctly. The ascendant sign of consonant-cluster-initial र् was not transformed correctly. The consonant cluster स्व was always replaced by ख. The consonant cluster with ङ् as in उल्लङ्घ्य or आलिङ्ग्य was always transformed into two syllables, उल्लङ्घ्य or आलिङ्ग्य. Many Devanāgarī characters are interchanged. The characters which are interchanged frequently are listed below:

- भ् and म्
- य् and प्
- स् and म्
- व् and ब्
- घ् and ध्

To summarize, we meticulously proofread the OCR output. Apart from correcting the text, we restored the original page numbers, footnote numbers in the text, and footnotes at the end of each page.

2.2 The digital character data of *BhM.* and proofreading

We did not need to perform OCR on the edition of *BhM.* because its digital character data is available at `https://www.ebharatisampat.in`. This website preserves this text in three different URLs. Only images in PDF format of the 1898 edition are available at `https://www.ebharatisampat.in/readSearch.php?id=NjQ2MDUzMjMyNjAxNjU1`. The devanāgarī character data in Unicode of the 1898 edition is available to read and download at `https://www.ebharatisampat.in/readbook3.php?bookid=ODYwMzk1ODIyMjAyNjU1&pageno=MjI0MjQyNjk5NTk=`. The website does not specify if the data is proof-read or not. The same text is displayed in a navigable HTML webpage at `https://www.ebharatisampat.in/read_chapter.php?bookid=MDIxOTY1Nzc4ODQxNDgy`, where the details of the source edition are missing.

As part of our online course UT102 "Character and higher-level encoding", we decided to perform an experiment. We downloaded the text available in Unicode Devanāgarī characters from `https://www.ebharatisampat.in/readbook3.php?bookid=ODYwMzk1ODIyMjAyNjU1&pageno=MjI0MjQyNjk5NTk=`. The total length of the text of the *BhM.* in the printed edition is 849 pages. We distributed it among six participants. Each participant got the character data of nearly 142 pages to compare against the original printed text. Our observations about the data and corrections to it are as follows:

1. The downloadable text lacked page numbers, footnotes, and footnote numbers. Therefore, our team added those according to the printed edition.
2. Eighteen pages (262, 263, 439, 460, 461, 465, 520, 530, 531, 576, 609, 614, 648, 690, 752, 798, 832, 833) were missing. The previous page repeated in place of these pages. Our team added the missing pages by doing OCR and proofreading them.
3. Errors such as missing visarga and anusvāra, interchange of devanāgarī characters, wrong mātrās, and omission of syllables are commonly observed.

The text, which is displayed as a navigable HTML page (`https://www.ebharatisampat.in/read_chapter.php?bookid=MDIxOTY1Nzc4ODQxNDgy`), is downloadable in a plain text file, but it does not preserve the original line breaks, making it difficult to use for further mark-up without regularization. We transcoded the devanāgarī character data into SLP1 by running The Sanskrit Library's TranscodeFile routine locally. The raw character data in SLP1 was then regularized using regular expressions. We will discuss the procedure of correcting this data in a subsequent section.

The Digital Corpus of Sanskrit (DCS `http://www.sanskrit-linguistics.org/dcs/`), brought to our attention by the reviewers of this article, contains most of the text of the *Bhāratamañjarī* digitized and analyzed by Oliver Hellwig. The DCS edition of the text is missing parvans 2–4 and the first eight verses of the first parvan. This edition is well proofread, incorporates word-separation as standard in Romanized text, and includes morphological analysis. Yet it has serious shortcomings. It does not identify its printed source, which we identify as based on the 1898 edition. The DCS edition eliminates all notes and editorial markup and inconsistently adopts or rejects editorial suggestions. Consider the following examples from the *ādiparvan* comparing the text as in the 1898 edition and DCS:

