

Issues and Challenges in digitization of Marathi Language Books

Chandrakant Dhutadmal¹, Mahesh Kulkarni², Nilesh Uke³, Rahul Borade⁴
and Deepak Dhore⁵

ABSTRACT

With the increase in easy and affordable availability of digital devices, importance of digital publication is increasing every day. People across the sections of the society are now accepting new media. Publishers have also started bringing out digital editions of their publications to cater to diversified needs of the readers. The scenario of eBooks in Indian languages is not very encouraging but there are efforts from various organizations under government towards bringing out publications in eBook format. The use of digital form of publications is also evident from quite a number of studies worldwide, especially use of eBooks in Schools and, Universities. While it all seems straight forward when converting already published printed versions of books into digital editions, there are various issues and challenges in converting the existing published/ printed books into eBook (ePub, PDF and Mobi) file formats. This paper brings out main issues and challenges in terms of inputting text (typing), validations and proof reading, formatting of the text, graphics, tables, formula's and others as per the originally published books and converting them into eBook formats in Marathi languages. We have made use of Unicode as the storage standard. Devanagari is the script used for writing Marathi language. We have also tried to address some of the non-technical challenges in coping up with the digitization process.

Index Terms: Digitization, Digital Publication, ePub, Mobi, PDF, eBook, Data Entry, Unicode, Marathi Books, eBook reader, Calibre, Sigil, Libre Office

1. INTRODUCTION

Centre for Development of Advanced computing (C-DAC) has been working with the Government. Of Maharashtra to digitize old published books of various departments of Govt. of Maharashtra and to bring out ePub, .Mobi and .PDF formats of these books (termed in general parlance as eBooks) [9]. There are more than 500 Books which are being digitized and brought out in eBook format. This paper tries to establish a typology of issues and challenges which were encountered in the process of digitization. The classification is illustrated through numerous examples under different category of problems. Some of the problems are associated with the technology and others are associated with the ignorance of data entry and thought process. Starting off with a general overview of the process of conversion of a non-digital document to a digital format, the paper presents next a typology of the problems encountered and pitfalls to be avoided.

2. BACKGROUND AND PREVIOUS WORK

The One of the biggest advantages of digitization of any media is of immediately accessible, portable, and easy to store and organize into libraries. Many music and film industries transformed by the fast developing

¹ Senior Technical officer, GIST, Pune, Email: chandrakantd@cdac.in

² Associate Director & HoD-GIST, C-DAC), Pune, Email: mdk@cdac.in

³ Professor Pimpri Chinchwad College of Engineering Pune, Email: nilesh.uke@gmail.com

⁴ Project Engineer-I, GIST, C-DAC, Pune, Email: rahulb@cdac.in

⁵ Senior Technical Assistant, C-DAC), Pune, Email: dipu@cdac.in

information and communication technologies in recent past. In music industry audio cassettes are replaced by MP3 files. A similar situation can be observed in the film industry where content distribution by DVDs and Blue ray Discs has been declined due to downloads through Peer-to-Peer (P2P) networks sites.

There is no specific definition for e-book, but many researchers have defined e-book in many different perspectives. [1] Simply referred e-book as text that is available in the electronic format. In general, e-book has been used to refer to hardware, software or document content [2]. Chen defines e-book in terms of four perspectives; the media used to preserve the books; the content; the device used to read the content; and the delivery channel [3].

Creating digitized books has a long history and is not new concept. Project Gutenberg, which was initiated by Michael Hart in 1971 at the University of Illinois, is considered one of the first steps in this direction. In 1990s, Many companies began selling books on CDs, but they had limited success as CD-ROM drives were not very common in that time [4].

In 2002, Stephen King launched a novella called *Riding a Bullet* in electronic book format which registered hundreds of thousands of downloads via Amazon and Barnes & Noble. This experiment popularized the idea of e-book and introduced the prospects that e-book might be commercially viable for book publishing [5].

3. DIGITIZATION PROCESS

First step towards the process is to segregate the available books into various categories of complexities. Class-I of the books are those books which contains mostly text and are small in size in terms of number of pages (typically less than 150). Class-II of the books contain books which has images, tables and references in them and are relatively large in size (between 151-500), and Class-III of the books has more complexities including complex, multicolumn layout, multiple language text, cross references within different pages of the book, different sizes of the pages, has more than 500 number of pages.

The flow chart illustrates the process followed for digitizing the books as shown in Fig 1. Once we classify the books, then we input the text with the help of data typist. Data typist is trained to type the text using a tool/ input mechanism which supports Unicode. This training is needed as it becomes difficult to find data typist who already knows typing text in Unicode format. Traditionally, data typist have been using nonstandard way of typing and storing the text. Since we wanted to take full advantage of eBook features, we decided to use standard way of inputting the same. There are many problems in this area of the process flow. Use of optical character recognition (OCR) systems for automatically converting scanned pages into text has lot of limitations for Indian languages. Hence this method is not considered for this effort.

Next step is to validate the typed text manually with the help of validators and proof readers. This is also a manual process since not many good spell checkers and grammar checkers are available which can handle this automatically. Moreover, there is a need to check the text with the already published and printed books. So it becomes more than just publishing the books from scratch.

We then format the text of the books according to the printed book. We make use of open source office softwares for formatting the books. We have tried few proprietary publication software but not used for our efforts. In this phase, various things such as handling tables, images, drop caps, mathematical and chemical formulas, references, etc is taken care of.



Figure 1: Digitization process

We use the formatted text documents to convert into various formats like PDF, epub and Mobi. When converting into epub and Mobi, there are multiple issues in terms of well formation of HTML/ XML files. At times, we had to clean up the markups to pass through the requirements. Finally, we have used ePub validators to check for validating ePub files generated. This gives confidence to the users that they will be able to read these files properly.

4. ISSUES AND CHALLENGES IN DIGITIZATION

4.1. Problem Typology

The problems encountered are classified into various categories.

1. Inputting, display of text and Authoring tools,
2. Validations and proof-reading,
3. Formatting of the text,
4. Miscellaneous

Each of these will be treated separately.

