

Quantitative Methods in the Humanities
and Social Sciences

Sukanta Chaudhuri *Editor*

Bichitra: The Making of an Online Tagore Variorum

 Springer

Quantitative Methods in the Humanities and Social Sciences

Series Editors

Thomas DeFanti

Anthony Grafton

Thomas E. Levy

Lev Manovich

Alyn Rockwood

More information about this series at <http://www.springer.com/series/11748>

Sukanta Chaudhuri
Editor

Bichitra: The Making of an Online Tagore Variorum

Editor

Sukanta Chaudhuri
Department of English
Jadavpur University
Kolkata, India

ISSN 2199-0956

ISSN 2199-0964 (electronic)

Quantitative Methods in the Humanities and Social Sciences

ISBN 978-3-319-23677-3

ISBN 978-3-319-23678-0 (eBook)

DOI 10.1007/978-3-319-23678-0

Library of Congress Control Number: 2015959948

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Foreword

There are more native speakers of Bengali worldwide than of Russian, Japanese, German, French, or Italian. One Bengali writer has won the Nobel Prize for Literature. The archive of his writings is larger than Shakespeare's, Goethe's, Proust's, or Faulkner's. His name is Rabindranath Tagore, poet, novelist, essayist and travel writer, dramatist, painter, composer, educator, translator. Furthermore, he promoted rural development and the improvement of agriculture and crafts. His archive of manuscripts and printed works, amounting to over 140,000 pages, is the largest archive for a major writer to be (almost) entirely digitized and posted to the Internet—"almost" because 40 rare books out of 450 books and 300 out of 3200 journal items could not (yet) be obtained for reproduction. The virtual archive was accomplished in two years by a team of 30 plus researchers and computer programmers funded primarily by the Indian government, which found itself justly proud of its Nobel Laureate on the occasion of his 150th birthday in 2011.

How they did it and why you should care is the subject of this book, *Bichitra: the Making of a Tagore Website*, by the project director Sukanta Chaudhuri. Readers of Chaudhuri's book, *The Metaphysics of Text*, are familiar with his elegant and clear prose, his attention to detail, his self-effacing grace, and his incredible stamina. Most of the world needs this book because we don't know Tagore well enough, we don't know Bengali, and we don't know how to build or use virtual archives. The onus is on us but *Bichitra*, the book, makes it easy to find out.

The first step is to understand the importance and achievements of Tagore himself. He is a recognized world figure, but few will know that his works (he wrote in both Bengali and English) exist in multiple versions. Sometimes he turned a play into a novel or vice versa, or he incorporated poems into novels or other works. Sometimes his works were both collected and anthologized under his supervision, for which he made changes. Sometimes he wrote the same work (more or less) in both Bengali and English. But more often he was discovering new things to say with his already written works—he changed his mind or he found a better way to say what he originally thought. The richness of Tagore's archive for the study of the genesis of thought and of literary works is unsurpassed by that of any writer anywhere. That is why it is called *Bichitra*, the various, the curious, the bizarre.

Obviously a reader needs more than just this book to explain *Bichitra*, the website. One needs to be able to work one's way around in the archive. So, there are tools: a search engine and a concordance engine bring Tagore's words and subjects together.

A bibliography with links to (nearly) every form of each work aggregates the related materials. A collation program identifies the variants in the different forms of each work.

It is an archive not an edition. At one point Chaudhuri modestly calls it a “mere archive” to explain why the site does not explore the genetic process or explicate the significance of textual variants—except for a small range of examples to show the potentials. He rightly points out what a major project that would be in itself. The site enables genetic study; it does not do it for us. There is nothing “mere” about this archive. For the first time, persons interested in Tagore can read any one of dozens of versions of his works, can read rare—not otherwise easily available—works, can read works in the context of collections of Tagore’s works or as originally printed, and can read the images of original publications or the transcripts made of them in order to be computer searchable. And readers can read manuscripts of works (mostly) published, but also versions that have never before been published.

Suppose, however, you are not interested in Tagore, you can still learn much about the Bengali language and its particular difficulties for keyboards, printing presses, and software for searching and collating. Even questions about fonts receive careful attention. In the absence of adequate software environments for major literary virtual archives (even for Roman alphabet languages), the Bichitra project invented its own standards for imaging, for transcriptions, and for collations. Everyone with a large text project confronts the delight and disaster of OCR (Optical Character Recognition) which even at 98 % accuracy produces an average two errors per 100 characters (counting spaces) or 40–50 errors per page and OCR is of no use at all for manuscripts, which have to be transcribed manually. Bichitra represents major accomplishments of interest to digital humanists everywhere—if they can just overcome their lack of interest in Tagore or Bengali. Ignorance is a comfortably debilitating condition, bliss—sort of.

For me the major accomplishment of the Tagore archive is the *images* of (almost) every version of every work. Digital collections of *transcriptions* are not archives, regardless of what anyone may claim for them. A transcription is a copy, a reset copy. It is different from its source text in every character because it is a copy susceptible to error at every character; it is not the original, it is not the same. Of course, a digital image is a copy also, but it is at least visually accurate. No one says that a picture of a person is the person. None should say that a picture of a book is the book. But digitally, images are as close as technology can get to providing surrogates for the material originals. Bichitra’s crown jewels are its images. No institution has all the documents, but in this website they are collected, photographed, and mounted. That is great not only for Tagore studies but also for all aspiring digital archives. The process, the cameras, the lighting, the negotiations for permissions to photograph, and the alternatives for storing, archiving, and displaying images are all so complex that anyone wanting to create a sophisticated archive website will learn much from the Bichitra experience. But it is so much more. Images cannot be searched, analyzed, or collated. For these operations transcriptions are needed, not just for the manuscripts but for the 90,000 pages of printed books as well. Bichitra provides them.

Those last three words were so easy to write. Over 47,000 pages of manuscript made transcription anything but easy. The chapter on manuscript transcription is easily the longest and most interesting because it deals so openly and sensibly with an extremely complex problem. Most readers will soon get over their unfamiliarity with the language as they get deeper and deeper into considerations of what every manuscript transcriber has experienced. Transcription is detective work, interpretive work, philosophical work, and practical work. Before the end of the day, decisions have to be made about how to proceed. Tagore was a rapid writer and inexhaustible reviser. Some of his assistants learned to emulate his hand. Is it a nightmare or a fertile field? Chaudhuri seems to know that it is the former but he treats it as the latter.

Every project director and every technical officer and computer science partner on a digital archive project will benefit from reading Chaps. 6 through 9 in particular. Chapters 6–8 do not shy from technical detail but even technically challenged textual scholars should have no difficulty understanding them.

They recount first the task of organizing the file structures required to keep track of hundreds of thousands of individual files of transcriptions and images. The project team devised a new content management system because there was none to hand adequate for the job. The description of Tagore's tangled bibliography is merely a prelude to describing the organizational system that brought digital order to it. Next they tackle the job of providing indexing and search capabilities to the website. Third, they describe the construction and function of a collation program that will handle Bengali language and multiple versions. These three back-end systems and tools represent a formidable accomplishment; given the time in which it was done it is like a miracle.

Chapter 9 describes the front-end user interface design and functions. Given the intricate and orderly content management system, display of content for the user is potentially infinitely malleable. The achieved system is not perfect but it is more than a very good beginning. The project was launched at a significantly high plateau of achievement.

Chapter 10 treats the entire project as a good start—it is far better than that—and addresses three areas for improvement: additions to the content, improvements of the internal synchronization of images and transcriptions, and additional analytical tools and uses for the content. The project, thus, fulfills the expectations of modern modular project structures, rejecting the intricate monoliths of early electronic projects. It is extendible.

The book begins and ends with acknowledgements to those who constructed or supported the project. It is fitting that this description of so large a project, with such high standards, should begin and end so. It takes a village to build a digital archive.

Peter Shillingsburg

Preface

This book tells the story of the making of Bichitra, the online variorum of the works in Bengali and English of the Indian poet and writer Rabindranath Tagore. To the best of our knowledge, it is the world's largest integrated literary database. By 'integrated' I mean that it was planned and created in a single operation, its various parts meshing with one another and, to a very great extent, accessible from one another. This huge operation, covering nearly 140,000 pages of primary material, was completed in a little over two years, which too must be something of a record. I do not wish to sound overly self-congratulatory. As this book should indicate, we are well aware of the flaws in what we have done, and the tasks that we have left undone. The former, at least, we hope to correct over time. We also hope to carry out the latter if given the opportunity.

The first chapter tells the more particular story of the execution of the project: educative, exciting, exhausting, sometimes frustrating, a little creepy when we turned away from our screens to survey the seemingly unreal prospect that lay ahead. Looking back now that it has turned real, I can allow myself the kind of self-indulgent shudder I firmly suppressed at the time.

Some salient persons have been named in Chap. 1 with (I hope) suitable appreciation and gratitude, but a few can never be thanked enough. Among them are Jawhar Sircar and Udaya Narayana Singh. Others are not named there at all, like Supriya Roy of Santiniketan and Saranindranath Tagore of Singapore; also the authorities and staff of the Indian National Library, C-DAC, CSSSC, the Calcutta University Library and the Bangiya Sahitya Parishat. Sankha Ghosh was an unfailing source of inspiration, scholarly advice and practical assistance.

Needless to say, the project could not have been taken up at all without the resources of Rabindra-Bhavana, Santiniketan.

Of my colleagues at Jadavpur University it seems invidious to name some and omit the rest, but I must run the risk. Thanks to the Vice-Chancellors waving us on at the starting and finishing lines respectively, Pradip Narayan Ghosh and Souvik Bhattacharya. Warmest and most affectionate thanks to Subha Chakraborty Dasgupta, Amlan Das Gupta, Samantak Das and Chandan Mazumdar. Thanks no less to Gour Krishna Pattanayak, Sanjoy Gopal Sarkar, and the members of the Major Projects Cell, the Central Library and the IT and Systems Management team.

It would be truly invidious to single out any one of the group that prepared the contents of this book for me to wrap in a shiny package. They have been named on

a separate page. Some of them, with a few others, also feature in the text. An Appendix lists the entire crew that worked on the project. I can now express, as perhaps I did not at the time, the love and appreciation I felt for them through those two memorable years. Bichitra has afforded the richest professional experience of my life, in human as well as intellectual terms.

Bichitra was funded by the Indian Ministry of Culture, graciously launched by the President of India, and dedicated to the nation. I may be pardoned for adding a personal codicil. My father Kanti Prosad Chaudhuri passed his childhood and youth during Tagore's later life, when his works appeared in a continuous stream to public acclaim. Brought up on that fare, my father always upbraided me for not devoting enough time and study to the poet. I have not done so to this day, as the example of Sankha Ghosh, Swapan Majumdar and others continually reminds me; but through Bichitra, I have tried to make good something of that lack. Belatedly and inadequately, I dedicate my personal part in the project to my father's memory.

Eleven years ago, some colleagues and I came together to set up the School of Cultural Texts and Records at Jadavpur University. It has grown from a single room (where not everyone could sit down at the same time, and a single computer might serve two projects) to spacious and enviably equipped quarters in a new building. It has also won acknowledgement as a 'top of the class' world centre of digital humanities. I hope it retains the structural and institutional freedom to allow the making of more Bichitras in the years to come.

Kolkata, India

Sukanta Chaudhuri

The Contributing Team

The data that went into this book was compiled by key members of the original Bichitra team, each contributing material relating to their roles in the project as detailed below. This data was recast, sometimes translated from Bengali, and put in final form by Sukanta Chaudhuri, who also wrote Chaps. 1, 2 and 10. The volume was text-edited by Debapriya Basu. The illustrations were prepared by Kawshik Ananda Kirtaniya.

Chapter 3 Fonts and OCR: Dibyajyoti Ghosh

Chapter 4 Images and Scanning: Purbasha Auddy, Kawshik Ananda Kirtaniya

Chapter 5 Manuscripts and Transcription: Smita Khator and Sahajiya Nath,
with contributions from Amrutesh Biswas and Aparupa Ghosh

Chapter 6 Data Management and Hyperbibliography: Purbasha Auddy, Debapriya
Basu

Chapter 7 Search Engine and Hyperconcordance: Dibyajyoti Ghosh in consultation
with Prakash Koli Moi and Arabinda Moni

Chapter 8 Collation: Spandana Bhowmik, Sunanda Bose

Chapter 9 Planning the Website: Ritwick Pal, Purbasha Auddy

Notes and Conventions

1. As explained in Chap. 6, Tagore's works are variously dated by three systems: the Common era (CE), the Bengali era and the Saka era. The last is not relevant to the material in this book. Where a book or journal item appeared with the Bengali date, that is given first, followed by the CE after a slash. In all other cases, only the CE year is given.
2. Titles of Tagore's Bengali works are followed by an English rendering except in a few untranslatable cases, or where the title is a proper name.
3. Manuscripts are indicated by the holding archive: RB (Rabindra-Bhavana, Visva-Bharati, Santiniketan) or HL (Houghton Library, Harvard University) followed by the shelfmark.
4. Bengali words have been transliterated by a simplified method avoiding diacritical marks. The same letter in the Roman (English) alphabet can thus stand for two or more Bengali letters like two i-s, two u-s, three r-s, three s-s, and hard and soft forms of the same consonants.
5. The city where we live and work is today officially called Kolkata. However, a few institutions, including one of India's oldest universities, still retain the form 'Calcutta' in their names. We have respected this practice in the interests of accuracy as well as tradition.

Contents

1 The Story of the Bichitra Project.....	1
Sukanta Chaudhuri	
2 Tagore's Text.....	7
Sukanta Chaudhuri	
3 The Bengali Writing System: Fonts and OCR	13
Sukanta Chaudhuri and Dibyajyoti Ghosh	
4 Images and Scanning	21
Sukanta Chaudhuri, Purbasha Auddy, and Kawshik Ananda Kirtaniya	
5 Manuscripts and Their Transcription.....	31
Sukanta Chaudhuri, Smita Khator, Sahajiya Nath, Amrithesh Biswas, and Aparupa Ghosh	
6 Data Management and Hyperbibliography.....	59
Sukanta Chaudhuri, Purbasha Auddy, and Debapriya Basu	
7 Search Engine and Hyperconcordance	93
Sukanta Chaudhuri, Dibyajyoti Ghosh, Prakash Koli Moi, and Arabinda Moni	
8 Collation: Prabhed and Its Predecessors	99
Sukanta Chaudhuri, Spandana Bhowmik, and Sunanda Bose	
9 Planning the Website	131
Sukanta Chaudhuri, Ritwick Pal, and Purbasha Auddy	
10 Beyond Bichitra.....	143
Sukanta Chaudhuri	
Appendix: The Bichitra Team.....	155

Sukanta Chaudhuri

Dreams, Plans and Prospects

Bichitra is the happy outcome of a number of people being in the right place at the right time, starting with Rabindranath Tagore's having been born in 1861. On the 150th anniversary of his birth, in 2011, the Government of India decided to sponsor a grand commemoration of India's de facto national poet. Among the projects they generously agreed to support was a comprehensive website of Tagore's works in English and Bengali in all available versions.

This book explains what a gigantic task it was—maybe more so than envisaged by the Indian Ministry of Culture, or indeed by the members of Jadavpur University who took up the task. Speaking as head of the project, I can say that though we knew what the work involved in quantitative terms, we had not, truly speaking, *imagined* it. Perhaps this was just as well: we may not have ventured upon it otherwise.

The School of Cultural Texts and Records had been set up at Jadavpur University, in the city of Kolkata (Calcutta), in 2004, as a centre for all kinds of textual studies, especially archiving, documenting and editing. As expected, our work engaged more and more with the electronic medium, till today (according to a survey by the Council on Library and Information Resources, Washington DC) the School ranks as one of the world's 'best in class' centres of digital humanities (Lewis 2015, 1, 7).

With our access to the Bengali language and to texts in that language, we were uniquely placed to explore the possibilities of electronic data collection, data

A personal account by Sukanta Chaudhuri.

S. Chaudhuri (✉)

Department of English, Jadavpur University, Kolkata, India

e-mail: schaudhuri@english.jdvu.ac.in

mining and editing in the works of Rabindranath Tagore. Chapter 2 explains the singular potential of the Tagore corpus as the all-time test case for virtually every issue of textual editing and data mining. For years before Bichitra, we had undertaken small exercises in variorum editing of Tagore's works, created some promising collation software, and issued (offline) an experimental electronic variorum of the play *Bisarjan (Sacrifice)* and a more elaborate one of the poetical collection *Sonar tari (The Golden Boat)*. We had joined a confabulation at Santiniketan, home of Visva-Bharati, the university founded by Tagore about 100 miles from Kolkata, to create a comprehensive Tagore database as the foundation for a scholarly print edition of his complete works. That hyper-ambitious plan did not come to pass, but Bichitra embodies the electronic part of the project. We understand Visva-Bharati is proceeding with plans for a print edition on chronological lines.