- 10c मान्धातृना(रा)म
  DCS *māndhātṛrāma*
  adopts the editorial correction of *nāma* to *rāma*.
- 17a तरत्तुरंग(त्तरङ्ग)
  DCS *tarattaraṅga*
  adopts the editorial correction of *turaṁga* to *taraṅga*.
- 26d न चिरादिति मा(सा)ब्रवीत्
  DCS *na cirāditi abravīt*
  eliminates the correction of *mā* to *sā* and instead creates hiatus *iti a*.
- 382c अभिम्वा(अनश्वा)
  DCS *abhimbā*
  eliminates the correction and changes *v* to *b*. The morpholical analysis identifies the form as masculine nominive singular of a non-existent *abhimban*.
- 389ab धवलोड्डु(?)ग्गूल
  DCS *dhavaloḍuggūla*
  eliminates the question mark which however is preserved in the morphological analysis of *uḍḍuggūla* '???typo?'. The 1918 edition corrects the reading to धवलो दुकूल.

Moreover, DCS originates fresh errors in the text. These errors replace legitimate text with similar more common forms, either deliberately or by an AI-driven process. Consider the following examples:

- 34b भैक्ष्यं भोक्तुं यतव्रतः
  DCS *bhakṣyaṁ* ....
- 174a सप्तमे ऽहनि संप्राप्ते
  DCS ... *samāpte*
- 174c निर्दग्धुं पृथिवीपालं
  DCS *nirdeṣṭuṁ* ....

Despite these shortcomings, Our procedure would have incorporated complete comparison of fresh OCR data with the DCS edition had we been aware of it; for comparison of two independent sources always leads to superior quality data.

3 Marking up in TEI

In the previous section, we discussed the procedure for obtaining character data, and proofreading and correcting it to match the printed edition. In this section we describe the procedure for transforming the corrected raw character data into XML mark-up according to the Text Encoding Initiative guidelines. A method using regular expressions in a text editor's `find` field—paired with a replacement statement (in the `replace` field) to transform text into TEI markup—is described at length by Scharf (2018) in his work on the TEITagger. We train our students in our course to use this effective method to transform a metrical text into TEI marked-up text. The following is a sample of the proof-read text, and the transcoded text.

> अपि नन्दनभृङ्गानामुद्भूतैः कमलानिलैः ।
> विदधानमिवामोदमहोत्सवनिमन्त्रणम् ॥ ११२८ ॥
>
> api nandanaBfNgAnAmudDUtEH kamalAnilEH .
> vidaDAnamivAmodamahotsavanimantraRam ..1128..

Figure 1 shows a sample of the regular expression and its replacement expression.

Figure 1
Regular expression and replacement expression

After applying the regular expression, the transformed text has the following structure. Each verse is wrapped in an `lg` element, and each line of verse in an `l` element. Each quarter is placed in a `seg` element. The meter-type is identified, and the verse number is inserted in and `n-attribure` and as part of the `xml:id-attribute` as shown in Figure 2. The process involves preserving the original space between the two quarters in a verse-line. We then replace this space with an empty `space` element to ensure it is represented explicitly and not just retained as whitespace. We must exercise caution with long texts. Since our texts contain both three-line and two-line Anuṣṭubh verses, we used the following procedure: first, mark up the three-line Anuṣṭubh verses by modifying the regular expression; then, run the regular expression for the two-line Anuṣṭubh verses. To identify verses in different meters, the annotator should adjust the syllable count in the regular expression. However, the regular expression does not work with verses containing brackets (for editorial corrections) or bracketed footnote numbers. With all these careful steps, over 70 percent of the verses in a single meter are successfully tagged with a single application of the regular expression in a large, metrically composed text.

Figure 2
Text marked up in TEI

```
<lg type="anuzwuB" n="1" xml:id="rm.k1.v1">¬
 <l>¬
  <seg type="foot" n="a">jitaM BagavatA tena</seg>¬
  <seg type="foot" n="b">hariRA lokaDAriRA .</seg>¬
 </l>¬
 <l>¬
  <seg type="foot" n="c">ajena viSvarUpeRa</seg>¬
  <space/>¬
  <seg type="foot" n="d">nirguRena guRAtmanA ..1..</seg>¬
 </l>¬
</lg>¬
```

After successfully running the regular expression, most of the well-formed verses are marked up in TEI format. This stage also exposes verses with typos, specifically extra syllables or lack of syllables. This stage also reveals any originally metrically ill-formed lines.