4.1.1. *Inputting, display of text and Authoring tools*

There are many issues which can be classified under this class, but the major ones are associated with text entry, text display and authoring tools used to create eBooks. They are further classified into technical (related to available technologies) and non-technical (associated with human resources, ignorance, general awareness, etc) issues.

a) Lack of standardized, fit to all text font

While beginning inputting the text in the book using document processor such as Libre Office, data typist selects one of the standard Unicode compliant fonts since in the beginning the typist does not know the variety of characters in the text of the book. Many a times, the text demands use of multiple fonts. The text in the book needs multiple fonts in order to render the text properly. However, the reading devices may not support the embedded fonts in finally created ePub 3 format. For example, in case of Zarathustra font, we can use it for text entry, but since the characters do not have Unicode value assigned to them, Unicode [10] values of Latin characters are used there instead. In such cases, searching text in the eBook becomes difficult and meaningless. Also some readers such as iPad do not have capability to use embedded fonts to display the text. Hence, one font fit to all is impossible and hence this is a big issue for digitizing text books.

b) Similarly confusing characters

There are multiple characters which look similar in printed text, but have different meanings. Hence, unless the data typist knows the context of the words in text, they often get confused.

c) Use of ZWJ and ZWNJ

Then there are issues associated with use of ZWJ (Zero Width Joiner) and ZWNJ (Zero Width Non- Joiner). These two characters are used when data typist has to indicate the processor whether two characters are to be explicitly joined or not to be joined. For example, if a font contains a rule containing combination of “द” and “व”, common result is formation of “द्व”. But in order to type words such as भगवदवाक्य, we need to use ZWNJ. The necessity of writing these words in this way comes from its appearance in the published text. In the example given below, the two words (figure 2 and 3) mean the same, but printed in two different ways.

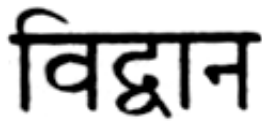


Figure 2:

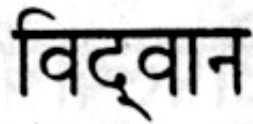


Figure 3:

Moreover a desultory use of these two character is carried out: since users also tend to type them where it is absolutely not required. For example, in cases such as (दृष्टि = U+0926, U+200C, U+0943 and & दृष्टि =U+0926, U+0943, उदगम = U+0909, U+0926, U+200C, U+094D, U+092E and उदम = U+0909, U+0926, U+094D, U+092E).

d) Horizontal and vertical character stacking

There is a problem associated with horizontal and vertical character stacking. For example, in some of the books both types of sackings are used for word display. We cite examples of two such words using “क”, “क”, and “क” in the same book. Though this facility of alternate glyphs is recently made available, it is available in few proprietary tools/ softwares. So purchasing the software packages and training the manpower to switch over to new software becomes a difficulty.

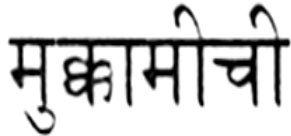


Figure 4:

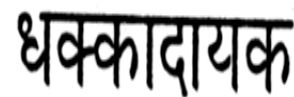


Figure 5:

e) Multiple ways to type same word

Data typists often tend to use different ways of inputting same text in Devanagari. This does not really change the way the appearance of the text and hence unaware of the correct sequence, they tend to use it wrong way. For example, to form words such as “जॉर्ज”, correct sequence of the word is – ज, ळ, र, ज”, but people also tend to use “ज, ळ, र, ज” to get the same word. Similarly for word “ऑण्ड”, correct sequence is “अ, ण, ळ, ड” and incorrect way is “अ, ळ, ण, ळ, ड”. For word “ऑड”, correct way is “अ, ळ, ड”, but people also tend to use “अ, ळ, ड”. This results in incorrect Unicode storage as well as it impacts the output as the speed of text entry is decreased to that extent. Another interesting case is that of the use of the candra in Marathi. in word ! M 0 K . E @ can be written in different ways. Those ways are a) “अ, ळ, ः, b) “अ, ः, and c) “अ”, “~”. It becomes very difficult to understand the difference between these three ways. Out of these three ways, only third way is interpreted by the eBook reader properly and rendered. The other two ways do not represent correctly in the ePub format.

f) Orthography of the text

There is an issue with the orthography of the text. There are some old books which use old orthography, but is currently obsolete. In cases where these types of books are to be referred, it is difficult to generate such words. We either need to change the glyphs of the font or tweak the inputting mechanism. For example, below are the words which are present in book related to Marathi language. The author of the book wants to show these words as they were used once upon a time, which are not acceptable today.



Figure 6: Old orthography

g) Equation or formula between running text

It is very difficult to type an equation or a formula when it comes in between running text. For example, in below example, there are equations in the running text. Hence, it is difficult to input as well as format such text. Text marked in red color is an issue.

अपूर्णांक : ख ने क ला निःशेष भाग जात नसेल तर $\frac{क}{ख}$ या संख्येस 'व्यवहारी अपूर्णांक' म्हणतात. क आणि ख ला अनुक्रमे 'अंश' आणि 'छेद' म्हणतात. जर क < ख असेल तर $\frac{क}{ख}$ याला 'युक्त अपूर्णांक' म्हणतात.

व्यवहारी अपूर्णांकांचे गुणधर्म : (१) $\frac{क}{ख} = \frac{प क}{प ख}$ हे समीकरण कोणत्याही प साठी खरे असते. यावरून एकच व्यवहारी अपूर्णांक विविध रीतींनी दर्शविता येतो असे दिसून येईल. उदा., $\frac{२}{३} = \frac{७ \times २}{७ \times ३} = \frac{१४}{२१}$

(२) $\frac{क}{ख}$ आणि $\frac{प}{फ}$ यांची बेरीज पुढीलप्रमाणे करतात : ख आणि फ यांचा ल. सा. वि. ल काढतात. समजा, ल = रख = गफ आहे, तर $\frac{क}{ख} = \frac{रक}{रख}$ आणि $\frac{प}{फ} = \frac{गप}{गफ}$ असे लिहून $\frac{क}{ख} + \frac{प}{फ} = \frac{रक}{रख} + \frac{गप}{गफ}$
 $= \frac{रक + गप}{ल} = \frac{रक + गप}{ल}$

Figure 7: Equation and formula in running text

h) Use of Vulgar fractions in Marathi Text

Consider the vulgar fractions $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, etc. These characters have Unicode values U+00BC, U+00BD and U+00BE respectively. If the same fractions are to be written in Marathi text, Devanagari unicode block does not have separate Unicode values for those characters. ($\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$). There are separate Unicode values for these fractions in some of the Indian scripts like Malayalam. Hence it is not possible to have both types of fractions in the same font.

i) Inability to type some characters

There are many characters which cannot be typed using current inputting mechanisms commonly available. In fact there is no Unicode representation for these characters. Few of the examples are given below. These characters are used to represent the currency units, demarcation and other special usage used in Devanagari text in old days.

j) Half characters

There are many instances where Half characters used in Devanagari needs to be typed. Currently there is no direct way of typing Half characters except for use of some other characters to form them.