The bounty of the Ministry of Culture allowed us to fulfil our dream of a comprehensive Tagore website: images of all manuscripts and authoritative print editions totalling nearly 140,000 pages, reading texts of every version, detailed transcripts of all manuscripts, a full bibliography, an in-depth search engine, and a new collation program to analyze Tagore's complex texts layer by layer as no extant program could do. All these components were to be interlinked within an integrated database. And we had to do it all in just over two years. The project was sanctioned in November 2010, work started in March 2011, and the site was launched just before Tagore's birthdate in May 2013.

One only has to spell out the project in these terms to see how crazy it sounds. But as I said, though we had to spell it out for the Government (and ourselves), we were crazy enough not to realize how crazy it was, or at least not to be deterred by the prospect. More improbably, the hard-nosed officials of the Ministry of Culture allowed themselves to be persuaded. We owe very special thanks to Jawhar Sircar, India's Culture Secretary at the time, who has somehow retained the capacity to dream dreams and steer them to fulfilment through the banks and shoals of the bureaucracy. For that brief period, the Culture Ministry was directly looked after by the Prime Minister of the day, Manmohan Singh. We profited by making our bid during that brief spell when culture featured exceptionally high on the India Government's agenda.

Closer home, we owe a great debt to another source, Tagore's university Visva-Bharati and its museum and archive, Rabindra-Bhavana. They were our chief project partners, as they must be in any project of this sort: they hold all the material. Rabindra-Bhavana is by far the biggest repository of Tagore manuscripts, and the biggest one-stop archive for print editions and journals containing his works. The deal was that they would provide the material, while we at Jadavpur processed it and set up the website. All material was supplied in digital copy: we did not touch the originals, nor did we need to.

This understanding, simple to state, could have taken ages and run into all kinds of problems, had it not been for the openness and enthusiasm displayed at both ends, Santiniketan and Jadavpur. Udaya Narayana Singh, that human dynamo, was then officiating as Director of Rabindra-Bhavana. I emailed him in early June 2010 suggesting we meet to discuss the project. He mailed back to say that, by good luck,

there would be a meeting of their governing committee in a few days' time. Could I send him a draft proposal immediately, and attend the meeting in person to follow up? I rustled up a formal document over the next 48 hours, and took the early morning train to Santiniketan on the day of the meeting. Rajat Kanta Ray, then Vice-Chancellor of Visva-Bharati, was in bed with a severe back problem, but he graciously invited Udaya and me to discuss the matter at his bedside. He also made it to the committee meeting, where we had a stimulating discussion on my proposal. The committee approved the participation of Rabindra-Bhavana and empowered Udaya to negotiate further, keeping the Vice-Chancellor in the loop.

At the Jadavpur end, I knew there would be no problem, though I had not yet discussed the matter with anyone in authority. It is this freedom of operation, this willingness to entrust the man on the spot with major negotiations, that has made our small and chronically under-funded institution into one of India's leading research universities. The jewels in its crown are a chain of 21 Interdisciplinary Schools, the School of Cultural Texts and Records among them. As I write, moves are far advanced to reshape the Schools on new lines. This book may provide a criterion, among many others by my brilliant colleagues, for the new model to match.

I received full support when I belatedly reported my initiative to the Vice-Chancellor of the time, Pradip Narayan Ghosh, a physicist by profession and a Tagore enthusiast like so many Bengalis. There followed a hectic round of paperwork, including the preparation of a 50-page, 20,000-word work proposal curiously called the Detailed Project Report (DPR), though at this point there was nothing to report except our dreams. This passed through the ranks of Delhi officialdom with amazing speed, and by November 2010 the first instalment of funds had landed in the university's bank account. As I have said, the Culture Ministry was granted exceptional importance at the time, with the Tagore Anniversary programme given top priority. There was a special committee to oversee the anniversary projects, chaired by Pranab Mukherjee, then India's Finance Minister. It was only appropriate that he should launch the finished website in May 2013 in his subsequent capacity as President of India.

The Ministry of Culture kept up its incredible level of support all through the project. Funds often crawl at snail's pace between Delhi and the corners of the country, but ours flew on wings. Any request or enquiry was promptly addressed. I should again thank Jawhar Sircar, and his successor when we completed the project, Ravindra Singh; also the successive Joint Secretaries Tuktuk Ghosh Kumar and Pramod Kumar Jain, and their colleagues in the Ministry.

We also have pleasant memories of the hospitality of Santiniketan and Rabindra-Bhavana; but sadly, our experience on that front ended in an impasse that left its mark on the project. After a change of guard at Rabindra-Bhavana, we were told by C-DAC, the agency scanning the material, that they had been instructed not to supply us with any more copies. The Rabindra-Bhavana authorities assured us that this was a misunderstanding, that they would not renege on their commitment; but despite many appeals from us, they never sent C-DAC the letter that would have resumed the supply.

Fortunately, by this time we had received scans of all the manuscripts and most of the printed material. Several other institutions rallied round to make good the loss: the Indian National Library, the University of Calcutta, the Bangiya Sahitya Parishat (Literary Academy of Bengal), the Centre for Studies in Social Sciences, Calcutta, as well as some private collectors and Jadavpur's own Central Library. We are deeply grateful to them all. But some gaps could not be filled. In round figures, we estimate that Bichitra lacks some 40 volumes out of 450, and 300 journal items out of 3200. Some of these are genuinely unobtainable; the rest are missing owing to this contretemps.

I must thank the Houghton Library, Harvard University, for their ready consent to supply copies of all Tagore's English manuscripts in their Rothenstein Collection, following a single conversation when I went there for my own work on Western manuscripts. Thanks also to the Senate House Library, University of London, where our team member Dibyajyoti descended during a brief visit to England to photograph that surprisingly elusive item, the first Macmillan imprint of the English *Gitanjali* (*Song Offerings*), the work that won Tagore the Nobel Prize in 1913. We value the support of these institutions abroad with respect to the author we call our *visva-kavi*, 'world-poet'.

Getting Down to Business

Such help and support from other quarters only increased the pressure on ourselves. Readers may be wondering why, if we received the funds in November, we could only begin operations in March. To start with, the School lacked proper premises at the time. Our first job was to ask the University for space—never an easy task. Once that battle was won, the rooms had to be fitted up with everything from air conditioners to enough wiring for 30 computer terminals and much else, not to mention tables and chairs. Those 30 terminals caused a minor diversion when audit inspectors came to ask why we needed that quantity of hardware to research the poetry of Tagore. They ended up listening eagerly to our account of the project. They too were young Bengalis addicted to Tagore.

At the same time, we had to go through the involved process of inviting applications and appointing members of the project team. Never let it be forgotten that Bichitra was as much about human management as computer operations; and we may still fairly claim that humans are more complex entities than computers. There was a core staff of 30, on modest but adequate salaries totalling a sum that, I am told, would have paid for just two hands in Britain. Except for two office staff, they were young scholars with good Master's degrees, nearly always in the humanities but with formal or informal computer training. In addition, we employed extra hands as needed all through the project: sometimes to catch up on routine work like transcription, sometimes to help with file and website management in the last phase. Others were engaged on contract for specialized tasks, above all to create the customized software and design the website itself.

Having retired, I could make it my full-time job as project director to co-ordinate this range of operations. While I might have known more about textual matters than

my young colleagues, I knew less about virtually every aspect of their hands-on work. My resort, which I would suggest to everyone in such a situation, was to respect their greater skill and knowledge, fall in with most of their suggestions or even invite them, guiding their efforts (with as light a hand as possible) only because I had the best view of the big picture, not to mention such matters as deadlines and the ways of a bad world. I do not know whether they would agree with this charitable view of my role: I am writing this chapter entirely by myself. For the rest (except Chaps. 2 and 10), this book has taken shape rather as the website did: with my erstwhile colleagues' input (which I emphatically could not have provided), co-ordinated and overseen by myself—in fact, reshaped and rewritten more radically for the book than I could have done for the website.

One of the joys of working after one's retirement, at least in India, is that the rules require someone else to look after the paperwork and accounts. My deepest gratitude goes to Subha Chakraborty Dasgupta and Samantak Das, the colleagues who successively took up this thankless task. In another direction, I am no less grateful to Sankha Ghosh, that doyen of Tagore scholars, who acted as adviser to the project and gave us the benefit of his encyclopedic grasp of all things Tagorean.

Barring one or two misfits we had to ease out, the team of 30 proved exceptionally committed, though a few left for various reasons and had to be replaced. The one occasional cause of strain was the need to adhere strictly to a schedule of work, especially with the transcription. When we started, my well-wishing colleagues feared for my rashness, and advised me to scale down the project before it was too late. I was armed with a set of figures showing that we could complete the transcription in time if we stuck to 20 pages a day per operator for prose, and 30–35 pages for verse. But of course this could not apply to manuscripts, where each item had to be scheduled on its own terms. Also, the number of variant texts far exceeded my first ballpark estimate. To compensate, we had the unexpected boon of a near-complete set of files using an OCR program devised by an erstwhile Jadavpur researcher, Anirban Raychaudhuri: so we did not have to key in each version from scratch, but only to modify this master copy. Anirban Raychaudhuri created the OCR files for a new edition of Tagore's collected works published by Visva-Bharati. Modifying the master copy was not as simple as it sounds. An account of the business is given in Chap. 3. All in all, my work schedule proved feasible, but it obviously allowed no margin for manoeuvre. It is not easy for intelligent people to copy thousands of pages of great literature month after month in more or less mechanical fashion without letting their minds wander, if only to thoughts inspired by those writings. They could have dealt more easily with an intellectually challenging task, but they successfully fought the demon boredom for months and years.

A greater worry by far was that we were committed to elaborate software programs that effectively did not exist. The search engine promoted itself to a hyper-concordance, but I knew we could handle that one way or another. The real challenge was the collation program. As Chap. 8 explains, we did have a program to start with, but it would not meet our vastly expanded needs under Bichitra. We worked for some time, along with colleagues in Chicago and Brisbane, to develop a new

approach; but promising as this might prove in time, it became clear that it would not serve our immediate purpose.

Like a gift from above, three young programmers arrived on the scene, one of them quite fortuitously, to create a working program in seven or eight months. They did more: they produced one of the world's most sophisticated multi-level collation programs to date. I cannot recount all the amazing coincidences that paved the way, nor the intensive work, through days and nights for many months, to make it happen. Chapter 8 affords a few brief glimpses of the story, besides a full account of the program itself. Chapter 9 indicates how we took the same approach in choosing the designer and planning the site: highly intensive exchanges in an informal low-keyed environment, without heed of work hours and formal obligations. Given the unusual nature of the task we set ourselves, we had to think out of the box. It was thus doubly appropriate that circumstances should force out-of-the-box methods of work upon us.

Major projects in India, perhaps specially in Kolkata, must often be carried out by very different means from the 'international', by which we usually mean the prevailing norms in the West. Kolkata's first 16-km Metro Railway was built by the 'cut and cover' method—that is, by digging open trenches from the surface down, in good part with pickaxe and shovel. This was not simply to save money, but because the soil and climate made tunnelling unviable. The operational technology was as sophisticated as any in its day, and needed further ingenuity to adapt it to local conditions.

Like much else in this resourceful environment, Bichitra may be said to reflect the same philosophy. Claude Lévi-Strauss contrasted the *bricoleur*, the ingenious designer by makeshift means, with the engineer of approved orthodox methods (Lévi-Strauss 1972, 17). The distinction fails in cases like ours, for we were pursuing—and dare I say achieving—the ends of the engineer by means that often smacked of the *bricoleur*. We thought, improvised, sometimes almost wished our way through problems: we could not hope for elaborate support according to the best practice. In our favour, our environment made it appropriate for us to think small even when planning something big.

Bichitra means 'the various'—an appropriate name for a variorum website. But the masculine form of the same word can also mean something like 'curious, bizarre'. That too seems appropriate for this improbable creation. All we can say is, it works. To borrow Galileo's words in a grander context, *Eppur si muove*—it nonetheless does move.

References

- Lévi-Strauss, Claude. 1966. *The Savage Mind*. Reprint, London: Weidenfeld and Nicolson, 1972.
- Lewis, Vivian, Lisa Spiro, Xuemao Wang, Jon E. Cawthorne. 2015. *Building Expertise to Support Digital Scholarship: A Global Perspective*. Washington, DC: Council on Library and Information Resources.

Sukanta Chaudhuri

In 1897, Rabindranath Tagore composed a famous love song that he revised extensively thereafter: 'Tumi sandhyar meghamala' ('You are like a cloudbank in the evening'). The lover addresses his beloved in these terms. But the original version, written in the poet's own hand in a notebook belonging to his niece Indira Debi, has a markedly different thrust. There we do not have the Sanskritic vocative *ayi*, addressed only to women or objects grammatically of feminine gender. The word *bijan*, an open space or wilderness, associated with the god Krishna, occurs three times though only once in the standard version. And the last word, addressed to the beloved, is *mohanamaranabihari*, 'One that moves in beautiful death', or by dividing the compound word differently, 'The beautiful one that moves in death'—either way, clearly referring to Krishna.

The differences affect eight words out of 68. They are enough to show that what in the standard text is a male lover's address to his beloved began as a woman's address to a male, in particular as Radha's address to Krishna. Radha was Krishna's chief consort among the herdswomen of Vrindavan, where Krishna grazed cattle. Her love for Krishna is conventionally spiritualized as the human soul's love for the divine. In other words, a poem of human love is underlain by one of erotic mysticism.

The song occupies half a page out of 955 pages of text (excluding notes, index etc.) in the final edition of Tagore's song collection *Gitabitan*. *Gitabitan* is a stand-alone collection, a large proportion of its pieces not included in the 32 volumes of Tagore's collected Bengali works (*Rabindra-rachanabali*) so far published by Visva-Bharati, the university he founded. At least two more volumes are due, of items scattered in journals and anthologies. Two others were published early on

A part of this chapter has appeared previously in *Towards Tagore* (Dasgupta et al. 2014).

S. Chaudhuri (✉)

Department of English, Jadavpur University, Kolkata, India

e-mail: schaudhuri@english.jdvu.ac.in

with the poet's juvenilia. His English writings fill four large volumes published by Sahitya Akademi, the Indian Literary Academy, but many English items remain uncollected or even unpublished. There is also a quantity of unpublished Bengali writings.

The intrepid scholar may see this vast corpus—surely the largest by far of any writer of comparable stature—as an immense terrain to explore; the faint-hearted, as a bottomless textual quagmire. Chapter 6 gives more details of the sheer volume of material and number of titles, ranging from short poems to large novels. The surface area of the corpus—that is to say, taking just one standard text of each title, as in the collected works—is multiplied many times, and problematized still more, by a number of factors as detailed below. All references are to the Bengali text of the work, not the English translation if one exists.

1. Tagore was a relentless reviser, both before and after publication. An outstanding instance is the early poem 'Nirjharer svapnabhanga' ('The Spring Wakes from Its Dream'). It first appeared in 1882 in a 201-line version in the journal *Bharati*. This was expanded to 267 lines in *Prabhat-sangit* (*Morning Songs*, 1883), the volume in which it next appeared, then scaled down to a version of around 150 lines which appeared in various recensions. Even this must have seemed to Tagore too long for a seminal but immature poem, for the version included in the popular anthology *Chayanika* (1909) was only 87 lines long, further reduced to 43 in the later collection *Sanchayita* (1931). This is the version that most readers know today: its opening line, known to all educated Bengalis, is not that of the original poem. Yet volume 1 (1939) of the collected Bengali works still provides a 154-line text—as part of *Prabhat-sangit*, which originally contained a much longer version.

Early poetical collections like *Manasi* (*The Woman of the Mind*) and *Sonar tari* (*The Golden Boat*) yield other striking examples. The poem 'Barshar dine' ('On a Rainy Day') in *Manasi* has 20 versions. Even very late volumes like *Rogshajyay* (*On My Sickbed*), *Arogya* (*Recovery*) and *Janmadine* (*On My Birthday*) show substantial revision over a much shorter span of time. The poet even expressed a wish to revise the last poem he wrote, dictated from his deathbed.

Tagore's lyric poems often exist in parallel 'read' and 'sung' versions. The drama and fiction can undergo radical recasting, with addition, deletion and relocation of whole scenes, chapters and even characters. In the plays, songs might be added, subtracted and relocated on a considerable scale, for productional as well as creative reasons. *Arupratan* (*The Invisible Jewel*) and *Shapmochan* (*The Lifting of the Curse*) provide striking examples, as do the various manuscript versions of *Raktakarabi* (*Red Oleanders*), and the early editions of *Raja* (*The King*, entitled *The King of the Dark Chamber* in the standard translation wrongly ascribed to Tagore).