3.1 Meters in *RM.*

The *RM.* is composed predominantly in the Anuṣṭubh meter. Table 1 provides an account of the total number of verses in each kāṇḍa, the number of three-line Anuṣṭubh verses, and list of other meters found.

Table 1
Meter identification

Kāṇḍa	verses	Additional information
1	987	4 three-line अनुष्टुभ् verses
2	1158	4 three-line अनुष्टुभ् verses
3	214	
4	573	4 three-line अनुष्टुभ् verses, some verses in स्रग्धरा, मालिनी, उपेन्द्रवज्रा
5	705	4 three-line अनुष्टुभ् verses,
6	1248	
7a	193	140 verses in अनुष्टुभ्, others in कीर्ति, ऋद्धि, इन्द्रवज्रा, हंसी, बाला, बुद्धि, श्यामा, सिद्धि, भद्रा, जाया, वाणी, उपेन्द्रवज्रा, आर्द्रा, माया, माला, वसन्ततिलका, मालिनी
7b	1263	

3.2 Meters in *BhM.*

The work of marking up the *BhM.* is in progress. We do not have the data for the whole text. Here we present the data for the *Ādiparvan*. The raw character data was obtained from the navigable HTML webpage at the following link `https://www.ebharatisampat.in/read_chapter.php?bookid=MDIxOTY1Nzc4ODQxNDgy`. We previously noted that this data is downloadable in a plain text file, which does not preserve the original line-breaks. We regularized the file after transcoding it into SLP1. Since SLP1 is an ASCII encoding, we ensured that there are no non-ASCII characters in the data. We found 13 non-ASCII characters in the file, which we removed. We used a regular expression and replacement expression to place each line of verse on a separate line to restore the layout of the printed edition.

Assuming that this text is fully proofread and is free of any typos, we marked up the data using the regular expression that identifies the Anuṣṭubh meter. The 1898 edition of *BhM.* has 1397 verses in the Ādiparvan. This procedure exposed the typos in the text. We report these below.

- Out of 1397 verses, 1249 were immediately identified as being composed in Anuṣṭubh.
- 27 verses are three-line Anuṣṭubh verses.
- 27 verses had incorrect punctuation, which prevented their identification during the initial use of the regular expression.
- Verses 6 and 376 are not correctly formed in Anuṣṭubh.
- 9 verses (569–577) are composed in Vasantatilakā and one verse (1397) is composed in Mālinī.
- 126 verses were corrected during this procedure by adding a syllable or eliminating an extra syllable.
- Verses 117, 802, 1205, and 1302 were intermixed with the subsequent verse so had to be retyped.
- Two unrelated single lines occured after verse 1369 and 1394. These lines are flagged with the element `sic`.
- During validation, verse number 1201 occurred twice; we corrected the second occurrence to 1202.

This effort of correcting the *Ādiparvan* has clearly exposed the textual problems in the *BhM.* The text of the *BhM.* has not yet been definitively established. The 1898 edition was based on a single manuscript. One more edition (Bhandare 1918), including only the *Ādiparvan* and *Sabhāparvan*, used multiple sources identified by various sigla in critical notes; however, no metadata was provided for these sources. The comparison of the text of the *Ādiparvan* in these two editions reveals variations in the total number of verses, the order of verses, and their numbering, as well as variant in reading. The text of the *BhM.* is therefore not yet definitely established and needs to be edited critically. There are at least ten manuscripts of this text listed in NCC vol. 17 (Dash 2007). There has not been a single attempt to publish this text taking into consideration all of the available manuscripts.

4 Adapting footnotes in TEI

We observe that web repositories of Sanskrit texts often display only the main text, omitting editorial notes and the critical apparatus from the navigable HTML page. This tendency significantly limits the resource's usability by academic scholars.

Retaining editorial notes in a digital edition is crucial because they provide essential transparency and context for the reader. These notes document the choices and procedures made during the transition from the source material (a print book or a manuscript). They record procedures employed by the editor(s); such as how errors in the original text were handled, any modifications or additions introduced, any significant variant recorded by a witness, and cross-references. Retaining such notes is vital for academic integrity and reliability, as they allow scholars and careful readers to understand and evaluate the textual history and overall correctness of the text. This ensures the digital edition is a trustworthy resource. Hence, we did not omit any footnote in the digital edition of the *RM.* Below, we discuss issues related to adaptation of footnotes in TEI.