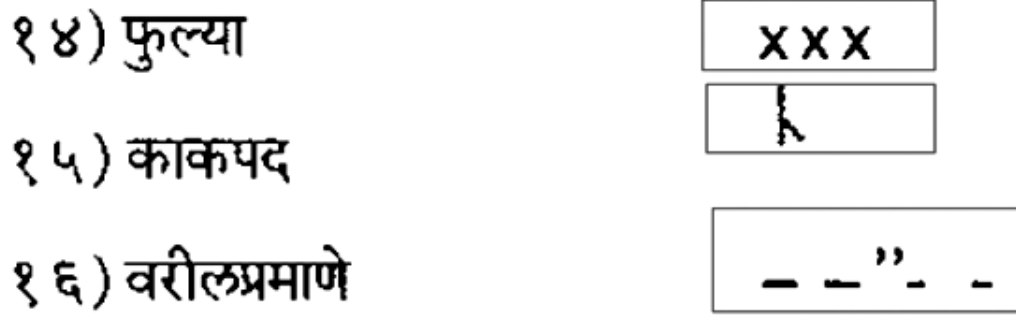


Figure 8: Different characters which are difficult to type

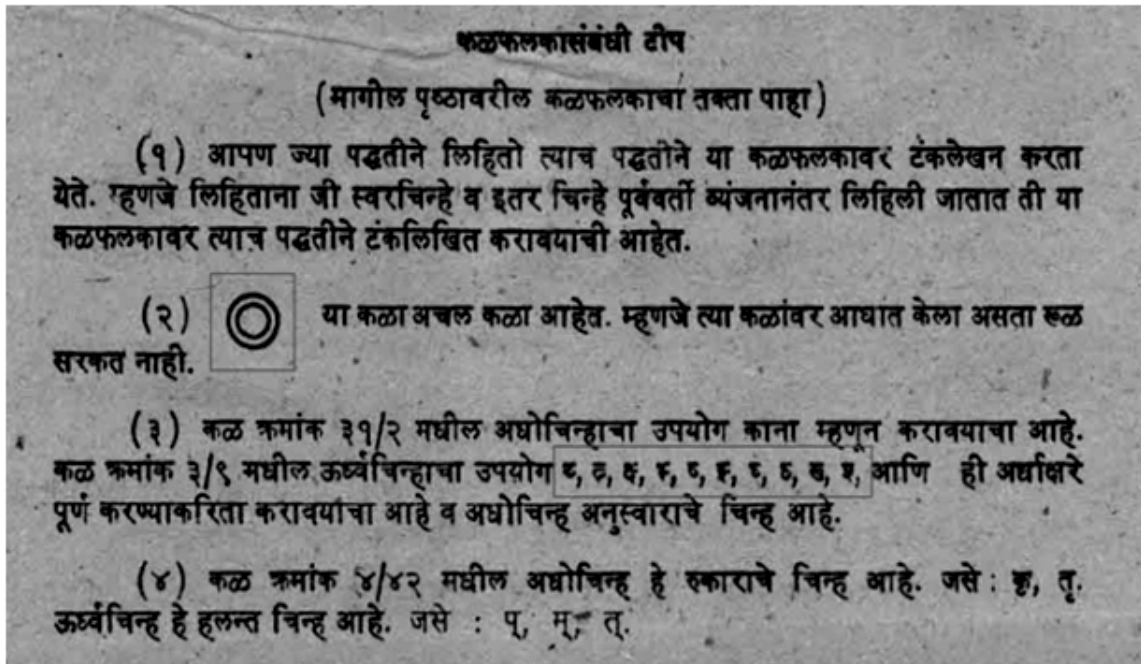


Figure 9: Half characters

k) Use of Superscript and Subscript

Superscripts and subscripts (Marathi and English) are difficult to handle. This is because there is no guaranteed converter for converting these superscript and subscripts into ePub (HTML) file tags. An example in case is when we have to represent degree symbol. Often data typist tend to make a superscript to the numbers while text inputting. When you convert into ePub, these superscript and subscripts are lost. Hence we need to make sure that in such cases, we have to explicitly type the degree symbol by copying from character map (charmap) or other tools.

l) Incorrect placement of HTML Tags

Incorrect placement of HTML Tags also creates problems in ePub file. For example instead of पुनः one places a tag like incorrectly after पुन (), This will create a problem in rendering of this word in ePub. But this is a difficult to track.

4.2. Validations and proof-reading

4.2.1. Capturing of Metadata and other Important Information

There are many published books where, on each page, the title of the book and page numbers are written. Also, some of the books has got chapter names written on the page to which the page belongs. This information is useful for a reader. While digitizing, these page numbers play a crucial role when they are

ती सहजसुलभ अशी असते कारण विश्वाशी एकरूप होण्यात अधिष्ठान व्यक्तीचे असते (पृ. १५०).

म्हणून आचार्यांची 'जीवनसाफल्याची' कल्पना (प्रकरण २) ही पारंपारिक कल्पनेपेक्षा वेगळी इहवादी अशीच आहे. कोणत्याही गूढ रहस्याच्या, अथवा कला किंवा आत्म्याच्या स्वातंत्र्याच्या मागे जाऊन मोक्ष साध्य होणार नाही असे रोमा रोलां यांच्या अवतरणावरून ते दर्शवितात (पृ. १५). जीवनाचे साफल्य जीवनातच शोधले पाहिजे. जीवनाचे मांगल्य व त्यावरील श्रद्धा वाढत गेली पाहिजे असा त्यांचा आग्रह आहे. शेवटच्या परिच्छेदावरून निबंधाची तात्कालिक प्रेरणा दिसते. त्या काळी 'कलेकरता कला', नवनीतीवाद, कॉम्युनिझम या कल्पना नुकत्याच आलेल्या होत्या व त्या सर्वांवरच आचार्यांनी आपल्या जीवनवादी भूमिकेवरून उपहासात्मक टीका करून शिवरामपंत पराजप्यांच्या ल्यबद्ध, क्वचित् प्रासयुक्त शैलीची आठवण करून दिली आहे.