As expected, the greatest range of variation is usually in the early works, which allowed most time for revision. The play *Bisarjan* (*Sacrifice*), first published in 1890, has eight (more closely viewed, 13) printed versions, the longest ten times the size of the shortest: testimony to continuous post-print revision over some 40 years. There can also be drastic pre-publication revision. *Raktakarabi* saw ten manuscript drafts in a short span of time before publication or performance.

Perhaps unexpectedly, the greatest revision occurs in the English collection *Talks in China*, where individual items might have 30 or more versions. On the whole, the English works show a degree of revision out of proportion to their bulk and importance. Perhaps Tagore was extra-cautious when writing in a language not his own: he also had to consider his growing international image. There are other types of complicated variation as well. Examples are scattered through Chap. 6.

2. Where there is so much major restructuring, it goes without saying that the quantum of local variants, affecting single words, phrases or lines, is incalculable. No doubt they were chiefly made by the poet himself, but many could be owing to the circumstances of printing and publication. His most popular works were (and are) continually reprinted. From 1923, all the Bengali works were printed and published by Visva-Bharati itself. Given this authoritative source, their variant readings must be taken seriously. Some may have been introduced by Tagore's lieutenants who oversaw the publication of his works. Some may simply be errors. But given the mass of little-explored documentation, there is always the chance of an authorial variant surfacing in a late edition. For instance, a reading in the poem 'Anadrita' ('The Unloved') was corrected from the manuscript only in the 1390/1983 edition of *Sonar tari*.

3. We therefore have to ask: at what stage of composition did the changes occur? Strictly speaking, a 'mere' database need not address this question: it is for the user to mine the data and find the answer. But an advanced database must present the data in appropriate form for such queries, and pre-process it as far as possible. The Bichitra bibliography records the dates of publication (not composition) of the printed versions. We have not taken up the immense challenge of dating the manuscripts, or (for the task to be worth doing at all) separately dating their individual items: that would be a major project in itself, involving much palaeographical and historical research. The pattern of variants thrown up by the collation engine obviously yields the basis for a stemma or 'family tree' connecting the versions, tracing the genetic history of the work. But Bichitra (or more precisely its collation software Prabhed) does not construct that stemma, as Collate (now housed at the University of Saskatchewan), the only collation program of comparable scope, can do.

In any case, as indicated in point 2 above, Tagore's publishing history is so involved that it obscures the compositional history. A collation program would have to be doubly complex to accommodate the two separately, incrementally more so to relate them. The revisions were not consistently incorporated in print. If we apply the dubious rule of thumb of adopting the last printed version in the author's lifetime, we may find three variant texts around that time: one in the collected works, a second in the separate volume containing the poem, a third in an authorized anthology.

4. The poet would shift his works from one setting to another, often more than once. Poems might move from collection to collection, or be embedded in plays or novels. Again, Chap. 6 cites many instances. A specially moving one is the song 'Samukhe shanti parabar' ('Ahead, the ocean of peace'), composed for a 1939 performance of *Dakghar* (*The Post Office*) but, at the poet's request, reserved to be sung after his own death.

5. Passing beyond mere relocation, we often find the same fable or plot used in, say, a narrative poem, a play and/or a novel or short story—perhaps more than one version of each. The novel *Rajarshi* (*The Royal Sage*) shares its story with the play *Bisarjan*, whose ramifications we have already seen; the poem ‘Parishodh’ (‘Reparation’) with a musical drama of the same name as well as the dance drama *Shyama*. There is a complicated tangle of relations between the plays *Raja*, *Arupratan* and *Shapmochan*; or the novel *Prajapatir nirbandha* (*The Marriage-God’s Decree*) and a play and a novel both named *Chirakumar sabha* (*The Society of Celibates*), so close to each other that they can be collated despite the difference of genre. Tagore even reworks passages from his letters into certain prose poems in *Punascha* (*Postscript*), but Bichitra does not include Tagore’s letters in its repertoire.

Such generic transformations are usually undertaken by later writers. Shakespeare reworked many novellas by earlier writers into plays, from which others in turn made novels or other plays. Seldom if ever has the same writer recast his material across genres on this scale. This gives Tagore’s works an unusually self-referential quality. His various works pick up one another’s threads, refer back and forth among themselves, extend one another’s meaning. He may directly refer to one work in another. His discursive prose writings contain countless references to his poems, plays and works of fiction. He may work himself into the fiction, as in the novel *Shesher kabita* (*The Last Poem*), either by name or under a thin disguise. This novel includes many poems, some ascribed to a fictitious opponent of Tagore’s style—and subsequently included separately in the verse collection *Mahua*, published of course in Tagore’s own name! These are the overt outworks of a web of correspondences that operate also at the level of the topos, the sentence, even an operative word.

6. The last major factor is translation. Tagore translated his own writings extensively into English, as did many members of his circle. Bichitra only includes the poet’s own renderings, often very free adaptations. Bengali and English versions of the ‘same’ poem thus offer insights and commentaries on each other. Short sections of several Bengali poems, sometimes little more than disjunct phrases, might be combined in a single English piece or scattered through a number of them. In literary merit, the English versions may compare poorly with the Bengali, but they extend the bounds of Tagore’s textual universe. There are reverse instances too. At least 28 English pieces do not have an identifiable source in Bengali. A Christmas poem entitled ‘The Child’ was originally written in English and later ‘translated’ into Bengali. The intertextual exchange becomes interlingual. This is not unique to Tagore—Nabokov and Beckett provide ready modern parallels—but it is specially marked in him.

These factors point in a clear direction. Even more than usually for a major author, a simple database of images and transcriptions cannot provide an adequate platform for engaging with Tagore. It must be reinforced by some basic resources for data mining. ‘Basic’ does not mean sketchy or rudimentary; it means conceptually fundamental, presenting (in full detail) the findings and configurations of textual data needed for an understanding of the works.

We therefore decided to equip the Bichitra site with three major tools: a hyperlinked bibliography; a search engine cum hyperconcordance; and a collation engine that would do justice to the uniquely far-flung, layered intricacy of Tagore's textual genetics. All three tools needed to be custom-built. There was precedent for the first two, though our strategies worked out as quite different. (The Society for Natural Language Technology Research, a consortium of scholars from Kolkata and nearby Kharagpur, had already created an online searchable database of the standard edition of Tagore's Bengali works.) The collation engine had virtually to be created from scratch, as existing models patently could not meet our purpose. Most of them would not effectively yield results for multiple texts on any scale, and none could conduct the three-tier collation we required (chapter/scene/canto, paragraph/speech/stanza, and individual word).

Imaginative use of a comprehensive Tagore concordance, hyperlinked for one-click access to full searchable texts of the works, can open up exciting lines of interaction within a uniquely large corpus of texts controlled by a single authorial intelligence. This is precisely what the Bichitra hyperconcordance affords. It also offers basic bibliographical information, filled out by the hyperbibliography. This is 'hyper' in that it allows one-click access to images of every version of a work, print and manuscript, as well as a clear-text file. It also has a link to the collation engine. Bichitra thereby offers a raft of linked resources illuminating not only Tagore's own works but, more basically, the growth and circulation of texts, the textual process itself.

Tagore's textual transactions expand in scope on a graduated scale: from local variant readings, through major structural revisions of the 'same' text, to trans-genre reworkings of the same topos, fable or narrative structure; ultimately, to works with no evident relationship to each other but linked by a common creative endeavour. The intratextual complexities in individual works open out seamlessly into a degree of intertextuality rare within the writings of a single author, magnified by the sheer scale and range of the corpus. While each version of each work has its own formal identity, we cannot clearly demarcate its range: every text participates in the being of every other text.

'Literature is an ongoing system of interconnecting documents.' This was said by Theodor (Ted) Nelson (Nelson 1981), ideologue of the Internet, to express the rationale of the World Wide Web and its perfection in the (as yet) unrealized dream of Nelson's Xanadu project. The remark is usually taken to imply a reduction of the author's role, a stress on the social, circulatory aspect of texts. One thinks of terms like 'multiverse' and 'docuverse', coined by theorists of electronic texts and the Internet, to convey this sense of an interactive textual cosmos. But we can apply these terms and concepts to the same process as enacted within the compass of a single author's work, a kind of local area network connecting an author's entire corpus.

Bichitra is designed to reflect in its structure the Tagorean textual universe. Reversing the relationship, we may say that Tagore's works offer a uniquely prominent instance of the complex ways of texts—their growth, change and interrelationships—whose best (though still inadequate) repository is an advanced electronic

database. Michelangelo said in a sonnet that he did not create the forms of his sculptures: he only released the form latent in the stone. We have tried to capture something of the latent form of Tagore's genius in a medium that would surely have fired his imagination had he lived to see it.

References

- Dasgupta, Sanjukta, Ramkumar Mukhopadhyay, Swati Ganguly, eds. 2014. *Towards Tagore*. Kolkata: Visva-Bharati.
- Nelson, Theodor. 1981. *Literary Machines*. Sausalito: Mindful Press. As excerpted in www.units.muohio.edu/technologyandhumanities/eng495/Ted%20Nelson%20What%20is%20Literature.htm. Accessed 14 December 2014.

Sukanta Chaudhuri and Dibyajyoti Ghosh

Bengali (or Bangla) is the world's seventh most widely spoken language, mother tongue of over 3 % of humankind. Its speakers are concentrated in a small but densely populated part of eastern South Asia: the sovereign country of Bangladesh, whose official language is Bengali; and the Indian states of West Bengal, Tripura and a part of Assam, where Bengali is one of India's 22 official languages. There is also a sizeable Bengali-speaking diaspora spread across the rest of India and the world.

But despite this strong demographic presence, and an impressive literary and cultural history with Tagore at its crown, Bengali is not a 'world language' in the wider sense. Hence the figures for the digital presence of Bengali are sadly different. Bengali ranks 80th in the list of Wikipedias, whereas a dead language like Latin is 48th as on 15 July 2015 (https://meta.wikimedia.org/wiki/List_of_Wikipedias).

This low digital presence of Bengali is owing to several factors. While Bangladesh has a national policy of developing digital resources for its official language Bengali, in India such attention is focused on the chief official language, Hindi. (Even there, the motivation is relatively low, as digital resources are plentiful for the other official language, English.) As explained in detail below, the Bengali script is cumbersome to render on the keyboard. Till very recently, indeed even today, these factors have restricted the growth of computing in Bengali (or other South Asian languages). Especially in India, most people with the education and resources to use computers at all habitually do so in English. This has been a major factor in promoting greater use of English instead of Bengali (or other Indian languages) in India. In fact, in order to use Bengali in the digital realm, one virtually needs some

S. Chaudhuri (✉) • D. Ghosh

Department of English, Jadavpur University, Kolkata, India

e-mail: schaudhuri@english.jdvu.ac.in; ghosh.dibyajyoti@gmail.com

© Springer International Publishing Switzerland 2015

S. Chaudhuri (ed.), *Bichitra: The Making of an Online Tagore Variorum*,
Quantitative Methods in the Humanities and Social Sciences,
DOI 10.1007/978-3-319-23678-0_3

13

knowledge of English. On top of this, very many people in both India and Bangladesh simply cannot afford to buy an expensive product like a personal computer.

The consequent lack of a large market has discouraged the computer industry from developing resources for Bengali. Microsoft Windows®, the dominant operating system in both India and Bangladesh, introduced a Bengali version only in 2011. Adobe provided support for text editing and photo/video editing in Indic scripts in 2012, with their CS6® package. Google, too, has introduced Bengali tools and webpages, but most users prefer the English version. The story is much the same, if not still more depressing, with all South Asian languages.

The bright side of this situation is that it leaves the field open for developing resources. Our work on Bichitra, and the activities of the School of Cultural Texts and Records generally, have been very exciting for this reason. It has also been challenging owing to many factors, beginning with the nature of the Bengali writing system.

The Bengali Alphabet and Keyboard Software

Bengali has a phonetic alphabet—in fact, following the Sanskrit model, a much more ordered and phonetically consistent one than the Roman alphabet used for English. There are about 50 letters. (We cannot fix the exact number as certain forms may or may not be classed as separate letters: for instance, hard *d* and *dh* trilled to become hard *r* and *rh*, or a truncated form of the soft *t*.) As there are no capitals, the total number of characters is much the same as the 26×2 set of the English alphabet. But given the 26 letter keys in the standard keyboard, about half the Bengali letters call for the shift key—i.e., the use of both hands. An English sentence, by contrast, might have just one upper-case letter at the start, and even that can be pre-programmed. This makes the Bengali keyboard slower and more cumbersome to handle.

But the real catch lies elsewhere. The Bengali writing system is much more complicated than that of the Roman or other Western alphabets. In the latter, each letter is written fully and separately, at most with an accent attached. Bengali, like most Indic languages, has what is called an abugida writing system, where a combination of two or more letters can constitute a single unit or glyph. This makes it hard for Bengali children to learn to read. (To compensate, Bengali spelling is vastly more regular than English.) It also makes it harder for computers to read Bengali.

First and foremost, a vowel is written in full only where a syllable consists entirely of vowel sounds. Where (as most often) it is combined with a consonant, only the consonant is written in full, and the vowel sound indicated by a tag or marker attached to it (see Fig. 3.1).

Modern Bengali has eleven vowels (one of them sounding rather like *r*, but a vowel in its root identity). This seems extravagant compared to the five in the Roman alphabet; but there each vowel can be pronounced in a confusing variety of ways, whereas only two Bengali vowels have two pronunciations each. The commonest vowel, somewhere between *a* and *o*, does not need a marker: it is assumed in the

Fig. 3.1 Bengali vowel markers attached to a consonant

ক	+	আ	=	কা
k	+	ā (long a)	=	kā
ক	+	ঈ	=	কী
k	+	ī (long i)	=	kī

Fig. 3.2 Aberrant forms of vowel markers. The same vowel *u* is attached by three different markers to the consonants *k*, *r* and *ś*.

ক	+	ঊ	=	কু
k	+	u	=	ku
র	+	ঊ	=	রু
r	+	u	=	ru
শ	+	ঊ	=	শু
ś	+	u	=	śu

absence of one. The other ten vowels have visible markers; combined with 40 or so consonants, they produce some 400 conjunct glyphs. These vowel markers have standard forms, but with many aberrations when added to particular consonants (see Fig. 3.2).

To complicate things further, though the vowel sound phonetically follows the consonant (and is keyed in after it), the vowel marker might be placed before, after, above or below it, sometimes combining more than one of these positions (see Fig. 3.3). This is confusing for the computer: it needs to be specially programmed for a vowel marker keyed in after a consonant to be displayed before it. The user’s computer, too, needs to be loaded with a program to read Indic fonts correctly. If a vowel marker normally placed before the consonant appears after it on the screen, it will not affect the computer’s operations, but will be unacceptable to the human reader.

The Bengali script has another, even more complicating feature. Other Indic writing systems (like the Devanagari alphabet used for Hindi) have it too, but not to the same extent. If several consonants occur without separating vowels, they may be written in a single conjoined glyph—perhaps rounded off with a vowel marker, placed as usual before, after, above or below the conjunct consonant. In other words, what appears to be a single character may actually be a combination of two, three or four (see Fig. 3.4). Quite often, as in some of the examples shown, the conjunct glyph bears little or no resemblance to the individual letters constituting it. A modern Bengali typecase will have two to three hundred such conjunct consonants; an old one, perhaps five or six hundred. The number is multiplied roughly tenfold by the possible vowel markers attached to them. Since the days of linotype, indeed of

Fig. 3.3 Vowel markers attached above, below, to the left and all round a consonant

ক	+	ই	=	কি
k	+	i	=	ki
ক	+	উ	=	কু
k	+	u	=	ku
ক	+	এ	=	কে
k	+	e (ey)	=	ke
ক	+	ঔ	=	কৌ
k	+	ou	=	kou

Fig. 3.4 Conjunct letters, often of irregular form

ক	+	ল	=	ক্ল				
k	+	l	=	kl				
ক	+	ত	=	ক্ৰ				
k	+	t (soft)	=	kt				
ট	+	ট	=	ট্ট				
t (hard)	+	t (hard)	=	tt				
ষ	+	ট	=	ষ্ট				
śh	+	t (hard)	=	śht				
স	+	ত	+	র	=	স্ত্র		
s	+	t (soft)	+	r	=	str		
স	+	ত	+	র	+	ঈ	=	স্ত্রী
s	+	t (soft)	+	r	+	ī (long i)	=	strī

intensive printing since the nineteenth century, there have been attempts to simplify the conjuncts, even to winkle them out altogether; but the Bengali public has demurred. Hence even a digital font has to incorporate them to win favour with users.