4.1 Adapting editorial corrections and footnotes in TEI in the digital text of *RM.*

4.1.1 Unclear text

The editors of the *RM.* use a question mark enclosed in parentheses (?) to denote unclear text. For example, in verse 304 in the first kāṇḍa, a question mark is placed after the word *mohaśyāmātimiratāṁ*

लज्जालतापरशुतां यशःशीतांशुमेघताम् ।
मोहश्यामातिमिरतां(?) यातः कस्य न मन्मथः ॥३०४॥

We use the TEI element `unclear` to document the doubt posed by the editors regarding the reading. However, the question mark placed at the end of the word does not fully indicate whether the entire word was illegible or just a couple of characters in it. We prefer to wrap the entire word in the `unclear` element to avoid any unnecessary speculation on the annotator's part.

Figure 3
Unclear element

```
<lg type="anuzwuB" n="304" xml:id="rm.k1.v304">¬
 <l>¬
  <seg type="foot" n="a">lajjAlatAparaSutAM</seg>¬
  <space/>¬
  <seg type="foot" n="b">yaSaHSItAMSumeGatAm .</seg>¬
 </l>¬
 <l>¬
  <seg type="foot" n="c"><unclear>mohaSyAmAtimiratAM</unclear></seg>¬
  <space/>¬
  <seg type="foot" n="d">yAtaH kasya na manmaTaH ..304..</seg>¬
 </l>¬
</lg>¬
```

4.1.2 Emendations

The editors of the *RM.* suggest preferable readings or corrections in footnotes. For example, in the *Yuddhakāṇḍa* (6th kāṇḍa) verse 749, the editors offer an emendation in a footnote as follows:

> अयं शरीरवात्सल्ये चिराद्विन्ध्यः समुद्गतः ।
> कर्णकोणे करोत्येष यस्याङ्कः[२] कुण्डलभ्रमम् ॥ ७४९ ॥

The editors suggest that the word *yasyāṅkaḥ* should be *yasyārkaḥ* in footnote no. 2, stating, यस्यार्कः इति स्यात्। So we adapt this suggestion by using elements `del` and `add`. The word originally in the text is wrapped in `del`, and the emendation is wrapped in `add`. So the text appears as in Fig. 4.

Figure 4
Emendation

```
<lg type="anuzwuB" n="749" xml:id="rm.k6.v749">¬
 <l>¬
  <seg type="foot" n="a">ayaM SarIravAtsalye</seg>¬
  <seg type="foot" n="b">cirAdvinDyaH samudgataH .</seg>¬
 </l>¬
 <l>¬
  <seg type="foot" n="c">karRakoRe karotyeza</seg>¬
  <seg type="foot" n="d"><del>yasyANkaH</del><add>yasyArkaH</add> kuRqalaBramam
..749..</seg>¬
 </l>¬
</lg>¬
```

There is one instance where the text is indicated as unclear, with a better reading suggested in a footnote. Specifically, in the second line of verse 1253 in the 6th kāṇḍa,

> निर्जने हरणे स्त्रीणां स्वा[1]पिनाम(?) प्रगल्भते ॥१२५३॥

the word *svāpināma* is marked as an unclear reading with a question mark in parentheses. In footnote 1, editors suggest the reading *cāpitā na* as an emendation for *svāpināma*. This case presents an overlap of two phenomena: an unclear reading and an emendation. We resolved this by choosing the second option (the emendation) and marked up the text as shown in Fig. 5:

There is one instance where the emendation suggested by the editors is metrically awkward. In verse 435 of the first kāṇḍa the reading is मृ[1]चेपुनाहं विक्रीतो. Footnote 1 reads: 'मृचीकनाम्ना' भवेत्. The editors, therefore, in this note suggest that the correct reading is *ṛcīkanāmnā* instead of *ṛcepunā*. If we accept the suggestion, the resulting verse quarter, मृचीकनाम्नाहं विक्रीतो, adds one extra syllable to the original Anuṣṭubh verse quarter, making it metrically faulty. If, on the other hand, the editors meant to