चातुर्वर्ण्यव्यवस्थेचा येथील उल्लेख मात्र गौरवर दिसतो. आश्रमव्यवस्था, धर्म, मोक्ष इत्यादी कल्पनांबरोबर ती वापरली जात असता ह्या सर्व कल्पनांनी जीवनाला समृद्ध केले ह्यात शंका नाही असे त्यांना वाटते (पृ. १५). मात्र पुढच्या काळात तीत वैगुण्य आले म्हणून त्यांनी चातुर्वर्ण्याच्या निषेधात्मक भूमिका संतांच्या आध्यात्मिक व सामाजिक कार्यांच्या

Figure 10: Page numbers referred within the page

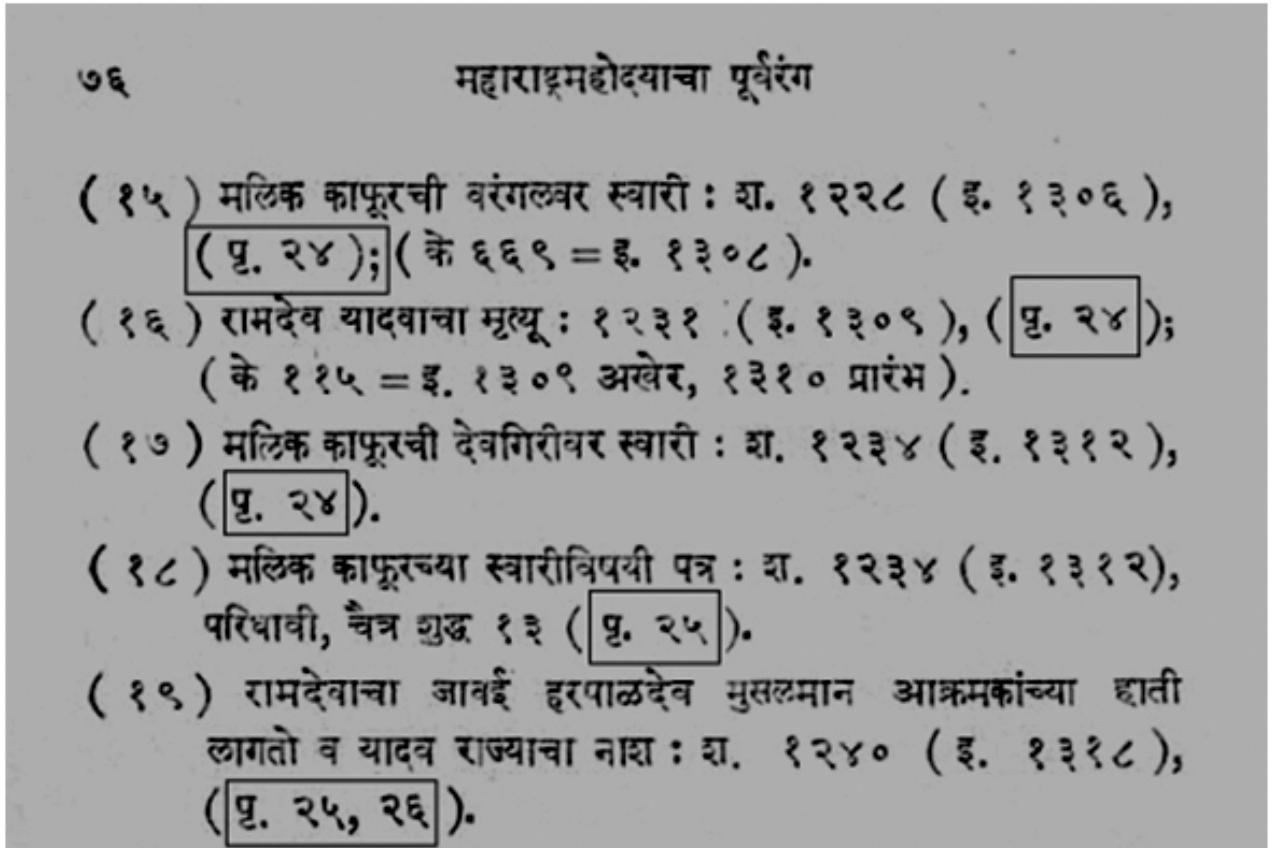


Figure 11: Page numbers referred within the page

referred in the text matter. In the below examples (Figure 10 and 11), text in red refers to the page numbers of the same book. Since, ePub does not have a concept of page numbers capturing this information is difficult and hence retaining all of its meaning in ePub is challenging.

In the examples below (Figure 12 and 13), the red text represents, superscript representing the references mentioned on that particular page only, the actual references and the chapter name and name of the book. It is important to note that there two examples are subsequent pages of the same book. Hence capturing this type of information is very difficult.

मूळच्या जागपासून हलवले गेले आहेत. टोप्रा व मीरत येथील स्तंभ फिरोजशहा तुघलक^१ याच्या कारकिर्दीत दिल्लीला आणण्यात आले. त्याबरील लिपी ज्ञात नसल्याने त्या चमत्कृतिजनक वस्तू म्हणून मानल्या जात असत. अलाहाबादचा स्तंभ मुळात कौशांबीला असावा असे मानले जाते^२ बैराटचा कोरीव लेख कनिगहॅमने कलकत्याला हलवला^३ यामुळे वस्तूच्या परिसराच्या संदर्भात अभ्यास करण्यास प्रतिबंध होतो व पुरातत्वशास्त्राच्या दृष्टीने ही खेदजनक बाब आहे.

हपूएन्संग या चिनी बौद्ध यात्रेकरूने, इसवी सनाच्या सातव्या शतकात भारतात प्रवास करीत असताना, राजाह, श्रावस्ती व इतर ठिकाणी पाहिलेल्या

१. इलियट-डाऊसन, हिस्ट्री ऑफ इंडिया, व्हॉल्युम तीन, पृ. ३५०, तळटीप.
२. सीआय्‌आय्, व्हॉ. एक, पृ. एकोणीस.
३. उपरोक्त पृ. पंचवीस.