This makes for big challenges in both writing and reading Bengali on the computer. Where the exact constitution of a conjunct is not a factor (e.g., in printing software, whose sole end is to reproduce the glyph on the physical page) an easy shortcut is to generate an arbitrary stand-alone glyph, unrelated to its constituent letters. The conjunct is produced by a sequence of keystrokes unrelated to those for its parts.

Obviously, this will not do where textual analysis is at stake: where variant spellings of a word, for instance, might be a point at issue. There, the computer needs to recognize the conjunct as the sum of its parts. Hence the conjunct glyph has to be created by keying in the separate letters one by one, indicating whether they should be conjoined. This message (whether to join or to separate) is often pre-programmed according to general spelling trends, but there are enough exceptions to keep keyboard operators on eternal alert—and slow down their work by demanding various double-handed keystrokes, or even a sequence of two or three strokes. It takes practice even to bear the combinations in mind: Bichitra workers often kept a keystroke chart clipped to their monitors.

Writing Bengali keyboard software poses challenges of an order that people processing the Roman alphabet can scarcely imagine. Early heroic attempts include Bijoy©, Indian Scripts Input System (ISIS) and iLEAP. Bijoy, made by Mustafa Jabbar of Ananda Computers, Dhaka, Bangladesh, is still marketed commercially (Jabbar 2014). ISIS, developed by Gautam Sengupta of the University of Hyderabad, India, is available as a free download (Sengupta 2014). It was sponsored by the Government of India, as more directly was iLEAP, developed by the Government's own Centre for Development of Advanced Computing (C-DAC). The ISIS and iLEAP repertoires cover other South Asian languages as well.

These were admirable pioneering attempts, but they had their problems. There could be lapses in phonetic rendering. Conversely, phonetic consistency could be offset by unstable display of certain conjuncts and curtailed letters (special consonant forms that do not combine with vowels, like functions of *r* and the soft *t*). Also, the background presence of other programs on the hard disk could sometimes affect the display of Bengali conjuncts.

Many of these problems have been overcome in recent versions of ISIS. But the real sea change was brought about by Avro (Khan 2003), created in 2003 by Mehdi Hasan Khan of Mymensingh, Bangladesh, and subsequently developed by a team at Omicron Lab, Dhaka. Omicron Lab is still a living concern, with new versions and ancillary programs continually developed, and a lively community centred on the Omicron website.

Avro was the first Bengali phonetic keyboard layout to produce Bengali text in the Unicode format UTF-8 (Universal Character Set Transformation Format 8 bit). Unicode is an inclusive encoding system for a vast number of characters across alphabets, as well as non-character signs like punctuation marks. Whereas its predecessor ASCII, addressing the Roman alphabet, could accommodate only 128 characters, UTF-8, a version of Unicode allotting 8 bits to the byte, can encode 1,112,064 characters. In other words, it can allocate a unique code to every character in virtually every phonetic alphabet: there is no risk of a character in one language overlapping with one in another. (Chinese, Japanese and Korean increasingly use 16-bit or 32-bit versions.) Thus a Unicode-compatible textual computing program can, in theory, work with any language in the world. All Bichitra programs and operations use Unicode UTF-8. Its programs should thus be applicable to most other languages, perhaps with some modification and, of course, assuming that the language in question has basic computing support starting with keyboard software.

Avro is almost totally consistent in phonetic terms. What is more, the phonetic logic is quite transparent. You see the conjunct glyphs taking shape beneath your fingers as you key in each constituent letter; and by reversing the process, the conjunct is dismantled step by step, allowing each component to be separately identified by both the computer and the human user. The double advantage of phonetic input and Unicode output has led to Avro's popularity. The fact that Avro is open source and free to use, in terms of its Creative Commons licence, is an added advantage in economies like Bangladesh and West Bengal.

A few residual problems remain, owing to inconsistencies in Unicode codes rather than the software itself: for instance, Unicode has two ways of creating the hard *r* and *rh* sounds. This puzzled us while creating text files for Bichitra, as we could not fathom why two glyphs that looked identical yielded different results during text analysis. We finally traced the divergence to different codes used in earlier files we had used to create our own.

These were marginal issues, affecting the relatively few instances where we did not create our own text files from scratch. As of now, Avro seems the best choice of keyboard software when creating a Bengali textual database, especially one intended for linguistic and textual data mining. We adopted it without hesitation for Bichitra.

We still had a major decision to take: which Bengali font should we choose? There are very many by now, developed both in India and in Bangladesh. Understandably, our first impulse was to go for the best aesthetic effect. But a check invariably revealed that such fonts displayed some conjunct letters in a confusing or unstable way, or could not generate them at all. We finally plumped for Siyam Rupali, 'hinted' by Muhammed Tanbin Islam Siyam from the Rupali font created by Solaiman Karim (both of Bangladesh), as being most uniformly clear and stable. Above all, it is least liable to displaying vowel markers in the wrong place. Omicron Lab's Avro troubleshooting site recommends that all files showing this problem be converted to Siyam Rupali. Mozilla Firefox is specially prone to such jumbled display. Omicron Lab can fix the bug (Haque 2011).

There was another factor as well: we needed a font offering not only Bengali but Roman characters in reasonably pleasing design. For reasons explained in Chap. 5, the texts were preserved in plain-text files, which cannot accommodate more than one font. Many of Tagore's writings have English words, phrases and sentences, in Roman characters, embedded in the Bengali text. The font we chose had to include both alphabets.

We also used a number of symbols for manuscript transcription. (Again, see Chap. 5.) These, expectedly, were not part of the Siyam Rupali repertoire: we imported them from Cambria Math. Here the anomaly does not cause a problem, because (unlike Roman characters) the symbols are totally absent in the Bengali font. On opening the file through any standard browser (as the site's end user will invariably do) the browser defaults to the most appropriate font of the range installed in it—in this case, Cambria Math—whereas for Roman characters, it draws on the dominant font, Siyam Rupali. There was a second reason why we could not use another font for the Roman alphabet: it needed to be read by our internal search and collation programs.

Transcription and OCR

When transcribing English texts, at least from printed copy, an Optical Character Recognition (OCR) program vastly speeds up the process. Such a program scans and ‘reads’ the page character by character, and converts it into a text file in the desired format. This file can then be edited, searched or mined like any other text file. No OCR program is perfect: the best might achieve 98 % or 99 % reliability, which sounds good enough but would not produce an acceptably accurate text even for reading, let alone data mining or analytic processing. A speck on the paper can cause a familiar letter to be read as an obscure mathematical symbol. The risk increases with languages like French, Spanish or German that use accents, not to mention distinctive punctuation marks. There is as yet no OCR program in wide use that checks the context to select the most likely word in case of ambiguity. Hence text files produced by OCR always need checking by human eyes and hands. All the same, OCR is a tremendous boon for text-file creation.

With a language like Bengali, creating an OCR program is vastly more difficult owing to the nature of the writing system, as described above. The program has to recognize several thousand glyphs—ten vowel markers added to single consonants (some 400 glyphs already, remember?), plus innumerable conjunct glyphs of two, three or four consonants, each combinable with ten vowel markers. The look and position of the vowel markers varies specially with conjunct consonants. The crowning problem is that these conjunct glyphs can look different, often totally different, from font to font. So even if you devise an OCR program for a particular font, it must be substantially modified to read any other: you might almost have to start from scratch. If one thinks of the scores of typefaces designed since Bengali printing began in 1778 (and the many in which Tagore’s works have been printed since 1874), full OCR resources covering the range ceases to be a practicable option, even if funds could be found for the gigantic task.

Our minds initially boggled at the only alternative course: to key in by hand, on the cumbersome Bengali keyboard, roughly 125,000 pages of text, nearly 40,000 of them in manuscript. (The remaining 15,000 were in English print or typescript, where OCR could be and was used.) An army of 25–30 project staff accomplished this marathon feat in roughly a year and eight months; a few continued to fill up the gaps till the very end of our two-year timespan. The transcribed files had also to be checked against the originals, doubling the work. Luckily, the material comprised much of the best literature in the language—though that could distract the copyist’s attention. We thought up many ploys to beat off boredom and keyboard fatigue: varying the genre and subject, alternating keying-in with checking, fitting in other kinds of work when opportunity allowed. There were tea and coffee breaks to recharge one’s batteries. A bit of conversation, though officially frowned upon, was tacitly encouraged, as staving off the tedium of withdrawn silence, even the risk of drowsing off on a sweltering Kolkata afternoon when air-conditioners hardly dispelled the heat. Manuscript transcription was actually better in this respect: being more varied in content, often affording the stimulation of solving puzzles, it kept the attention keenly engaged. Amritesh describes how he would feast his eyes on the manuscript doodles to refresh his mind!

We did have the benefit of a basic set of OCR-generated text files. OCR has been developed for a few major Bengali fonts: first by Bidyut Baran Chaudhuri of the Indian Statistical Institute, Kolkata, and then his pupil Anirban Raychaudhuri during the latter's association with the Department of Computer Science at Jadavpur University. Luckily for us, Anirban's output included a near-complete set of OCR-generated files of Tagore's Bengali works in the standard collected edition. This product of an earlier Jadavpur project afforded a set of text files of one version of the works. Thus we did not need to key in the other versions in full, only make necessary changes in this 'master text'. For the English works, we created our own master text by applying OCR to the standard edition of Tagore's *English Writings* (Das 1994–2007) or to other printed texts of works not included there.

This makes our task sound much easier than it was. As any user of OCR knows—even with the vastly more amenable Roman alphabet—it's rather like that proverbial sweet, the famous *laddu* of Delhi: you pine if you can't get it, but you have indigestion if you do. All told, Anirban's text was extremely accurate, having already been manually checked; but converting it to other versions was a challenge. The 'necessary changes' sometimes meant keying in a virtually new version from scratch. Elsewhere—no less challengingly—the variants were so few and unobtrusive that to spot them called for unflagging concentration. Checking such subtly different transcripts could sometimes take longer than making them in the first place. And, of course, there was no 'master text' of works found in manuscripts alone, or in Bengali print versions other than the collected *Rachanabali*.

We therefore end with an anomaly. The world's biggest integrated literary database incorporates a writing system unsuited to the standard keyboard, and adapted to it only through extensive research and development. Adequate machine-reading support seems a distant and perhaps unfeasible prospect. The Bichitra project rooms gave concrete shape to a paradox: a computerized version of the medieval scriptorium. To make the paradox real, it had to engage in new ways with digital resources, bringing humans and computers into a new kind of contact. The story begun in this chapter continues in the rest of the book.

References

- Das, Sisir Kumar ed. 1994–2007. *The English Writings of Rabindranath Tagore*. Vols.1–3. Vol. 4 ed. Nityapriya Ghosh. New Delhi: Sahitya Akademi.
- Haque, Nippon. 2011. 'How To: Fix Bengali Display Problem in Firefox'. <http://www.omicronlab.com/blog/tips-and-tricks/bengali-firefox-problem/>. Accessed December 19, 2014.
- Khan, Mehdi Hasan. 2003. Avro Keyboard and Bangla Spell Checker. <https://www.omicronlab.com/avro-keyboard.html>. Accessed December 19 2014.
- Jabbar, Mustafa. 2014. Bijoy bangla keyboard o software. Last modified December 19 2014. <http://www.bijoyekushe.net/index.php>.
- Sengupta, Gautam. 2014. Indian Scripts Input System. <http://bangla.name/isis/>. Accessed December 19 2014.

Sukanta Chaudhuri, Purbasha Auddy,
and Kawshik Ananda Kirtaniya

The ‘Browse Collection’ menu is the pathway to the digital images of all manuscripts and printed versions of all Tagore titles in Bichitra. They add up to 139,157 pages. This chapter describes how we acquired and processed this huge number of digital images. It was obviously vital to obtain them fully and present them clearly: they are the primary sources of the texts, the foundation on which the whole site rests.

A crucial factor must be kept in mind. Except in a small fraction of cases, we were not creating our own images but obtaining them from various sources in various formats and resolutions. On the one hand, this saved us an immense deal of labour and expense: capturing nearly 140,000 images would have been a major task in itself. On the other hand, it meant we had no control over the nature and quality of the images, though we did what we could to ensure a minimal standard. (Rabindra-Bhavana, for instance, insisted on providing images of their manuscripts only in 100-DPI resolution.) It also meant we had to standardize the file format and byte size—often a complicated task, sometimes requiring more than one step. In a word, we had to negotiate a path full of challenges.

Gathering the Material: Sources and Formats

The images were obtained from three types of sources: institutions, individuals and web resources. More trickily, they were obtained in various formats like JPEG, TIFF and PDF: we had to decide which format to use when uploading images to the website.

S. Chaudhuri (✉)

Department of English, Jadavpur University, Kolkata, India
e-mail: schaudhuri@english.jdvu.ac.in

P. Auddy • K.A. Kirtaniya

School of Cultural Texts and Records, Jadavpur University, Kolkata, India
e-mail: pauddy@gmail.com; keuekjon@gmail.com

© Springer International Publishing Switzerland 2015

S. Chaudhuri (ed.), *Bichitra: The Making of an Online Tagore Variorum*,
Quantitative Methods in the Humanities and Social Sciences,
DOI 10.1007/978-3-319-23678-0_4

The most important source was [Rabindra-Bhavana](#), the Tagore library and archive at Santiniketan. Their manuscripts had already been digitized: a full set of copies was made over to us at the very start of our work. With the printed material (both books and journals), the situation was more fluid. These were being digitized at the same time as the work on Bichitra, under a project sponsored by the Raja Rammohan Roy Library Foundation, the Government of India's apex body for library and archiving services. The digitizing was done by another state agency, the Centre for Development of Advanced Computing ([C-DAC](#)). We at Jadavpur were not a party to this project; but as the material was scanned, a copy would be made over to us.

Visva-Bharati felt they could not supply high-resolution images of the manuscripts, for fear of unauthorized downloading and publication. They provided 100-DPI images in coloured JPEG format. But our first use of the images was to transcribe them as text files, and such low-resolution images (especially of manuscripts) were not always clearly decipherable. Enlarging them could make matters worse, as the images would pixelate and become totally unreadable. We did what we could by using Microsoft Office® Picture Manager to adjust brightness and contrast. We also installed a virtual magnifying glass to zoom in on a small section of the manuscript. The only advantage—a somewhat doubtful one—of the low-resolution images was that they were small in byte size.

Meanwhile, C-DAC began supplying images of the printed material. At first, we collected them on pleasant trips to the idyllic Santiniketan campus; later in a more businesslike way from the C-DAC lab in Kolkata. We warmly recall the courtesy we received from the C-DAC officials on these visits: Mina Desai, Utpal Saha, Sanjeev Kumar. C-DAC provided TIFF images in two formats—original TIFF (OTIFF) and processed TIFF (PTIFF). We started with the PTIFF images, but found they contained missing pages, unclear images and illegible portions in the gutter area. By contrast, OTIFF images were generally more legible; moreover, they retained comments and other markings in both gutter and text block areas. These were eliminated in PTIFF because C-DAC applied its image-enhancement procedures in large batches: hence any image which needed special attention was ignored (see [Fig. 4.1](#)).

We were using OTIFF images for transcription anyway; we further decided to use them as the basis for the images we would upload. There was no question of uploading the TIFF images themselves. Apart from the large file size, most browsers like Google Chrome, Mozilla Firefox, Opera and Internet Explorer do not support the TIFF format. Had we persisted with this format, a TIFF image viewer would have had to be separately installed in the Bichitra website.

From quite early on, we had to open another front in our hunt for material. Despite the unique richness of the Rabindra-Bhavana collection, it did not contain everything. The most significant gaps related to rare journals. This was a crucial sector, as a good proportion of Tagore's writings first appeared in journals. We had to look elsewhere for a good deal of this material.