Figure 5
Emendation in the verse 1253

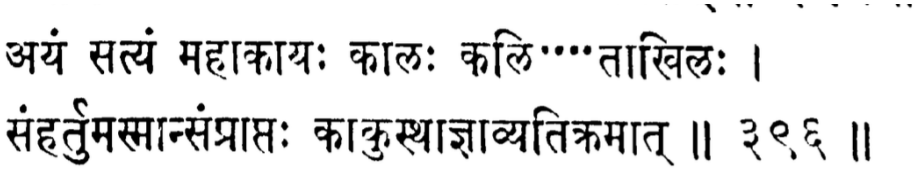
```
        <lg type="anuzwuB" n="1253" xml:id="rm.k6.v1253">¬
         <l>¬
          <seg type="foot" n="a">samare yadi vIro'si</seg>¬
          <seg type="foot" n="b">tatsaMdarSaya pOruzam .</seg>¬
         </l>¬
         <l>¬
          <seg type="foot" n="c">nirjane haraRe strIRAM</seg>¬
          <seg type="foot" n="d"><del>svApinAma</del><add>cApitA na</add> pragalBate
..1253..</seg>¬
         </l>¬
        </lg> ¬
```

suggest that *ṛcepu* should be replaced by *ṛcīka*, it fits the meter and aligns with the choice of reading in verses 426 and 428 where the witness ग reads *ṛcepu*. There is confusion in the text of the *RM.* with a sage named *ṛcīka*. We also find him called *ṛcepu* in the editor's reading of ms. ग in 426 and 428. The reading *ṛcepu* is a clear confusion of the devanāgarī character प् with य्, the latter of which is attested as *ṛceyu* in three other texts by Böhtlingk and Roth. *ṛcīka* is abundantly attested. However, if the editors meant to suggest that *ṛcepu* be read as *ṛcīka* in 435, the note should have read 'ऋचीकेन' भवेत्. What they actually suggested we indicate by using the `del` and `add` elements, and we indicate the correct emendation by using the `sic` and `corr` elements.

4.1.3 Gaps

Gaps in the text are indicated by dots. In TEI, this gap is shown using the `gap` element. The mandatory attributes for the `gap` element are `quantity`, `unit`, and `reason`. We must be very precise in reporting the number of missing syllables, unlike in the printed edition. For example; in kāṇḍa 4 verse 396, four dots are printed to indicate a gap in the second quarter. When we examined this verse, we found

Figure 6
Missing syllables


अयं सत्यं महाकायः कालः कलि····ताखिलः ।
संहर्तुमस्मान्संप्राप्तः काकुस्थाज्ञाव्यतिक्रमात् ॥ ३९६ ॥

that only one syllable is missing because the verse is composed in the Anuṣṭubh meter which requires eight syllables per quarter, but the second quarter has seven syllables, one fewer than the required eight. Therefore, we precisely indicated the gap as follows.

4.1.4 Variant readings

Variant readings are a very common phenomenon in Sanskrit texts which are preserved in various handwritten copies of the text. These readings range from minor differences, such as the omission of a syllable, a missing visarga or anusvāra, to major differences, including the omission or addition of words, sentences, or larger portions, that alter the meaning of the text. The 1903 edition of *RM.* is based on multiple manuscripts, which is evident from the variant readings mentioned in the footnotes. Unfortunately, the details of these manuscripts are not given by the editors. This missing information makes this edition less reliable academically. The sigla क, ख, ग, शा and occasionally कश्मीर point out that the editors had access to multiple manuscripts, but the missing details of these manuscripts make it impossible to know the exact sources.