८ : अशोक आणि मौर्यांचा न्हास

Figure 12: References used, number of the chapter and name of the book

वरून दिसून येते. कंदहारच्या कोरीव लेखाचा अपवाद सोडल्यास प्रत्येक आज्ञा-लेखातील भाषा ही अशोककालीन प्राकृत आहे. स्थूलमानाने तिचे पुर्वेकडची व पश्चिमेकडची प्राकृत भाषा असे प्रादेशिक भेद आहेत. संस्कृतसारख्या सस्कृतीच्या भाषेऐवजी प्राकृतसारख्या लोकभाषेचा वापर अशोकाने सातत्याने करावा ही गोष्टही या ठिकाणी लक्षात घेण्याजोगी आहे.

अशोकाचे नसलेले, परंतु मौर्यकालाशी प्रत्यक्ष संबंध असलेले असे इतर कोरीव लेख आहेत, त्यांच्यापैकी तक्षशिला येथील 'प्रियदर्शि' कोरीव लेखाचा

१. बि सार्फ ऑफ हपूएन त्संग वाय शमन हुइलि, बुक तीन (भाषां-वील पृ. ९३)
२. गाइल्स, ट्रेन्व्हल्स ऑफ फा-हियान् पृ. २५, ४८.

९ : पार्श्वभूमी आणि आधारसाहित्य

२५-२५
२

Figure 13 : References and number of the chapter and name of the book

4.2.2. Cross References

This is a class of problems associated with referencing some other text, in the same book or any other book. In the example given below, the superscripts (used for tooltip) are referring to a tool tip given at the end of the page. The tool-tip text at the end of the page itself is referring to some other text on previous page of the book. Capturing this type of information is also crucial, but currently, there is no concrete solution to this problem.

आत्मदेशास ।
देई सौख्याला ।
जो पुनः स्थापि धर्मा^९ ।

नानांच्या संगीताच्या अभिरुचीसंबंधाने त्यांचे स्वतःचे उद्गार देऊन हा विषय संपवितो. “ सुप्रसिद्ध बाळकृष्णबुवा इचलकरंजीकर हे ‘ख्याली,’ आपले चरित्रनायक विष्णुपंत छत्रे हे ‘भृपदे,’ नारायणबुवा गोगटे, फलटणकर, हे ‘टप्पे,’ आणि सत्वाराम्बुवा हे ‘लावणीवाले’ ह्या चौदांही गवय्यांचे गाणे ऐकण्याइतके भाग्य ह्या लेखकास ईशकृपेने प्राप्त झाले आणि नारायणबुवा फलटणकरांचे गाणे तर, त्याने अगदी लहानपणापासून – त्याला कळावयास लागल्यापासून – तो ते दुर्दैवाने असत्या आयुष्यांतच कैलासवासी होईपर्यंत मिळून, शंकडां वेळां ऐकिले आहे. स्वतः या लेखकाला गायनकला मुळांच अवगत नाही. केवळ थोडीफार त्या कलेविषयी नादलुब्धता मात्र आहे, इतकेच.”^{१०} ”

जुने, प्राचीन मराठी काव्य साक्षेपाने व त्रिकित्सेने अभ्यासिणारे बहुश्रुत विद्वान् म्हणून नानाना मिळालेल्या लौकिकाचा एक भाग म्हणून १८९९ साली मुंबईस मोरोपंतांची सार्वजनिक पुण्यतिथी (चैत्र शु. १५) साजरी करणाऱ्या कार्यकर्त्यांत नाना एक प्रमुख कार्यकर्ते होते, या गोष्टीचा उल्लेख करा-

९. मागील पृष्ठात दिलेली ८ ही टीप पाहा.

१०. पं. विष्णुपंत छत्रे यांचे चरित्र, पृ. ६७-६८

Figure 14: cross references

Given below is the page from the translated version of some other book. In this case, the page numbers (red marked) refer to the page numbers of original book from which current book is translated. The page number of this page is 20 (translated version), while red marked text represents page number 23 of the original book.

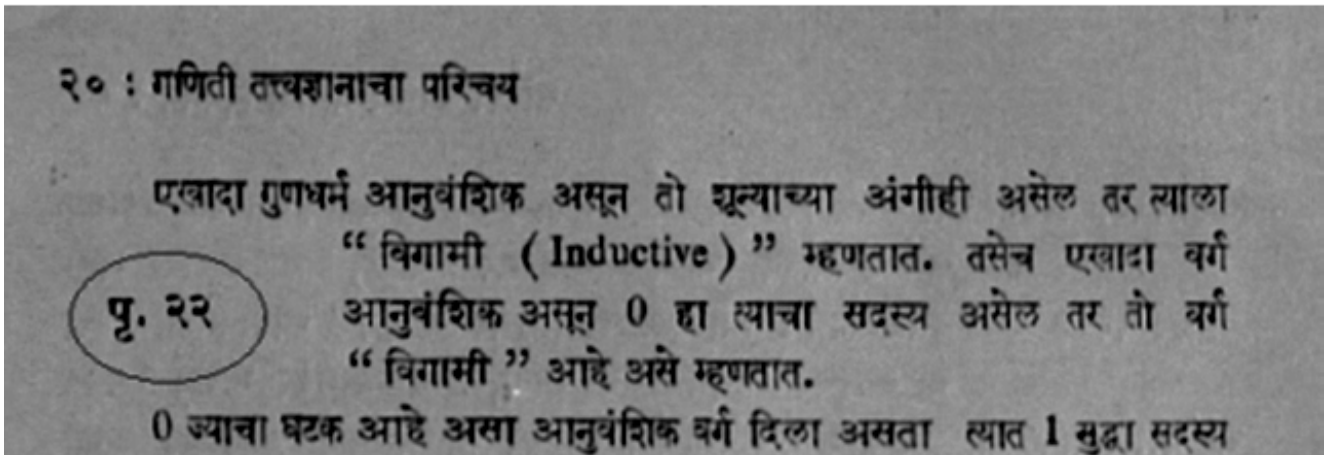


Figure 15: references in the text block

4.2.3. Manual Scribbling

Many of the books contain stamps of the owner library and some other handwritten text, for example, register number. This is important for knowing the source/ owner of this book. Such metadata is difficult to capture apart from incorporation of an image.