Three online databases proved invaluable here. The first was purveyed by the West Bengal Public Library Network (WBPLN [2014](#)). This program uses [DSpace](#) architecture as the repository and access provider of their digital resources, while

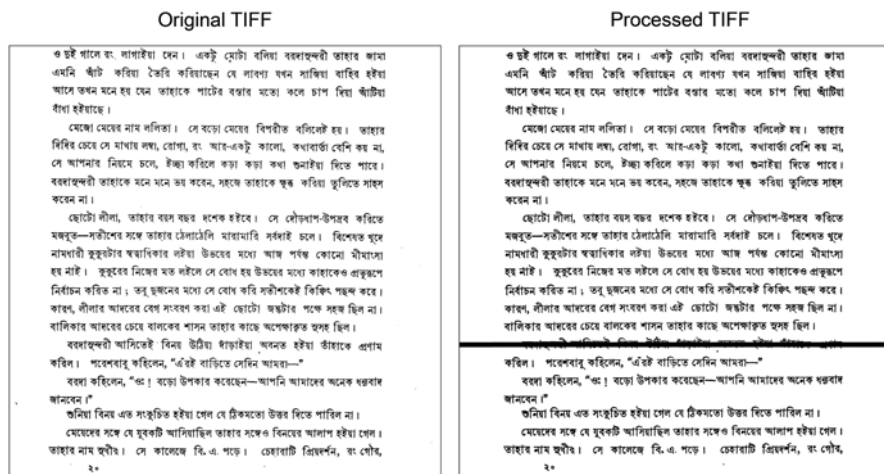


Fig. 4.1 The difference between OTIFF (*left*) and PTIFF (*right*)

the pages are stored in PDF format. Helpfully for us, the thick physical volume containing each year's run of a journal was usually divided into smaller PDF files. Having located the reference through our bibliographical research, we only needed to download the small file with the relevant issue, then extract the few pages we needed using Adobe Acrobat© Professional 7.0.

The second database was set up by the South Asia Institute ([Savifa 2014](#)), Heidelberg University. This collaborative effort of the Centre for Studies in Social Sciences, Calcutta ([CSSSC](#)) and Savifa contained many Bengali journals in the former's [Hitesranjan Sanyal Memorial Collection](#). A problem was that, unlike the WBPLN site, Savifa stored each thick yearly volume in a single large PDF file. We could only download them early in the morning, when there was less demand on the Internet bandwidth. Having downloaded a file, we then had to sift through the whole issue to locate a particular item. We also visited CSSSC in person for material not yet placed on the web but available offline. We are especially grateful to them (in particular their archivist [Abhijit Bhattacharya](#)) for supplying us with a PDF file of a special issue of the journal *Bharati* edited by Tagore himself.

Finally, we used the database of the Digital Library of India ([DLI 2014](#)), hosted by the Indian Institute of Science, Bengaluru. DLI is a consortium that tries to bring together significant Indian cultural and historical material on a single platform. Several scanning centres across India have contributed to the effort. Among them was C-DAC, Kolkata, which scanned Bengali books and periodicals from old public libraries like the Uttarpara Jaykrishna Public Library and the North Bengal State Library at Koch Behar, among others. To view the images on this website, one needs to install a plug-in: for Windows (the operating system we chiefly used) this was [AlternaTIFF](#).

An advantage with DLI was that we did not have to download an entire file. The few necessary pages could be downloaded one by one. But this ceased to be an

advantage when downloading a large work: a book of 100 pages meant hitting the ‘Save’ button 100 times, then gathering the single images in a folder. Moreover, DLI only provided PTIFF files. To add to our woes, our AlternaTIFF image viewer did not work properly at the outset: we could not proceed straightaway to the desired page number, but had to click through the pages to reach the one we wanted.

By this point—about halfway through our 2-year schedule—we had acquired digital image files of several different types from different sources:

- The manuscripts of Tagore in JPEG format.
- Books from Rabindra-Bhavana scanned by C-DAC in OTIFF and PTIFF format.
- PDF files from the WBPLN and Savifa websites, downloaded in two different ways.
- PTIFF files scanned by C-DAC from the DLI website.

About 70 % of the material was in TIFF format. We had already started planning the website structure. We needed to decide which file format to use for uploading images to the site.

No less urgently, we had to take stock of what material we had collected and what remained. There was a substantial deficit, some of it not available in Rabindra-Bhavana. Some, sadly, was available there, but not provided to us owing to a last-minute failure of communication described in Chap. 1. We had to enter the field ourselves to fill the sizeable gaps.

Digitizing books, even rare and fragile books, was not a problem for the Bichitra staff. They included veterans who had (among much else) already completed three projects digitizing printed material under the British Library’s Endangered Archives Programme. The School boasted a range of scanners and cameras with custom-made lighting and other ancillaries, devised by Kolkata’s endlessly ingenious craftsmen and technicians (see Fig. 4.2). The bounty of the Bichitra project funded an advanced cradle scanner, the *Atiz* BookDrive Pro© archiver, and a superior Nikon D800E SLR camera. The problem, thus, was neither the equipment nor the skill to handle it; it was time management, as staff had to be diverted from the major tasks of text transcription, file management, and software writing. Once the images were captured, they had to be processed and organized: this ‘post-processing’ could take up much more time than the actual scanning or photography. The in-house digitization, which we had not bargained for on this scale, truly placed a strain on our human resources.

We had to proceed with it nonetheless. The first task was to locate the material. By combing the work spreadsheets (see Chap. 6), we listed the items of which we had no scan, and looked for them in catalogues of major libraries like the National Library, Kolkata, the Calcutta University Library, the Bangiya Sahitya Parishat (Bengali Academy of Letters), and of course the Central Library of Jadavpur University itself. A few private collectors also lent items for scanning. The National Library helpfully scanned its own material and supplied us with copies. For the rest, items that could be brought to our premises were scanned on the *Atiz* archiver; the



Fig. 4.2 (a) Atiz BookDrive Pro© archiver scanner. (b) Locally made camera stand and lights

others in situ using the Nikon digital SLR camera. Both methods allowed old and fragile material to be scanned without any contact or pressure, with the book open at a gentle angle to protect the binding. The Atiz BookDrive Pro© archiver can scan the two angled pages of an open book using two separate cameras, and combine the results in a single flat image (though we preferred separate images of the two pages).

It can capture images in RAW and JPEG format. As RAW files needed further processing, we opted for high-quality JPEG files.

We must not forget to mention a major source of material from abroad. Tagore sent early drafts of much of his early English poetic translations to his friend the artist William Rothenstein. Rothenstein's vast collection of papers has come to rest in the Houghton Library of Harvard University. It included a valuable cache of Tagore manuscripts, besides letters and other material. One or two manuscripts, like that of the English *Gitanjali*, had already been rendered in facsimile; but for the most part, these manuscripts were now offered for the first time to scholars who could not personally visit Harvard. Sukanta sealed the deal during a trip to America. He has grateful memories of the welcome he received from Robert Darnton, Sugata Bose and William Stoneman, and their offer to digitize the material on a priority basis. They provided images in both TIFF and high-resolution (600 DPI) JPEG format. We used the first for transcription and the latter when preparing files for uploading.

On the subject of help from abroad, we should also record the unexpected difficulty in finding a copy of the first 1913 Macmillan imprint of the English *Gitanjali*. (A later reprint of the same year was available, but we insisted on using first impressions of all works.) A copy was located in the Senate House library, University of London. Dibyajyoti was briefly in London on a training tour. He photographed the book using a point-and-shoot camera and mailed home the JPEG images.

A last surprise item arrived during the final phase of work. Samantak had unearthed eight loose sheets in Tagore's handwriting in an old file among his late grandfather's papers. They contained two poems that eventually became songs, and a new text of a short story. These pages marked the last reprographic activity of the Bichitra team. But we will gratefully accept more such offerings, digitize and upload them to the website, and return the originals in good condition to the owner, along with a set of digital copies by way of saying 'Thank you'.

Processing the Acquisitions: File Formats and Conversions

Having obtained the images, we needed to process them. In particular, images made with a camera had to be cropped, either manually or through a batch process using suitable software. This is because the camera scans both facing pages of an open book laid flat under the lens: these have then to be made into two separate images. No cropping is necessary with the Atiz BookDrive Pro© archiver, as it has an option to capture the two facing pages separately. We also needed to adjust the brightness and contrast in a lot of images, as they had been captured in low light in various libraries.

We have said how the images we obtained were in various file formats and byte sizes: high and low quality JPEG, compressed original and processed TIFF, uncompressed TIFF, and PDF. Standard archival practice recommends the uncompressed TIFF format as ideal for storage, as it includes maximum information. This is how

we produced and archived master images for the British Library's Endangered Archives Programme. The size of an uncompressed TIFF file can run to 50 MB, but it can be compressed without loss to just 60 kB. This is the way C-DAC stores their images, the final output being processed TIFF (PTIFF) files. The Digital Library of India also uses PTIFF for sharing images online, but as explained earlier, the user has to install a TIFF image viewer.

Seventy percent of images in the Bichitra image archive were in PTIFF format. If we were to adopt this for the website, we would need to convert other formats like JPEG and PDF to PTIFF. More crucially, users would have to install an additional plug-in, which we wished to avoid. These factors made us decide against PTIFF. The next option was PDF; but this would involve the user's downloading large files containing an entire long work to access even a single page—a tedious or, with limited bandwidth, impossible operation. We needed a file format that allowed access to one image at a time.

We next considered JPEG. This format can accommodate millions of colours and is excellent for photographs and for online exchange. But it undergoes considerable loss of output when compressed to the size required for the website, which must be 250 kB–280 kB at most. With such images, users would not be able to zoom in as much as they wished. Hence reading the text within the image might prove impossible, especially in the case of manuscripts.

Clearly, we needed something completely different. After several meetings, we decided to try Graphics Interchange Format (GIF). This format supports fewer colours, but images of texts do not call for much colour information. It can support 8 bits per pixel, which was enough for our purpose. Experiments showed that it produced lossless results, even for images of manuscripts. We therefore decided on black-and-white or greyscale GIF images for printed texts, and coloured GIF images for manuscripts (see Fig. 4.3 for a flow chart of image collection and conversion in various formats).

We were relieved to have found the file format we needed. An image viewer was built into the website that enabled users to see the images one page at a time (see Chap. 9). We now faced the task of converting 139,157 images in JPEG, TIFF and PDF format into GIF.

These images were contained in 4437 folders comprising books, periodicals and manuscripts. The enormous task of conversion was carried out with various softwares like Adobe Acrobat© 7.0 Professional, Atiz BookDrive Pro© Editor 5.0 and IrfanView, through automated batch processing without much manual intervention. With PDF files, there was no option for saving the images directly in GIF, so the individual pages were first separated into TIFF images using Adobe Acrobat© 7.0 Professional. The Atiz BookDrive Pro© Editor 5.0 (provided with the scanner) came with some very useful software that greatly aided batch processing. Books scanned in-house or photographed in libraries were edited—cropped, resized, their background removed, brightness and contrast adjusted—very quickly by means of this software. It also enabled us to change the colour depth of the black-and-white images as well as some greyscale images, enhancing their legibility; but the

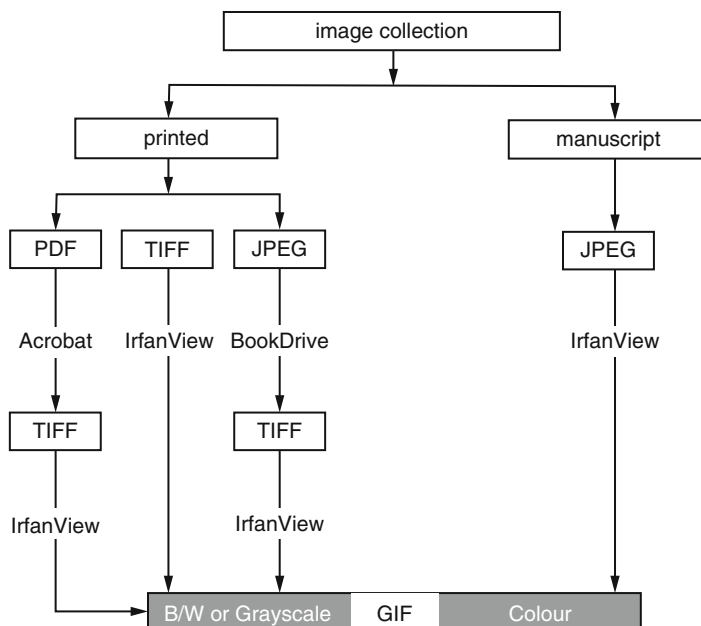


Fig. 4.3 Image collection and conversion: flow chart showing formats and conversion software

resulting images were only available in JPEG and TIFF format. In the rare cases where the Atiz BookDrive Pro© Editor 5.0 failed to give good results (usually with images of old moth-eaten pages), Adobe Photoshop© 7.0 Professional was used instead. All images of printed material were converted into black-and-white TIFF format as a first step. Images of manuscripts, obtained from Rabindra-Bhavana and Harvard in coloured JPEG format, were not touched, as any editing might have impaired the registration of handwriting, additions, deletions and other marks.

We had chosen the file format, resolved the question of byte size, and carried out the required editing. We now needed software to convert the images into GIF format, and rename all the image files with an eight digit file number (see Chap. 6 for details). Obviously, we had to resort to batch processing. The above-mentioned programs could not meet this requirement. We had recourse instead to the freeware [IrfanView](#). With its help, we could convert all the images to GIF and rename them simultaneously.

As this account shows, processing the vast and varied body of images was an immense challenge, calling for decisions at every stage. But once we had solved the problems, we were able to build up an image archive notable not only for its size but also, we hope, for its technical quality, given that we had to deal with material provided by other parties in heterogeneous formats—above all, manuscript images of only 100 DPI.

Accessing the images is simple. Using the drop-down menu under ‘Browse Collection’ (see Fig. 4.4), you choose the language and genre, then opt for either



Fig. 4.4 Merged screenshot showing opening stages of drop-down menu and Alphabetical Index

the ‘Alphabetical Index’ or the ‘Full Table’. The latter affords the better means of accessing the images. Each cell in the table records a particular version of the text in manuscript or print. You click an icon in the cell to open the first page. You then click the ‘Toolbar’ button at the top right to open the toolbar, and navigate by either using the left and right arrows or typing in the image number in the appropriate box.

We admit to a problem in navigation. You cannot go directly to a particular item in a book or manuscript: you must scroll through the pages. With images of a printed book, you can look up the Table of Contents near the start or the Index, if any, at the end, and try to open a page at approximately the right point. But even this is a hit-or-miss procedure—the more so as one can only call up a particular image, not a particular page number in the original book. Also, the images of the manuscript pages do not scroll synchronously with the transcript: one has to scroll separately through the latter to arrive at the correct point.

A tip: To reach page 175 (possibly image 185 or thereabouts), first open page 5, then 85, then 185 in the page-search window on the toolbar. You thereby save yourself 182 clicks. Finally, use the arrows to scroll to the exact page.

For the time being, we can only ask the user to bear with us for these shortcomings. We look forward to a chance to set them right.

References

- DLI. 2014. The Digital Library of India. <http://www.dli.ernet.in/>. Accessed December 19, 2014.
- Savifa. 2014. 'Bengali Periodicals and Newspapers'. <http://www.savifa.uni-hd.de/thematicportals/periodicals/overview.html>. Accessed December 19, 2014.
- WBPLN. 2014. West Bengal Public Libraries Network. Directorate of Library Services, West Bengal. <http://www.wbpublibnet.gov.in/>. Accessed December 19, 2014.

Sukanta Chaudhuri, Smita Khator, Sahajiya Nath,
Amritesh Biswas, and Aparupa Ghosh

The Manuscript Material

The term ‘manuscript’ is not so easy to define as seems at first sight. Strictly speaking, it can refer only to a handwritten document, but is customarily used to include typescripts and computer printouts. A number of Tagore ‘manuscripts’ (chiefly English but a few Bengali) in the Bichitra archive are in fact typescripts, sometimes with handwritten revisions and insertions.

There are other hidden complexities. Even handwritten manuscripts, to use a tautology, are of many different types that blend into each other, reflecting every stage of composition from an author’s first draft to the final fair copy. They can be in the author’s handwriting (holograph) or someone else’s (scribal copy), or a combination of the two. Paradoxically, the coming of print has created a new category of manuscripts: press copies, sometimes with actual signs of presswork. There is also a curious but not uncommon class of manuscripts copied from printed material.

Even this is putting it too simply. It is not always easy to identify an author’s hand. Tagore’s handwriting varies strikingly across tens of thousands of pages, according to date and circumstance (hurried first draft, thoughtful revision or fair copy) as also the writing materials used. Many manuscripts were copied by

S. Chaudhuri (✉)

Department of English, Jadavpur University, Kolkata, India

e-mail: schaudhuri@english.jdvu.ac.in

S. Khator • S. Nath • A. Biswas

School of Cultural Texts and Records, Jadavpur University, Kolkata, India

e-mail: smita.khator@gmail.com; sahaj.hiya@gmail.com; tataiputu@gmail.com

A. Ghosh

Department of Bengali, Jadavpur University, Kolkata, India

e-mail: mail.aparupa.ghosh@gmail.com

members of his circle like Amiya Chakrabarti, Sudhindranath Datta, Satyendranath Datta, Rani Chanda and Priyambada Debi. Sometimes their own compositions have mingled with Tagore's: these had to be spotted and excluded. The copyists' spelling and punctuation practices may have coloured their transcriptions of Tagore. In an opposite category is a famous manuscript, the *Paribarik smritilipi pustak* (*Family Memorial Volume*, MS RB 272) collectively created by Tagore's family when he was young. Only a few pages of this precious document were composed and inscribed by the poet. In view of the importance of the manuscript, Bichitra provides images of all the pages, but a transcription only of the poet's own contribution.