According to the TEI guidelines, the `TEIHeader` should include the details of the manuscripts used for the edition. These details are included under `listWit` under `sourceDesc`. Since the sigla are

Figure 7
Gap element marked up

```
      <lg type="anuzwuB" n="396" xml:id="rm.k4.v396">¬
       <l>¬
        <seg type="foot" n="a">ayaM satyaM mahAkAyaH</seg>¬
        <space/>¬
        <seg type="foot" n="b">kAlaH kali<gap quantity="1" unit="syllable"
reason="illegible"/>tAKilaH .</seg>¬
       </l>¬
       <l>¬
        <seg type="foot" n="c">saMhartumasmAnsaMprAptaH</seg>¬
        <space/>¬
        <seg type="foot" n="d">kAkusTAjYAvyatikramAt ..396..</seg>¬
       </l>¬
      </lg>¬
```

available in the edition but not the details we tag them `not known` wrapped under the `witness` element.

Figure 8
Siglum in preamble

```
    <listWit>¬
     <witness xml:id="ka">not known</witness>¬
     <witness xml:id="Ka">not known</witness>¬
     <witness xml:id="ga">not known</witness>¬
     <witness xml:id="SA">not known</witness>¬
     <witness xml:id="kaSmIra">not known</witness>¬
     <witness xml:id="kaSmIra">not known</witness>¬
    </listWit>¬
```

TEI offers the following elements and attributes to mark up a critical apparatus:

1. The `app` (apparatus) element is used to group together each lemma and all its variations; it has two child elements: `lem` and `rdg`.
2. The `lem` (lemma) element is an optional child of the `app` element. In this context, the term *lemma* signifies the accepted reading in the base text.
3. The `rdg` (reading) element is a required child of the `app` element used to indicate variations in the base text.
4. The `wit` (witness) attribute specifies which witness supports the reading. This attribute is used in both of the elements `lem` and `rdg`.

So the tagged text looks as shown in Fig. 9.

The editors are inconsistent in recording the witnesses in footnotes. We found many instances where a variant reading is mentioned but the witness is missing. In such cases, we assign the value `xxx` to the attribute `wit`. For example; the first line of the 130th verse of the first kāṇḍa reads,

स्वस्थः श्रीमान्वरं प्रादाद्देशयोर्धन्यता[3]वहम्.

Footnote 3 indicates that there are two variants namely *dhṛṣyatāvaśam* and *vaśyatāvaham* for the word *dhanyatāvaham*; however, there is no mention of the witnesses. Therefore, we record these variants for the mentioned word and assign the value `xxx` to the attribute `wit` as shown in Fig. 10.

Figure 9
Critical Apparatus

```
      <l>¬
       <seg type="foot" n="c"><app><lem>gItasvarAvizwapadO</lem><rdg
wit="ga">gItasvarAvizwaparO</rdg></app></seg>¬
       <space/>¬
       <seg type="foot" n="d">viBaktamaDurasvarO ..29..</seg>¬
      </l>¬
```

Figure 10
Missing witness

```
       <seg type="foot" n="d">ddeSayor<app><lem>DanyatAvaham</lem><rdg
wit="xxx">DfzyatAvaSam</rdg><rdg wit="xxx">vaSyatAvaham</rdg></app> ..130..</seg>¬
      </l><!-- witness of this reading is unknown -->¬
```

5 Grammatical corrections in the digital edition of the *RM.*

Sandhi-analyzed data for digital Sanskrit texts is of fundamental importance for their processing, analysis, and access. Besides creating the raw character data of the *RM.* as described above in Section 2.1, we are creating sandhi-analyzed data for the *RM.* to accompany its digital edition. We have so far completed the sandhi-analysis of the first kāṇḍa (987 verses). The metrically tagged text (saṁhitā) is stored in a separate file, and the sandhi-analyzed data is in the pada file. We corrected numerous grammatical errors in the text during this task. These range from minor phonological corrections to more complex morphological corrections.

5.1 Correcting retroflexion

The edition contains many instances where retroflexion is either unnecessarily added or required but omitted. In some instances, the editors notice these mistakes and suggest the correction in a footnote. We denote these corrections using the `del` and `add` element. If these mistakes are not noted by the editor, we restore the correct form using the `sic` and `corr` elements. While we did not notice these errors in the continuous text, we mark them up in the sandhi-analysed text. For example, in verse 521 of the first kāṇḍa, the instrumental singular form of the word *rāghava* erroneously occurs without retroflexion as *rāghavena*. The editors note this and state in the footnote: राघवेण इति स्यात्। We represent this correction using the `del` element around the dental *na* and the `add` element around the retroflex *ṇa*. The reverse is the case in verse 906 of the first kāṇḍa, where the instrumental singular form of the word *muni* erroneously occurs with retroflexion as *muniṇā*. This is corrected using the `sic` element around the retroflexion *ṇā* and the `corr` element around the dental *nā*.