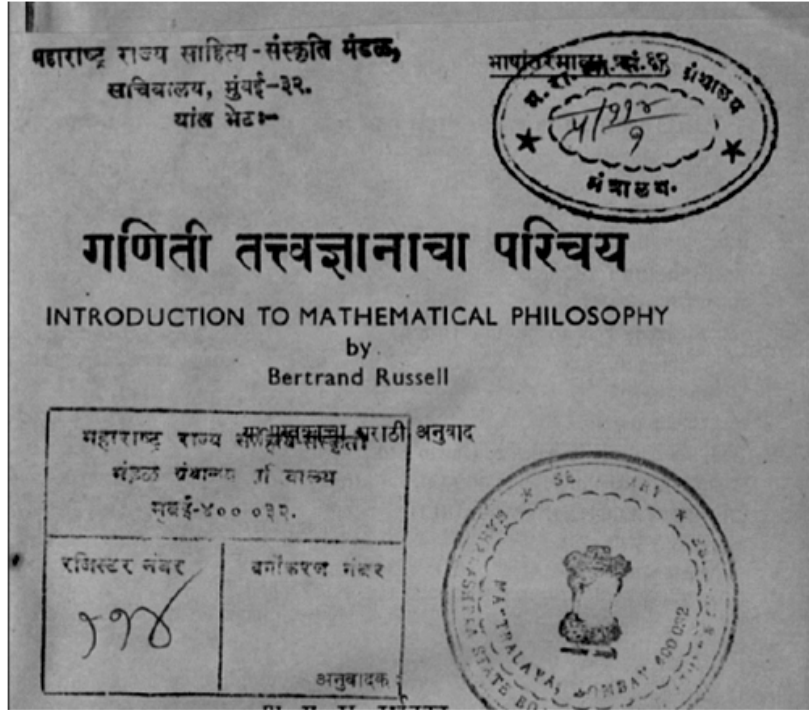


Figure 16: Manual scribbling

4.2.4. Abstract of the text

In the example below, the red marked text is an abstract of what is being described in the adjacent paragraph. This gives reader an opportunity to just read abstract instead of complete adjacent text. When converting into eBook, it is difficult to capture this context in ePub format.

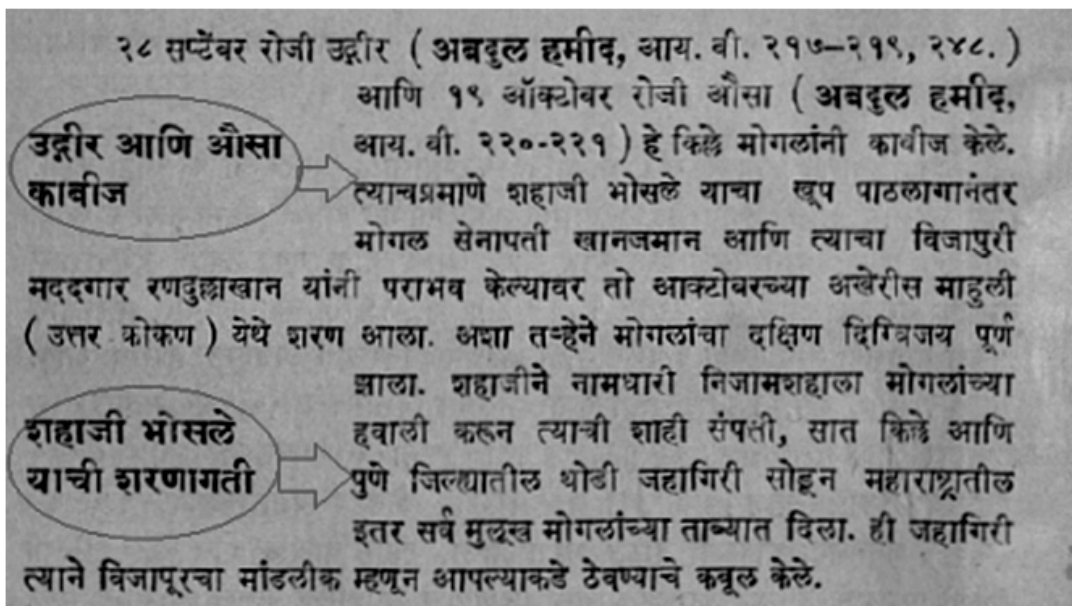


Figure 17: abstract of the paragraph

4.3. Formatting of the text

Complex layout (Multicolumn) with lots of Images, Tables, Diagrams are difficult to handle as they are not properly displayed on the devices (eReaders). Multicolumn and Single Column layout

यंत्रालयाचा ज्ञानकोश 421 प्रशाणन यंत्र

```

graph TD
    A[विशेष प्रमापी साधने] --> B[कानगिरीनुसार]
    A --> C[व्यवर्था श्रेणीनुसार]
    A --> D[मान्यशास्त्रीय गुणधर्मानुसार]
    B --> B1[रेषामापन व कोनमापन]
    B --> B2[भूमितीय आकार मापन]
    B --> B3[पृष्ठ पक्यता मापन]
    C --> C1[व्यवर्था श्रेणी - 1]
    C --> C2[व्यवर्था श्रेणी - 2]
    C --> C3[व्यवर्था श्रेणी - 3]
    D --> D1[प्राप्त मापन मूल्य]
    D --> D2[परिमाण प्रभाग मूल्य]
    D --> D3[परिमाण अंतर]
    D --> D4[संवेदनशीलता]
    D --> D5[वाचन काटेकोरपणा]
    
```

1. प्रत्यक्ष प्रमापी साधने : ज्या प्रमापी साधनांवर त्यांच्या साहाय्याने घेतलेल्या मापाचे प्रत्यक्ष परिमाण दाखविणाऱ्या खुणा वा/व अंक लिहिलेले, छापलेले अथवा कोरलेले असतात त्यांना प्रत्यक्ष प्रमापी साधने म्हणतात. प्रत्यक्ष प्रमापी साधनाच्या साहाय्याने विशिष्ट घटकाचे माप प्रत्यक्षात 'वाचता येत असल्याने' प्रत्यक्ष प्रमापी साधने 'वाचिक प्रमापी साधने' म्हणून ओळखली जातात. विशिष्ट असे

नसतात त्यांना अप्रत्यक्ष प्रमापी साधने असे म्हणतात. विशिष्ट अशा एखाद्या घटकाचे माप अप्रत्यक्ष प्रमापी साधने घेतल्यावर 'ते वाचण्यासाठी' ठरावीक अशा एखाद्या वाचिक प्रमापी साधनाचा उपयोग करावा लागतो, तरच 'घेतलेले माप वाचता येते'. अशा प्रकारे, घटकाच्या परिमाणाची वाचिक प्रमापी साधनावरील मापाशी तुलना करण्याचे तुलनात्मक साधन म्हणून अप्रत्यक्ष प्रमापी साधने, तीलनिक प्रमापी साधने म्हणून ओळखली जातात. - उदाहरणार्थ, कैडार (caliper).