It seems fair to assume that a heavily revised page is authorial: no one else would venture to change the readings. We cannot be so sure with a neat fair copy. Some of Tagore's followers deliberately cultivated 'Gurudev's' hand, so what appears to be a holograph may well be a scribal copy. Tagore routinely left it to his entourage to make fair copies for the press. Hence minor variants in a fair copy might have been introduced by the scribe or amanuensis, in error or by way of well-meant correction.

It was not our task to identify the handwriting of a manuscript, except if it had a bearing on the authorship (see the section on 'The English Corpus' in Chap. 6, on the problems of tracing the translators of some English writings). Any manuscript of Tagore's known works, preserved in the 'home' archive of his own university, was grist to our mill irrespective of the scribe. Nor was it our task to track the occasion or motive behind every change in the text: we were simply presenting the material from which scholars could make such deductions.

We were working not with the actual manuscripts but their electronic images. Rabindra-Bhavana insisted on supplying low-resolution 100-DPI images to prevent unauthorized publication after downloading. (The Harvard images are at 300 DPI.) It is hard to decipher 100-DPI images for transcription, even after on-screen enlargement. It is even harder to detect variations in colour and density of ink that might indicate the sequence of changes or layers of revision (see Fig. 5.1). We could do so only tentatively, where the sequence of deletions and revisions was absolutely clear. Otherwise, we treated all revisions on a par.

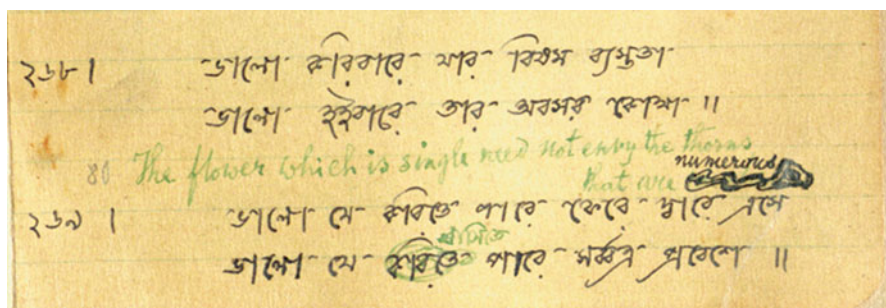


Fig. 5.1 Section of manuscript page showing entries made at different times, including some in faint green ink. Bengali epigrams with English translation inserted later in the gaps (MS RB 008 page 81; Bichitra image 47)

Proper genetic criticism (i.e., tracking the stages of composition) calls for physical access to the manuscripts. Not one reader in a thousand carries out such study, and even that one scholar can make a start with the Bichitra images. For the rest, the Bichitra manuscript archive opens up a realm they would otherwise never have entered.

To map this territory of 47,520 pages was a stupendous challenge. Every manuscript is, by definition, unique: not only in the text it offers but, more often than not, in the way that text is laid out. At one end of the spectrum we have much-scribbled, worked-over drafts that scholars of Elizabethan drama aptly called ‘foul papers’; at the other, fair copies intended for the press. But another range of differences relates to the overall contents and structure of a manuscript.

In many cases, a manuscript book is not really such, but a collection of stray sheets of different nature and origin, bound together (not always correctly or systematically) at a later date. An integrated manuscript book might consist of a quire of loose sheets, bound or sewn by hand. In Tagore’s case, such manuscript books were often of foolscap size. He sometimes divided a foolscap page into two vertical column spaces by a fold down the middle. He would write down one side or column, leaving the other free for revisions and comments. Elsewhere, he uses facing pages in the same way. Many other manuscripts consist of commercially sold exercise books or printed diaries (see Fig. 5.2). There is, of course, no question of dating the manuscript entries by the printed dates in a diary. We can only assume a *terminus ab quo*: they must be later than 1 January of the year in question, perhaps even a little earlier if the diary was bought in advance. Even this holds good only if the version in question is a first draft.

But the really crucial point relates to the distribution of items within the manuscript. Life is easy if it contains a single long work like a novel or play, or an ordered series of short poems or essays. But very many manuscripts are totally haphazard in arrangement. First of all, their contents might relate to components of several different published volumes. Second, they might be distributed in a confusing way. A single page might contain three poems in whole or part (see Fig. 5.3). If in part, one or more other parts might surface some pages later. In several cases, one work—say a novel—might be written on the right-hand pages, starting at the front, and another work—say a play, or more confusingly a set of poems—on the left-hand pages, perhaps upside down or in reverse order starting at the back (see Fig. 5.4). Revisions to a page might be entered on its verso (reverse: see Fig. 5.5), perhaps along with other revisions to the page facing the verso! There were also times when, for whatever reason, Tagore economized on paper by cramming the margins with additions and revisions whose sequence is not always apparent (see Fig. 5.6). Such inconclusive locations have been marked in the transcript by the sign \pm . The standard printed version—if there was one—might give a clue, but that is to assume that the poet’s intent in the manuscript coincided with his final plan.

Sometimes, excitingly, these features were numerous and intensive enough to represent a radical revision: we could see one text grow into another. There are several variant manuscript versions of many epigrams in the Bengali *Sphulinga* (*Sparks*), the English *Stray Birds* and *Fireflies*, and the bilingual *Lekhan* (*Writing*).



Fig. 5.3 Multiple entries on a manuscript page. There is a cancelled poem spread across both pages, and a second poem crammed into the right margin of the left-hand page, with one line at the top of the right-hand page. (MS RB 159 pages 492–3; Bichitra image 250)

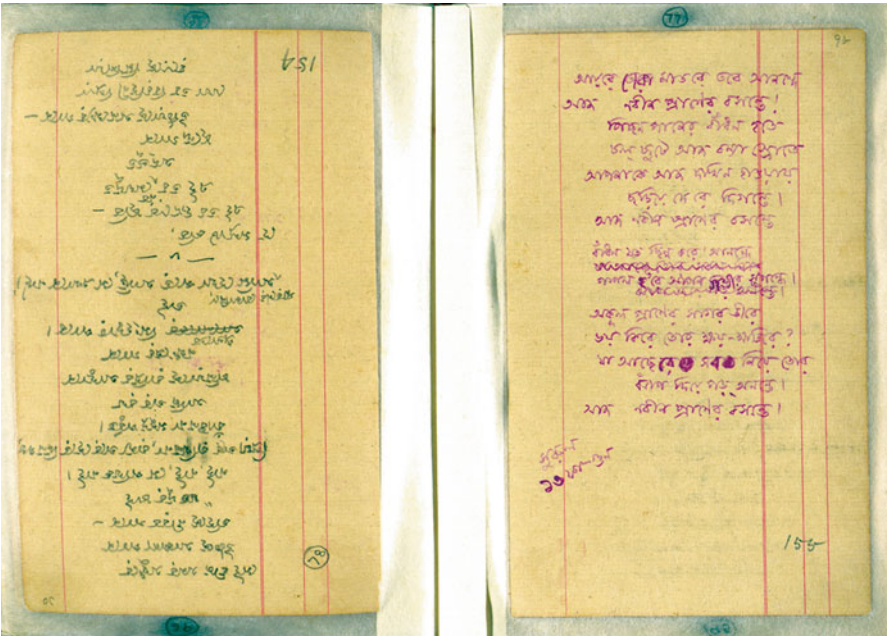


Fig. 5.4 Manuscript with different sequences of poems on left-hand (upside down) and right-hand pages in different inks (MS RB 131 pages 96, 77; Bichitra image 82)

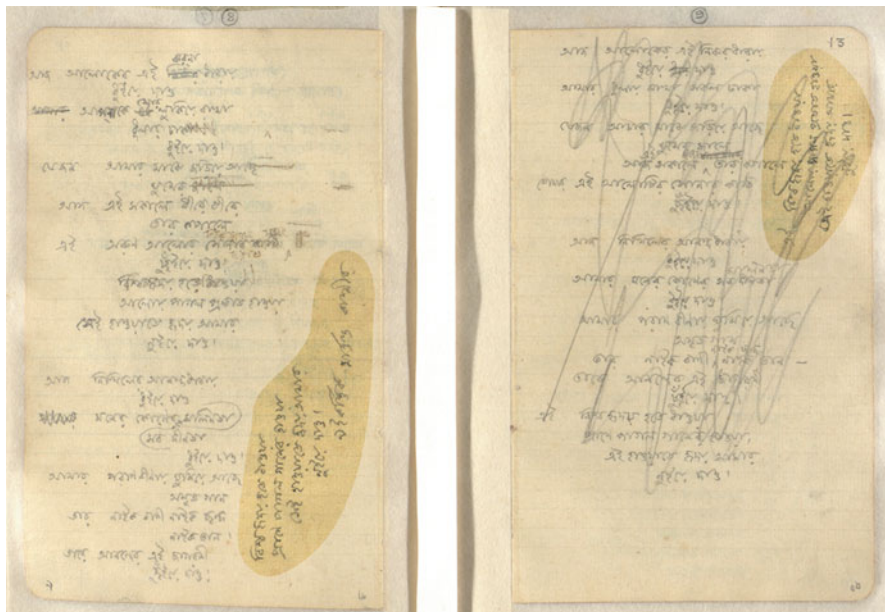


Fig. 5.5 Revised text entered on opposite page (corresponding sections highlighted by us) (MS RB 111 pages 8–9; Bichitra image 17)

More dramatically (pun intended) MS RB 230 shows the play *Achalayatan* growing into its recast version *Guru* (see Fig. 5.7), and MS RB 84, a new manuscript text for the play *Paritrān* (*Salvation*) generated by revising and adding to the printed pages of its earlier avatar, *Prayashchitta* (*Penance*) (see Fig. 5.8). Or as in Fig. 5.9, a page left intact in the revised version might be inserted from the earlier printed text (here a proof copy) among handwritten pages with radical revision.

The Rationale of Manuscript Transcription

A manuscript, especially an authorial draft, is a multi-layered thing. It is created in instalments, as the author thinks and re-thinks the material and revises his original text. Other kinds of addition may follow, such as instructions for printing, or for staging a play text (see Fig. 5.10). The total process might be completed in an hour, or it might stretch over months and years with fallow periods in between. The manuscript is a static witness to this process in time.

Its transcription, on the other hand, is created entirely in one go: all its phased components are reduced to the same level, even literally by eliminating differences in ink, handwriting or layout. To put it another way, a manuscript is (even metaphorically) a three-dimensional object: one can probe beneath its surface to reach layer after layer of composition. The transcription is, so to speak, its two-dimensional

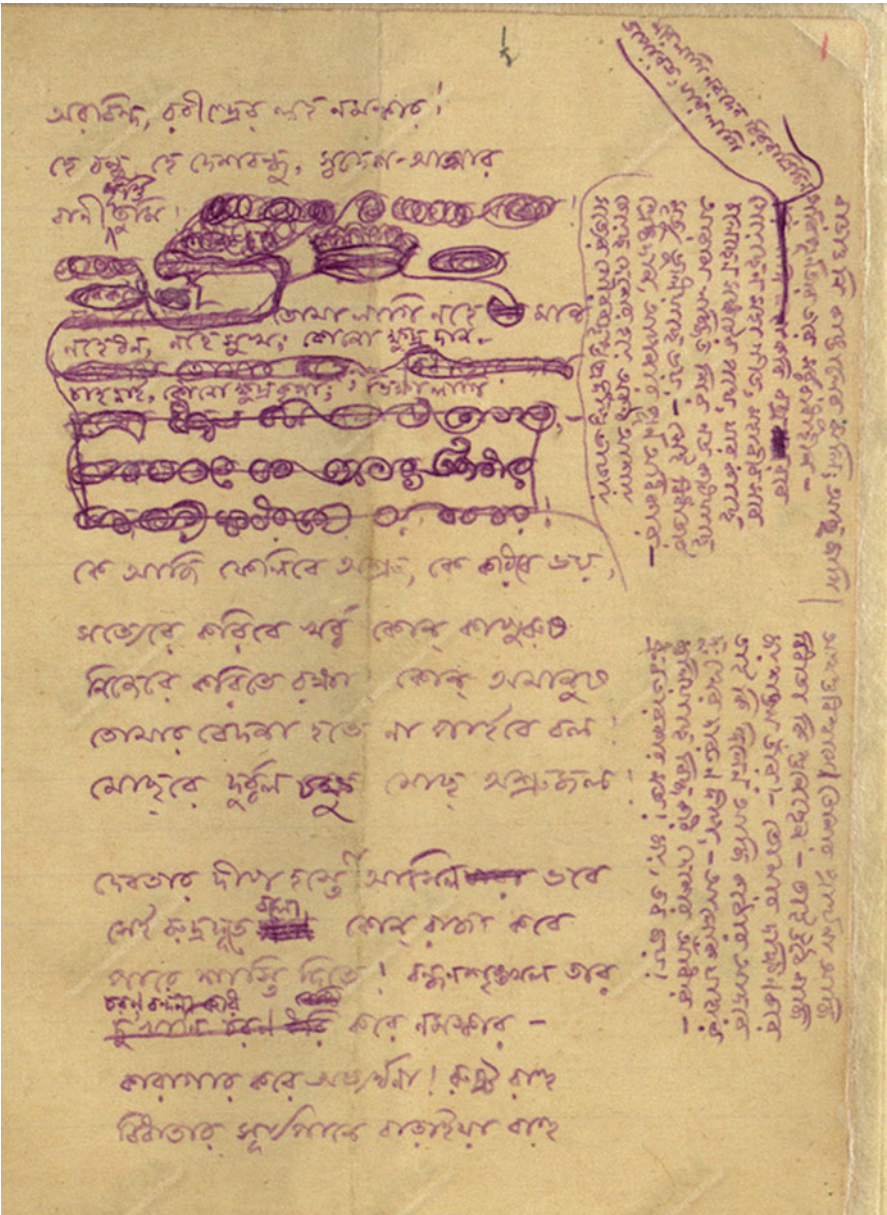


Fig. 5.6 Extensive marginal additions of indeterminate position in the poem (MS RB 351 page 1; Bichitra image 2)

representation, as a painting represents three-dimensional objects on a two-dimensional surface. In order to do this, the painter resorts to certain strategies, above all the use of perspective. What could we do to bring perspective to our rendering of the original manuscript, to present its many layers and facets?

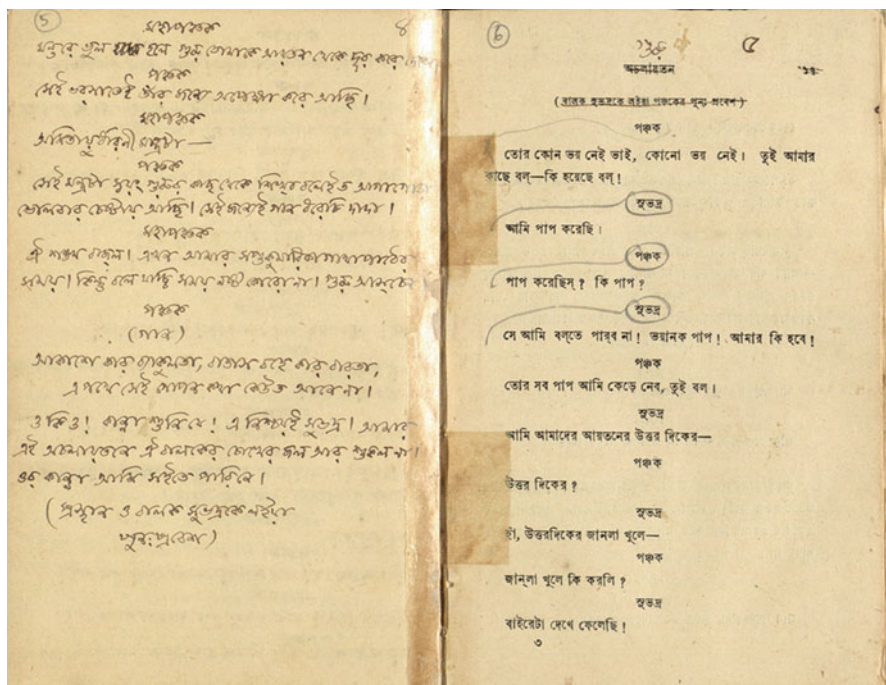


Fig. 5.7 Printed text of *Achalayatan* being converted to its revised version *Guru* by manuscript additions on the facing page. The original printed text has been pasted on the right-hand page of the manuscript book. (MS RB 230 pages 5–6; Bichitra image 5)

There is another theoretical angle to the rationale or, to use a grander word, philosophy of manuscript transcription. Each manuscript is unique, however closely it may resemble another, in a way printed books are not (though there can be variants between copies of the same imprint). This has a practical consequence in that except for a few routine copies (chiefly typescripts), the 711 manuscripts we processed each presented some unique problems. The transcribers had constantly to consult each other and the project leader on how to render some novel feature at a particular point. This sometimes happened with printed texts as well, but there we could assume a basic level of uniformity across the corpus.