5.2 Restoring correct vibhakti

The edition records grammatically incorrect forms, and these are mostly unnoticed by the editors. For example; verse 428 of the first kāṇḍa reads as follows:

ततोऽब्रवीच्छुनःशेपं गच्छ राज्ञो महाक्रतौ ।
गाथाद्वयं पठित्वेदं मां स्मृत्वा शुभमाप्स्यति ॥४२८॥

Here the last word *āpsyati* in the second line does not agree with the elided agent *tvam* which is the agent of the verb *gaccha* in the first line. The correct reading should be *āpsyasi*. We correct the verb form using the `sic` element around the word *āpsyati* and the `corr` element around the correct form *āpsyasi*.

Similar is the case in verse 67 of the first kāṇḍa. The verse reads as follows:

ते तमूचुः सुरारातिं वरदानेन वेधसः ।
करोति देवानसुरान् दुर्जयो देव रावणः ॥६७॥

"They said to him who was the enemy of the gods: O Lord, Rāvaṇa who is invincible by the boon of the creator is making all the demons gods."

In this episode all the gods have gathered to convince Viṣṇu to take incarnation on the earth to kill Rāvaṇa. However, as edited, the masculine accusative singular form *surārātim* occurs in apposition with the masculine accusative singular form *tam* which refers to Viṣṇu, and thus Viṣṇu is erroneously described as being the enemy of the gods. In fact, it should be an adjective of Rāvaṇa which occurs in the nominative at the end of the second line. So we corrected the verse with an *r* instead of anusvāra at the end of the word *surārāti* as follows:

ते तमूचुः सुरारातिर्वरदानेन वेधसः ।
करोति देवानसुरान् दुर्जयो देव रावणः ॥६७॥

At some point in the transmission or editing of the text, the ascending cluster-initial *r*, was changed to the sign of an *anusvāra* — a common copyist error. This is corrected using the `sic` element around the syllable *tim* and the `corr` element around the syllable *tir*.

There are many such cases in which the text is improved by correcting forms.

6 Newly discovered words

Sandhi-analyzed data is a stepping stone for further morphological and syntactic analysis. The Sanskrit Library developed a method of annotating morphology and syntax in TEI format. The identified `lemmas` in morphologically annotated file are then linkable to our digital dictionaries. We plan to annotate the *RM.* morphologically and syntactically. So far in our work, we have encountered at least two words in the *RM.* that are not listed in our dictionaries. Occurrences of these two words are as follows: the 658th verse in the fifth kāṇḍa has the genitive singular word *pāvaneḥ* derived from the nominal base *pāvani* which is not listed in any of the printed dictionaries.

धृतिं लेभे परिष्वज्य स श्लाघ्यां पावनेस्तनुम् ।

'He (Rāma) gained courage after hugging the great body of pāvani.'

The word *pāvani* is a taddhita derivate derived from the word *pavana* 'purifier, wind' in the meaning of 'offspring', i.e. 'offspring of the wind'. The word refers to hanūmat, the son of the wind in the *Rāmāyaṇa*. Reviwers suggested checking for the occurrence in DCS. The DCS dictionary does indeed include the word *pāvani* and the DCS corpus includes two occurrences in the *Bhāratamañjarī*.

- 7.464ab तद्बाणदारिततनुः पावनिः कोपकम्पितः।
- 7.775ab इति गर्जन्तमायान्तं पावनिं द्रोणनन्दनः।

Unfortunately the dictionary defines the word as 'Bhīṣma' while both occurrences refer to Bhīma, the son of the wind in the *Mahābhārata*, in accordance with the taddhita derivation.

Another example of a discovered word is an emendation from editors. Verse 1253 in the 6th kāṇḍa,

निर्जने हरणे स्त्रीणां स्वापिनाम(?) (चापिता न) प्रगल्भते ॥१२५३॥

'Abduction of women in uninhabited place is not real heroism (Bow-armedness).'