(टीप : निरनिराळ्या वाचिक प्रमापी व तीलनिक प्रमापी साधनांची माहिती इतरत्र अकारविल्ले दिली आहे, ती पाहावी.)

काट : 1. मापन व मापन पद्धती, 2. मान्यशास्त्र

प्रशाणन यंत्र (Honing Machine) : प्रशाणन यंत्र हे एक विशेष प्रकारचे यंत्रोपकरण असून ह्याचा उपयोग विशिष्ट घटकांच्या भोकांचे अत्यंत काटेकोर यंत्रण करण्यासाठी करतात. सामान्यतः प्रशाणनाच्या साहाय्याने विशिष्ट मापाच्या भोकाच्या निर्मितीतील पुढील प्रकारच्या दोषांचे निराकरण केले जाते :

1. भोकाचा आकार इच्छित प्रमाणात वृत्ताकार नसणे;
2. भोकाचे माप किंचित कमी असणे;
3. भोक शुंडाकार तयार होणे;
4. भोकाच्या पृष्ठभागावर तरंग उठणे;

Figure 18: complex layout of the page

4.3.1. Capitalization (Drop Cap) and graphics for decorating the text

Drop caps and its decoration is a challenging thing in normal word processors. Use of specialized authoring tool also does not guarantee other features.

भाषाविकारी गुन्हे असतात, तसेच तर्कशास्त्रीय असत
विभागणारी रेषा स्पष्ट नसते. पण, 'पूर्वचितित' म
संकल्पनेने दंडसंहिता त्यांच्यातील भेद प्रदर्शित करते. आपण पूर्वचितन
ज्याचा मागमूस मागे राहात नाही अशा पटाईत गुन्हेबांच्या यु
आहोत. आमचे आजचे गन्हेगार गन्हेबांचे समर्थन म्हणून प्रेमभाव

Figure 19: Drop case

4.3.2. Complex Tables and other hierarchical structures.

Often the printed pages contain text in vertical way rather than horizontal way in order to accommodate big and complex tables. Representing such tables in eBook format is an issue when it comes to proper display on the reading devices. User has to swipe the area of this table to read the text in the table and hence difficult to read the data at once.

TABLE 15
POPULATION BY AGE AND MARITAL STATUS

Age Group	Marital Status 1961											Total Population	
	Never Married		Married		Widowed		Divorced or Separated		Unspecified Status				
	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Persons	Males	Females
All ages ..	4,06,327	3,05,644	2,67,226	3,21,855	22,715	99,987	2,083	3,832	204	232	14,30,105	6,98,555	7,31,550
0-14 ..	3,18,793	2,91,979	1,075	13,307	12	113	1	85	71	43	6,25,479	3,19,952	3,05,527
15-24 ..	76,656	11,618	25,586	1,02,924	184	1,854	202	1,232	34	55	2,20,345	1,02,662	1,17,683
25-44 ..	8,977	1,868	1,38,630	1,57,618	3,041	22,766	1,153	2,039	51	85	3,36,228	1,51,852	1,84,376
45-54 ..	976	83	49,976	33,343	22,639	23,691	12,391	332	15	17	1,13,463	55,997	57,466
55-69 ..	717	63	40,589	13,184	9,103	35,052	269	124	22	24	99,147	50,700	48,447
70+ ..	186	25	11,348	1,465	5,736	16,509	67	20	6	6	35,368	17,343	18,025
Age not stated	22	8	22	14	2	5	2	75	49	26
1971													
All ages ..	5,03,139	3,82,952	3,25,486	3,95,088	18,367	98,844	1,050	2,330	50	70	17,27,376	8,48,092	8,79,284
0-14 ..	3,82,064	3,56,211	565	7,150	10	10	10	7,46,040	3,82,649	3,63,391
15-24 ..	1,09,376	25,444	26,206	1,07,439	100	1,336	35	620	15	20	2,70,591	1,35,732	1,34,859
25-44 ..	9,949	1,025	1,60,402	2,04,964	1,815	16,432	480	1,395	20	40	3,96,492	1,72,666	2,23,826
45-54 ..	720	130	64,256	49,948	3,229	22,058	240	210	5	1,40,796	68,450	72,346
55-69 ..	705	80	58,053	23,086	7,511	38,634	235	95	1,28,399	66,504	61,895
70+ ..	270	15	15,986	2,531	5,703	20,374	60	10	44,949	22,029	22,920
Age not stated	35	47	18	9	109	62	47

Figure 20: Complex tables

4.3.3. Creating Graphics

Grabbing Images from scanned images and Manual method of creating a fresh. There are so many incomplete images (specially the book covers of old books) which needs to be recreated. While creating Book Cover from scratch is relatively easy to undertake, patching up the half images to create full image is time consuming and difficult task. We undertake to recreate the covers from whatever portion is available.

4.3.4. Navigation within PDF files

Consider the situation that user is at page n of the book. Enabling him/her to go from that page of the book to Table Of Contents (TOC) and Vice versa in a single click would be really helpful. This is a unique thought which we have incorporated in all the digitized books (PDF) of our work. Currently user can navigate from TOC to the page where the text appears and vice versa using Bookmarks on the left side of PDF window. When one has to go back to TOC, he/she has to use Navigation panel. It would be better if you have a mechanism to go to TOC by clicking a link at the end of the page. This is easier for navigation. Making this facility was difficult for us.