This meant that at ground level, we had to make different rules for transcribing manuscript and print. For printed works, misprints in the original were deliberately preserved. (If you come across an apparent error in transcription, do check the original—and tell us if the mistake is on our part!) In manuscripts, however, some ‘errors’ are endemic, and obviously owing to idiosyncrasies on Tagore’s part or his copyists’. To take the plainest example, the Bengali letters for *b* and *r* differ only in a dot below the latter character; so do *y* and a form of *j*. When writing fast or carelessly, Tagore often left out the dot. The intended word is hardly ever in doubt.

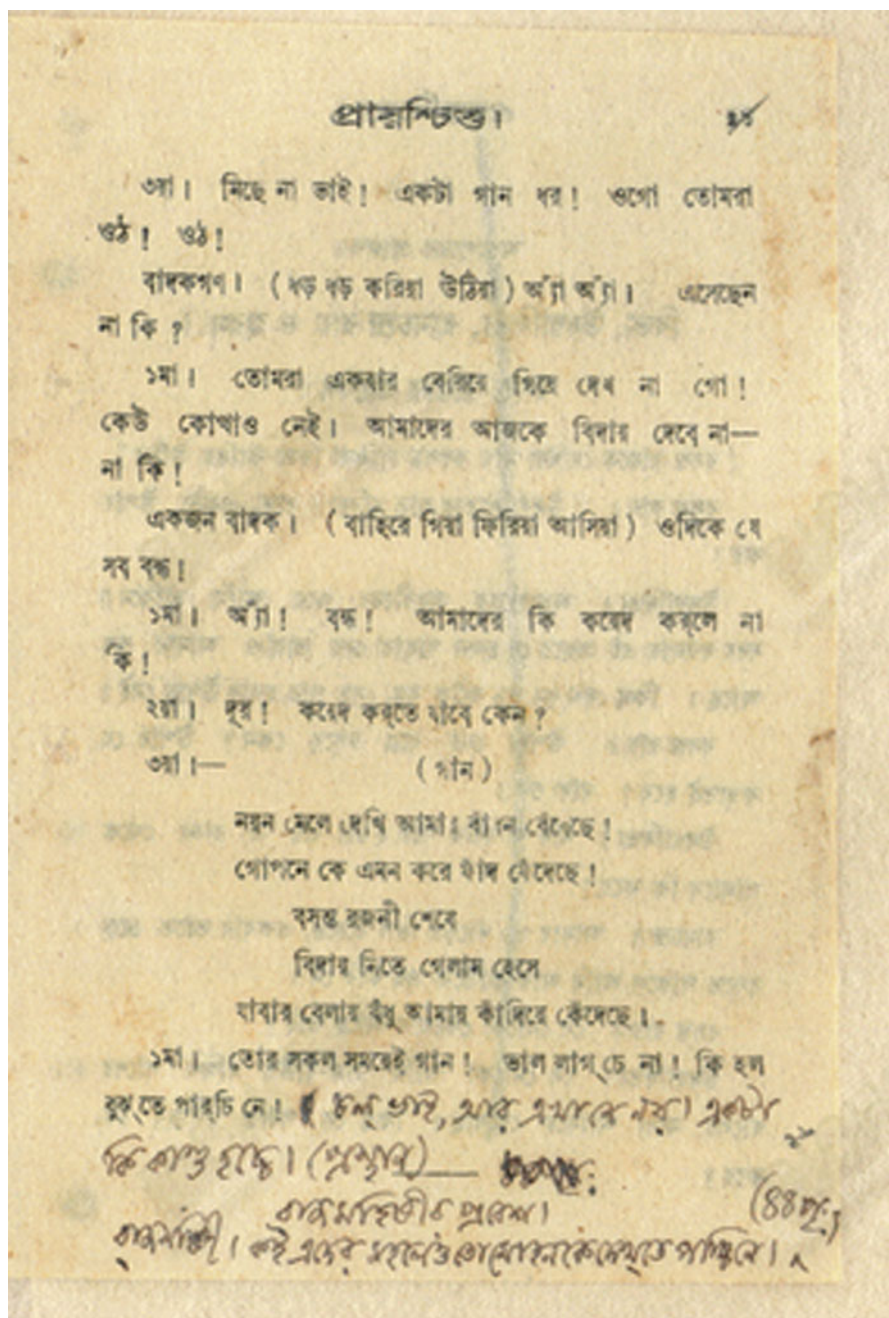


Fig. 5.8 Printed text of *Prayashchitta* being converted to its revised version *Paritrān* by manuscript additions (MS RB 084 page 31; Bichitra image 30)

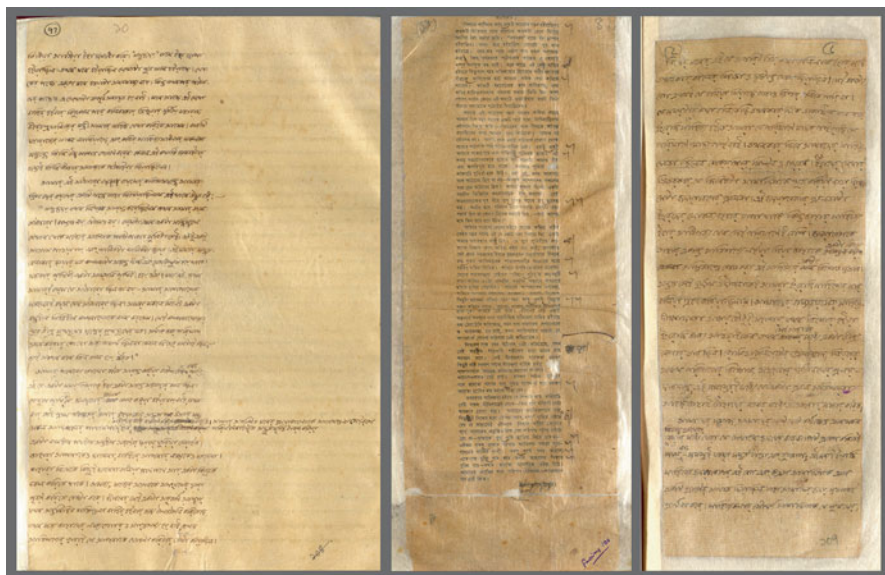


Fig. 5.9 A printed page, where the text remains the same, pasted among manuscript pages drastically revising the earlier text (MS RB 146(iii) pages 97, 99, 100; Bichitra images 58, 59, 60)

It would simply clog the reader's progress if, in such cases, every *r* were rendered as *b*. On the same principle, we also silently corrected obvious cases of wrong or missing pen strokes, even the occasional instance of a wrong conjunct. To take a trickier case, there are two *n*'s in Bengali, and Tagore habitually formed the last conjunct in *chinha*, 'sign', with the wrong *n*. We let this stand. Of course we did not make any change where there could be any ambiguity in the reading. In such a case, all possibilities were left open.

There were tricky problems that called for great attention. Tagore often wrote the vowel markers for short *i* and long *i* in confusingly similar ways. From mid-life, he also backed a spelling reform movement to consistently replace long *i*'s by short *i*'s in Sanskritic words. Many people would want to know which system he followed in which work, at what stage of his life. His holograph manuscripts obviously yield the best evidence, but it is not always easy to interpret.

Transcribing a manuscript, above all a draft in the author's own hand, involves close engagement with his personality and work habits as reflected in the text. This makes for a greater personal entry on the transcriber's part as well: it is a meeting of two minds, on however different terms or levels. The project staff who worked on these manuscripts for 18 months and more—covering their entire range, probing every nook and cranny—had the rare opportunity to explore the mind of an all-time literary genius.

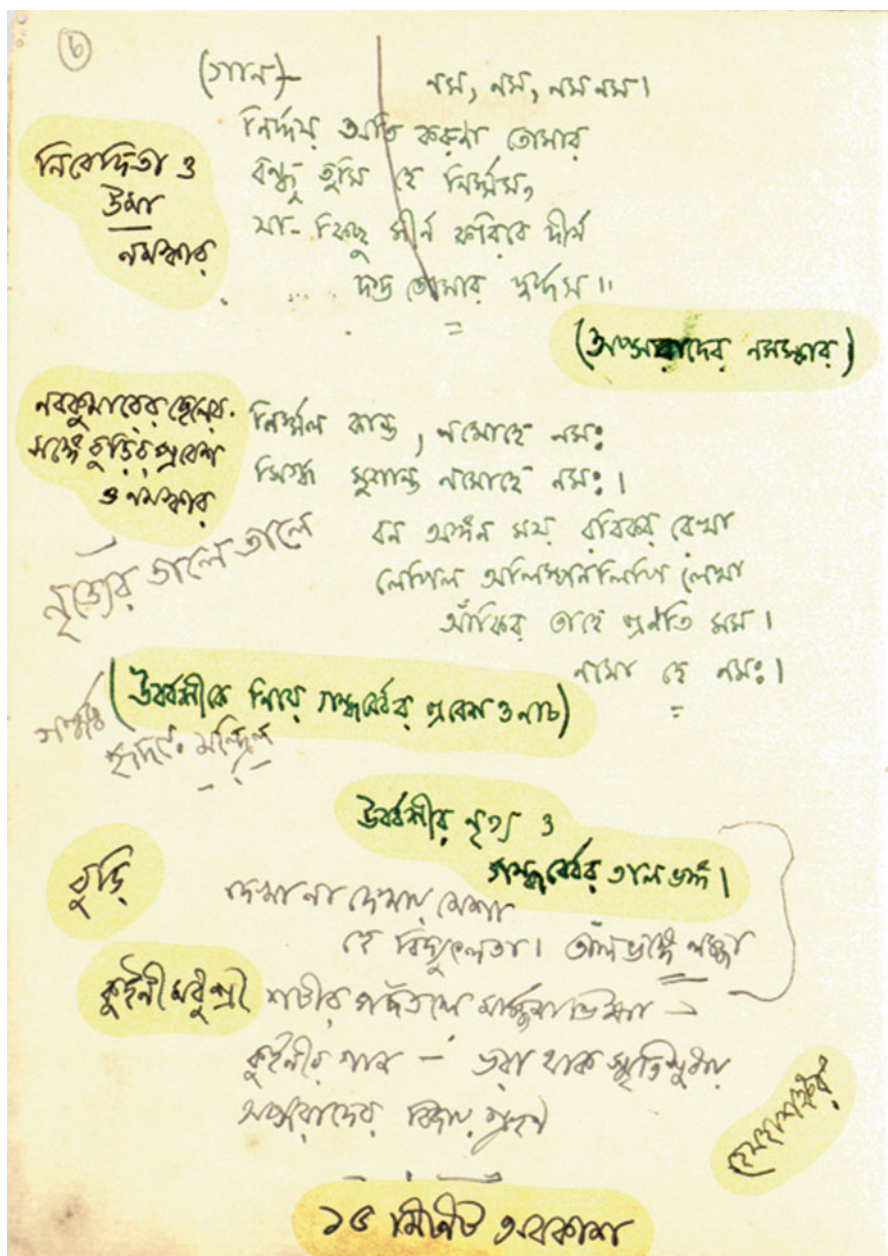


Fig. 5.10 Manuscript of the play *Shapmochan* with instructions for performance (highlighted by us) in margin (MS RB 252 page 6; Bichitra image 4)

The Task of Transcription

Every time we took up a complex manuscript, we had to spend some time—maybe half a day or more—simply to get its bearings, like an animal sniffing out a new territory. This was done to locate the items it contained, separate the different ones and join up the scattered parts of each. To identify them would be a later and more challenging task.

Rabindra-Bhavana provided us with a full set of manuscript images at the very start of the project. But given their complexities as detailed above, we waited a few months to tackle them, while transcribers trained themselves on printed material. On a memorable day some six months into the project, two of our most skilled colleagues, Rohan Islam and Sahajiya, made a start with MS RB 159. We had deliberately chosen a large and complex volume containing poems from no less than ten verse collections, parts of two plays and a number of songs. Yet more crucially, the texts were often rough drafts with confusing deletions, insertions and other revision (see Fig. 5.11). There were also many doodles, and pages full of random jottings. It took our two colleagues a month's hard work to master the contents and make the transcription.

If even a quarter of the manuscript pages were as complex (some proved to be more), it would have taken two operators 40 months to transcribe them at this rate. But the month had been well spent. By plunging in at the deep end, we had

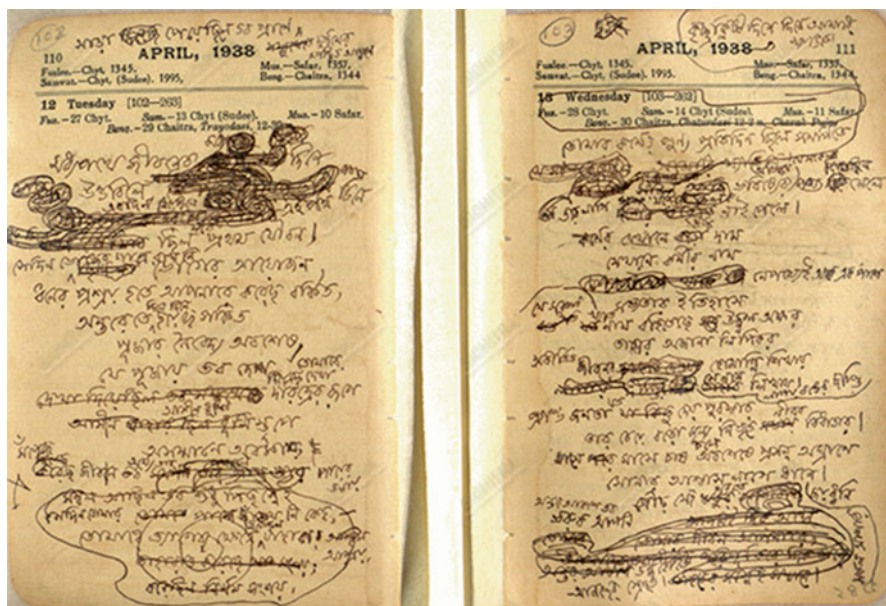


Fig. 5.11 Manuscript draft of an unidentified poem showing various kinds of addition, deletion and revision (MS RB 159 pages 244–5; Bichitra image 126)

identified most of the general problems and found tentative solutions. There were still endless surprises in store: the solutions sometimes proved inadequate or, rarely, unworkable. But we had taken the measure of the task and worked out a broad *modus operandi*. We now set six hands to work on the manuscripts, then eight and finally ten. Of course, there were also many clean fair copies, easier to transcribe than some printed books. These were allocated to yet others. To be honest, relief tempered our disappointment at having no manuscripts of Tagore's longest work, the novel *Gora*, although there are reported to have been 16 drafts.

The rest of this chapter describes how we set about transcribing the manuscripts. But first, we should state the many aims of the exercise, as we defined them to ourselves at the start:

- To provide a full transcription of each manuscript, page by page.
- To incorporate all revisions and additions: deletions, insertions, transpositions, comments.
- At the same time, to extract the final text emerging from a particular manuscript version, after taking stock of these changes.
- To determine the structure of the manuscript: which pages contained what material, and what titles or items they constituted.
- To identify each of these items: a challenge in itself, as some works had never been published, and many others published in versions so different as to elude recognition.
- To indicate the special features of a manuscript like Tagore's celebrated doodles, or directions for staging plays or printing the work.

We should also clarify two things we did *not* set out to do:

- To identify the author of each hand. This would have called for years of painstaking research, with physical access to the manuscripts, by expert palaeographers familiar with Tagore and his circle.
- To analyze the reasons behind the revisions. Bichitra is essentially a database. It provides the material for literary analysis; it does not undertake it.

What, then, were the stages by which we tackled each manuscript?