The word *cāpitā* is derived by addition of the suffix *tal* to the nominal base *cāpin* 'bow-armed'. Even though the word is explicable by derivational procedure it does not occur in dictionaries.

7 Conclusion

In the present project of making TEI digital editions of the *Rāmāyaṇamañjarī* and *Bhāratamañjarī*, we encountered various issues. First of all, the text of the *Rāmāyaṇamañjarī* was not available in digital character data. It was available only in the form of images. Our project made it possible to create its systematically marked-up data along with its sandhi-analysis. The *Bhāratamañjarī*'s text was available in digital form but with many errors. So far, we have corrected the *ādiparvan*. Our project produced data of the same in XML in accordance with the TEI guidelines, true to the printed edition which was its source, with the metrical patterns identified.

We learned during this process that the transition from print to digital text is far more than a simple conversion; it is a meticulous, multi-stage process that fundamentally transforms the nature and accessibility of a document. This paper has outlined the essential workflow, beginning with obtaining images and continuing through the crucial stages of OCR and marking-up the text in accordance with the TEI guidelines in the standard set by The Sanskrit Library. The mere mechanical translation of characters, however, is insufficient for scholarly and archival purposes. The true value and reliability of a digital edition are realized through the application of scholarly editing and academic critical analysis. This process, encompassing careful text correction, verification against the source, and the systematic encoding of the textual apparatus, ensures that the digital text is not just a reproduction, but a faithful and critically informed representation of the original. We respect the efforts and scholarship of the editors in bringing these texts into print form; therefore we preserve their notes. In this paper, we showed how every editorial note is represented in the TEI format.

By rigorously applying these steps—from initial capture to final structurally marked-up text, we not only preserve textual heritage but actively enhance it, creating a durable, searchable, and interoperable digital resource. The resulting digital text becomes a powerful foundation for future scholarship, digital humanities projects, and broad public access, ultimately bridging the gap between historical print culture and the demands of the modern information age.

8 Acknowledgements

We would like to thank Madhura Inamdar for her contribution preparing the sandhi-analyzed data of the *RM.*, and Rasika Vaze for her work preparing the TEI edition of the same. We also thank all the students registered for our course UT102 in the fall semester 2025 for their work correcting the digital character data for *BhM*.

References

Bhandare, M. S., ed. 1918. *The Bhāratamañjarī of Kṣemendra ādi and sabhā parvas. with introduction, full translation, exhaustive notes, appendices and various readings*. Girgum, Bombay: The Standard Publishing Co.

Dash, Siniruddha. 2007. *New Catalogus Catalogorum. An alphabetical register of Sanskrit and allied works and authors*, ed. by Siniruddha Dash. Madras University Sanskrit Series 17. Madras: University of Madras.

Neill, Tyler. 2024. "OCR Options". URL: `https://tylerneill.info/blog/ocr-options`.

Paṇḍit, Śivadatta and Kāśīnātha Pāṇḍuraṅga Paraba, eds. 1901. *The Bṛhatkathāmañjarī of Kṣemendra*. Kavyamala 69. Bombay: Nirṇayasāgara Press.

Śarmā, Śivadatta Paṇḍit and Kāśīnātha Pāṇḍuraṅga Paraba, eds. 1898. *Kāśmīrikamahākviśrīkṣemendrakṛ̥tā Bhāratamañjarī*. Kavyamala 64. Bombay: Nirṇayasāgara Press.

Śāstrī, Bhavadatta Paṇḍit and Kāśīnātha Pāṇḍuraṅga Paraba, eds. 1903. *The Rāmāyaṇamañjarī of Kṣemendra*. Kavyamala 83. Bombay: Nirṇayasāgara Press.

Scharf, Peter M. 2018. "TEITagger: Raising the standard for digital texts to facilitate interchange with linguistic software." *Computational Sanskrit and Digital Humanities. selected papers presented at the World Sanskrit Conference, University of British Columbia, Vancouver, 9–13 July 2018*, ed. by Gérard Huet and Amba Kulkarni, pp. 229–257.