4.3.5. Size of eBook

Imagine there are many images in the book. If there is no source of those images, it is but obvious to scan them from the available book and reuse them in the eBook. In some of the books, there are so many images



Figure 21: Sample page having Images and text

that you cannot avoid integrating in the book. For example, in a book regarding Short Hand, it has got Marathi languages sentence in one column and equivalent short hand on the other. When this has to be digitized, all the short hand matter has to be captured as Images. Given below is the example. Hence this becomes an issue in terms of size of the eBook. Currently we do not have any other way to solve this issue.

4.3.6. Mathematical Formulas and chemical equations

Currently, we explored three different approaches to deal with the representing mathematical formulas and chemical equations. First one is straightforward and involves scanning of the relevant mathematical formula from the book being digitized. An example below explains this.

$$\begin{aligned}
 y_{k0} + ay &= y_0 + \frac{a(\Delta y_0)}{1!} + \frac{a^2(\Delta^2 y_0)}{2!} + \frac{a^3(\Delta^3 y_0)}{3!} + \dots
 \end{aligned}$$

Figure 22: Complex Mathematical equation in Marathi

Advantage of this approach is that it is easy to use and relatively less time consuming. However, since the images are not scalable, it becomes problematic in reading on eBook readers/ devices. Also, the size of the eBook increases based on the number of images in the eBook.

The second approach is applicable, if the mathematical formula is simple, we can try to represent the formula using the table rows and columns. An example of this is shown below.

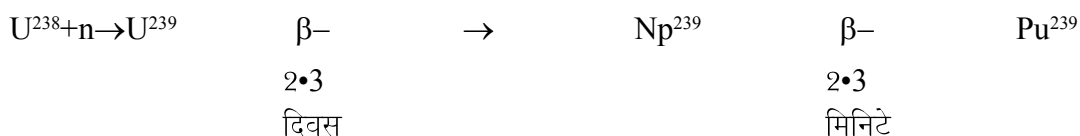
$$\eta = 1 - \frac{1}{\gamma} \left\{ \frac{\left(\frac{\text{सं}}{\text{प्र}}\right)^{\gamma-1}}{\text{सं}^{\gamma-1} \times \left(\frac{\text{स}}{\text{प्र}} - 1\right)} \right\} \quad \text{या सूत्राने}$$

Figure 23: Relatively simple mathematical formula in Marathi

This table can be converted to displayable equation

$$n = 1 - \frac{1}{\gamma} \left\{ \frac{\left(\frac{\text{सं}}{\text{प्र}}\right)^{\gamma-1}}{\text{सं}^{\gamma-1} \times \left(\frac{\text{स}}{\text{प्र}} - 1\right)} \right\}$$

Similarly, for chemical equations, making use of tables is one option.



This method is time consuming even for simpler formula and equation. Third method is to use MathML. There are challenges in creating the formula's using MathML, and even when the formula's are generated using Equations Editor in Libre Office for example, finally it has to be exported as Image into ePub file. More important challenge was to find people who would be able to generate formula using MathML. Hence, we used combination of first two methods for dealing with formula and chemical equation.

4.3.7. Non-Technical Issues

Our study shows that the books to be digitized are domain specific and they need resource with expertise and domain knowledge. It is difficult to get either the original authors or resources with such expertise, given the budget and time constraints. We also found out that there are very few people who understand how to type in Unicode. This may be because in India, the software available has mostly been supporting non-standard mechanisms. So trying to locate or bring on board such people is a difficult task.

4.4. Miscellaneous

Last but not least, because of unavailability of good and useful word processing tools such as spell checker, grammar checker, etc, typed text cannot be checked for spelling mistakes. We have also noticed that versions of the software's used for Authoring such as MS Word, Libre Office, etc, can create problems while importing files from one version to another.

5. CONCLUSION

As we have seen from various issues listed above, it is challenging work to digitize the already published books into eBook format. All of these issues are very critical and work needs to be done to solve these issues. There is plenty of scope of research in this area of digital publishing and digitization. We would be happy to work in collaboration with individuals/ organizations interested in bringing out the standards to solve these issues.

ACKNOWLEDGEMENTS

We would like to acknowledge Govt. of Maharashtra and its departments (Marathi Vishwakosh Nirmiti Mandal, Bhasha Sanchalanalaya, Rajya Marathi Vikas Sanstha, Sahitya ani Sanskriti Mandal), Principal Secretary (IT), Government of Maharashtra for showing faith in C-DAC for undertaking this mammoth work. We would like to thank Dr. Raymond Doctor for his immense help while researching this work.

REFERENCES

- [1] S. Saurie, M.Kaushik, "Electronic publishing" *Electronic publishing IT*, *encyclopedia.com*. New Delhi, Pentagon Press, 2001.
- [2] M. Letchumanan and R. A. Tarmizi, "Utilization of e-book among university mathematics students," *Procedia - Soc. Behav. Sci.*, vol. 8, no. 5, pp. 580–587, 2010.
- [3] Y. Chen, "Application and development of electronic books in an e-Gutenberg age," *Online Inf. Rev.*, vol. 27, no. 1, pp. 8–16, 2003.
- [4] A. Maxim and A. Maxim, "The Role of e-books in Reshaping the Publishing Industry," *Procedia - Soc. Behav. Sci.*, vol. 62, pp. 1046–1050, 2012.
- [5] C. T. Kenneth C. Laudon, *E-commerce. Business, technology, society(4th ed)*. Upper Saddle River: Pearson Education International, 2008.
- [6] महाराष्ट्र महोदयाचा पूर्वसंग, लेखक- कै. नारायण कृष्ण गद्रे, संपादक प्रा गणेश, हरी खरे महाराष्ट्र राज्य साहित्य संस्कृती मंडल, 1971
- [7] अशोक आणि मौर्यांचा रहास (Translation of Dr. Romila Thapar authored, As'ok and The Decline of Mourya's), अनुवादिका- डॉ. सौ. शरावती शिरगावकर
- [8] गणितीतत्वज्ञानाचापरिचय (Translation of Bertrand Russell authored "Introduction to Mathematical Philosophy"), अनुवादक- प्रा. म. रा. राईलकर.
- [9] <https://www.maharashtra.gov.in>
- [10] <http://www.unicode.org>
- [11] <http://marathibhasha.maharashtra.gov.in>