Numbering the Pages

A preliminary chore was to number the pages consistently. As with most manuscripts everywhere, the Tagore manuscripts were paginated by librarians or archivists at a later date. But with the Rabindra-Bhavana manuscripts, this had been done at different times by different people in different ways. Sometimes there were more than one set of numbers: we had to track each one through to find the most consistent and sustained sequence. This would be the one to record in the transcription. And of course, these page numbers would never match the image numbers, as the latter included things like covers, endpapers, annexures and blank pages.

Hence locating a manuscript page in the Bichitra archive can be tedious work. You must scroll through the transcription till you find the item you want, and note its page number (indicated by a double asterisk). You then click through the images till you find that page.

A tip: To reach page 175 (possibly image 185 or thereabouts), first open page 5, then 85, then 185 in the page-search window on the toolbar. You thereby save yourself 182 clicks. Finally, use the arrows to scroll to the exact page.

We know this is not good enough. Our time and resources did not allow page-by-page matching of image and transcript. If we ever get the chance to radically improve the site, we know where to start.

Detailed Survey of the Manuscript

We would then make a detailed survey of the manuscript. If it was a complicated one with many items, this meant (a) separating each item, perhaps by bringing its scattered pieces together; and (b) trying to identify them. Many manuscript entries carry no clue to their identity, not even a title. The first line of a poem may be quite different from its standard printed and indexed version. A passage in a manuscript may prove to have come from the middle of a poem, which would not be indexed anywhere. The only way to identify it is by trying to match specific words through a search engine. (Until our own was available, we used www.tagoreweb.in and www.rabindra-rachanabali.nltr.org.) Even this method might fail where the manuscript version is radically different. Scholars like Sankha Ghosh and Subimal Lahiri identified some divergent versions from their deep knowledge of the works.

Titles occurring in the manuscript have been transcribed without a sign. Those inserted from other sources are indicated by a single asterisk.

We were finally left with some untraceable pieces that are almost certainly new additions to the canon. (See ‘The Material and Its Challenges’ in Chap. 6.)

Base File Compilation

Before starting to transcribe, we would compile a ‘base file’ of as many items as possible from our transcripts of the printed texts, usually the OCR version of the Collected Works (see Chap. 3). This saved the time and labour of keying in the manuscript text from scratch: we only needed to modify the ready text as required. Needless to say, where there was no transcript of any printed version, we had no option but to key in the whole text from the manuscript.

Transcription

We were now ready to start the actual transcription. This would initially take the form of a page-by-page copy of the entire manuscript, irrespective of content. If, say, a poem was scattered across five separated pages, we would not attempt to join up the pieces at this stage. This **Pagewise Transcript (PT) file**—a .txt version of the manuscript itself—afforded a platform for any subsequent recasting or extracting. (Why .txt? That is a question we answer below.)

We then rearranged the material in an **Itemwise Transcript (IT) file**. This joined up the disjunct pieces of a work if any, then arranged the items by date and genre as in the Collected Works. Thus if a manuscript had several poems each from the collections *Sonar tari* (*The Golden Boat*), *Sphulinga* (*Sparks*) and *Kshanika* (*Momentary Pieces*) plus a scene from the play *Tasher desh* (*The Land of Cards*), the poems from each collection would be placed together, followed by the extract from the play. Finally, we would create separate files for **Individual Items (II)**. Items from the same printed collection would be placed in a single sub-folder. The entire process is illustrated in a flow chart in Fig. 5.12. Unidentified items were placed in a separate **Untraced Folder (UF)**, to be followed up later on. It was cause for celebration every time an item could be taken out of the UF and restored to its proper place, often thanks to Sankha Babu or Subimal Babu's help.

The UF aside, we thus arranged the contents of a transcribed manuscript in three ways: **pagewise (PT)**, **itemwise (IT)**, and by **individual item (II)**. The PT is the transcript the user sees alongside the image of the manuscript. One can adjust the type size or change the display to bold for easier viewing; conversely, one can hide it altogether for a full-screen view of the image.


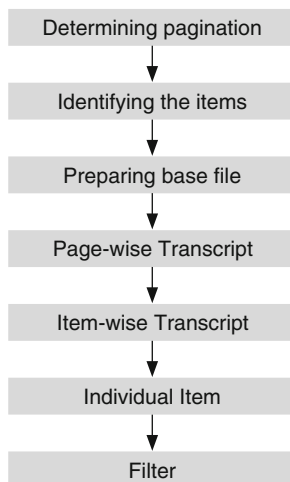
The II contains the text opened by clicking the text icon  against that item in the Full Table in the Bibliography. The IT does not reach the end-user at all: it is needed solely for organizing the material, but is crucial to that task.

Fig. 5.12 Flow chart showing the procedure for manuscript transcription



Even the II is not quite what the user sees. Moreover, as a glance will confirm, the texts viewed in the PT and the II are not the same. That is because before the II is uploaded, it has been cleaned up using a **filter**. Before we see what the filter is or does, we must take a long detour to explain why it is necessary.

Markup and Markdown

How does one transcribe a much-revised handwritten draft with countless deletions, insertions, transpositions, and changes of purpose? The printed texts also needed transcribing, and had some features that called for record. The commonly accepted method today would be to mark up the transcription with suitable encoding for each type of entry: most likely XML markup according to the protocol of the Text Encoding Initiative (TEI).

Some of our School staff had considerable experience of XML transcription of manuscripts. We had transcribed Thomas Hardy's manuscript of *The Return of the Native* and all the manuscripts of the Australian poet Charles Harpur in this way. But these were in the Roman alphabet. For Bichitra, XML was not a viable option. Our primary corpus came to nearly 140,000 pages. Over a third was in manuscript, most of it in Bengali with its cumbersome keyboard. It was hard enough to find enough capable operators to make and check the transcriptions in the time available. It would have been unrealistic to demand training in XML markup and TEI encoding as well. We had to look for a simpler solution.

We finally decided on a set of symbols found on the standard keyboard or, rarely, among the symbols of the extended keyboard of any word-processing software. Shortcut keys could be allocated to these. The commonest symbols we employed were < > to enclose deleted text, and { } to enclose later insertions. Deleted but restored text is enclosed within . Alternative readings, neither deleted, are enclosed within >> and << respectively. Undecipherable text (whether deleted or simply faint) is indicated by +++; if smudged or torn, by «+++». Wherever possible, we have tried to rescue deleted readings, placing them within < >.

Transposed passages are enclosed within ~ ~ and ~ ~ respectively. Passages (usually words or short phrases) shifted from their original locations are placed at their final position within ^ ^ for an upward shift and v v for a downward shift. If there is more than one such instance in close proximity, they are distinguished by numbers: ^ 1^, ^ 2^ and so on. The same principle is followed where a passage is inserted from another (say, the facing) page. In both manuscripts and printed texts, a note at the bottom of the page is indicated by placing the text after *↓*.

That is virtually the lot! Figures 5.13, 5.14, 5.15, and 5.16 illustrate the use of these signs, and Fig. 5.17 gives the full list of signs. The advantage of our system is that the transcribers only need to learn a small number of symbols, all of them inserted by a single (perhaps double-handed) keystroke. This list is small enough to be stuck to the screen front, which is what the operators often did. Or they kept prototypes of the symbols in a small window in a corner of the screen, to be copied

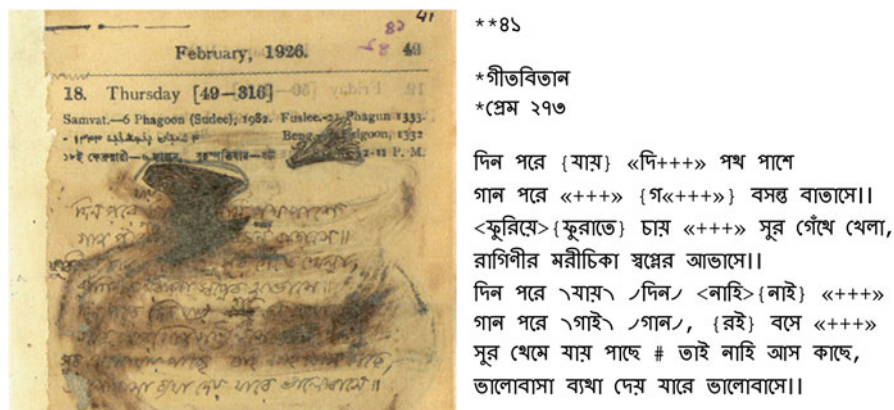


Fig. 5.13 Manuscript text with transcript showing various transcription symbols (MSS RB 027 page 41, Bichitra_image_24)

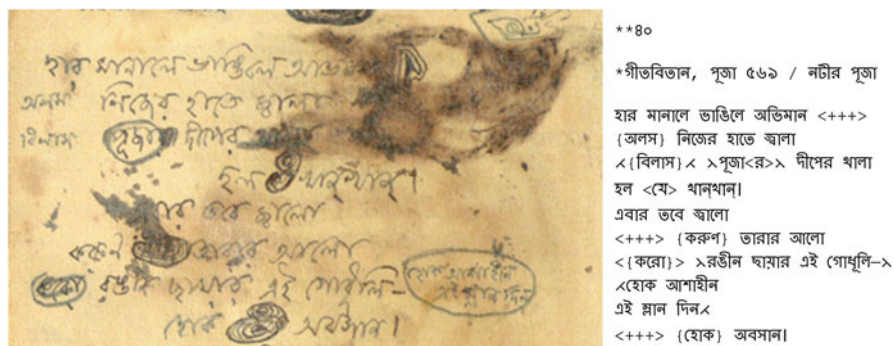


Fig. 5.14 Manuscript text with transcript showing various transcription symbols (RB 027 page 40, Bichitra image 24)

and pasted as required. It was a simple and stress-free system, but it covered most palaeographical features likely to occur in any manuscript. As more and more texts are transcribed across the world in a variety of languages, it may be a good strategy to follow such simple markdown protocols, standardized as far as possible. We ventured on this route from practical compulsions, but we would now propose it as a method of choice for its intrinsic advantages.

As both a logical and a practical consequence, we decided to save all transcriptions in plain-text (.txt) files. These would have minimal markup even at the level of embedded formatting. We thereby ensured that whatever program was applied to the text files—even later, outside the ambit of Bichitra—the file markup would not interfere with its execution, as might happen with WordPad© files.

The .txt format met all our needs. It sufficed, for instance, to feed Prabhed, our complex three-tier collation program: gaps at ‘segment’ level (between prose paragraphs, verse stanzas or dramatic speeches) or ‘section’ level (cantos of a poem,

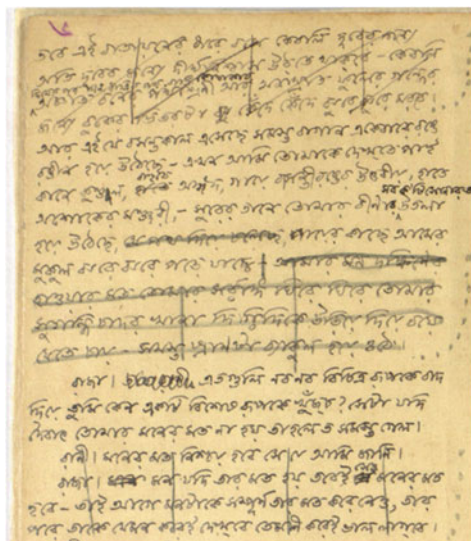


Fig. 5.15 Manuscript text with transcript showing various transcription symbols (RB 143 page 6, Bichitra image 6)

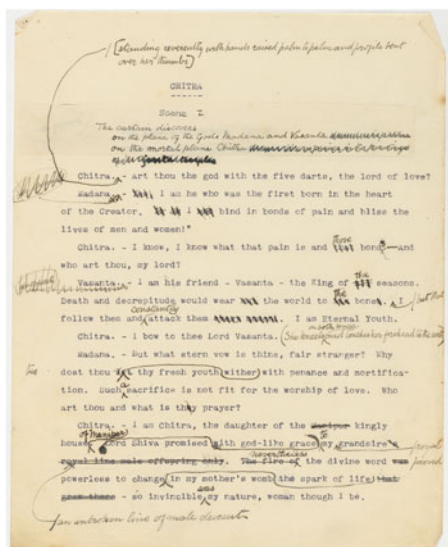


Fig. 5.16 Transcript of a typescript of the English play *Chitra*. The reader can match the variety of transcription symbols to the revisions in the original. (HL 8 page 5; Bichitra image 5)

scenes of a play or chapters of a novel) were indicated simply by using the Enter key—more exactly, the ‘newline character’ (`\n`)—once, twice or three times: that is to say, by starting a text block on a new line, perhaps keeping one or two blank lines in between (see Chap. 8, ‘Parsing’ for details). This simple recourse could

**6

এতবে এই বাতায়নের ধারে বসে কেবলি দুয়ের জন্যে অতি দুরের জন্যে দীর্ঘনিশ্বাস উঠতে থাকবে— কেবলি (দিনের পর দিন, রাত্তির পর রাত্তি কোথাকার) অজ্ঞাত বনের পথশ্রেণী আর অনাস্রাত ফুলের গন্ধের জন্যে বুকের ভিতরটা <বু+++> কেনে কেনে খুঁরে খুঁরে মরবে। আর এই যে বসন্তকাল এসেছে সমস্ত বাগান একেবারে রঙে রঙীন হয়ে উঠেছে— এখন আমি তোমাকে দেখতে পাই কানে কুঞ্জ<++++>[দ]ল, <যাতে> {বাহতে} অন্দ, গায়ে <ব>[বা] সত্তী রঙের উত্তরীয়, যাতে অশোকের মঞ্জরী, — সুরের তানে তোমার বীণার {সব কাটি সোনার তার} উতলা হইয়ে উঠেছে, <যে পথ দিয়ে চলেছে> পায়ের কাছে আসের মুকুল ঝরে ঝরে পড়ে যাচ্ছে— <আমার মন দক্ষিণের হাওয়ায় মত তোমার সর্বাঙ্গ ঘিরে ঘিরে তোমার সুগন্ধি চাদর খানা দিখিদিকে উড়িয়ে দিয়ে বইয়ে যেতে চায়— সমস্ত প্রাণটা ব্যাকুল হয়ে ওঠে।>
রাজা । <তবে রাণী> এতগুলি নব নব বিচিত্ররূপকে দেখে বাদ দিয়ে তুমি কেন একটি বিশেষ রূপকে খুঁজছ? সেটা যদি দেবায় তোমার মনের মত না হয় তাহলে ত সমস্ত পেল।
<রাণী । মনের মত নিচয় হবে সে যে আমি জানি।
রাজা । <মনে> মন যদি তার মত হয় তবেই <সে> {সেও} মনের মত হবে— তাই আপে মনটাকে সম্পূর্ণ তার মত করে নেও, তার পরে তাকে যেমন করাই দেখবে তেমনি করাই ভাল লাগবে।>

**5

[CHITRA]

[Scene I

The curtain discovers

on the plane of the Gods Madana and Vasanta <dressed as in pictures> on the mortal plane Chitra <dressed like a prince in the paintings of the Ajanta temple>]

Chitra . [(standing reverently with hands raised palm to palm and profile bent over her thumbs)] Art thou the god with the five darts, the lord of love? Madana . <[see opposite] Yes, I am he who was the first born in the heart of the Creator. <It is> I <who> bind in bonds of pain and bliss the lives of men and women!>

Chitra . I know, I know what that pain is and <that> (those) bond[s]. (—) And who art thou, my lord?

Vasanta . <[see opposite]> I am his friend— Vasanta—the King of <all> (the) seasons. Death and decrepitude would wear <out> the world to <its> (the) bone<s>. (but that) I follow them and (constantly) attack them <every moment>. I am Eternal Youth.

Chitra . I bow to thee Lord Vasanta. [(She kneels (on both knees) and touches her forehead to the earth)]

Madana . But what stern vow is thine, fair stranger? Why dost thou <let> \ wither \ thy fresh youth \ with penance and mortification. Such (a) sacrifice is not fit for the worship of love. Who art thou and what is thy prayer?

Chitra . I am Chitra, the daughter of the <manipur> kingly house of <the> Lord Shiva promised with god-like grandeur <the> <royal line male offspring only> [an unbroken line of male descent]. <The fire of> (Nevertheless) the divine word <was> [proved] powerless to change \ my nature, woman though I be. <that grew there>— so invincible (was) my nature, woman though I be.

<text>	deleted text
{text}	inserted text
+++	illegible text
±text±	text whose position is uncertain
↖text3↖ text2 text2 text2 text 2 text2 ↗text1↗	text which has been transposed
[text]	underlined text
>text of version1> <text of version2<	two juxtaposed versions of the same text
<text>	stet: retention of text earlier marked for deletion
[~] OR [~]	If a note, comment, instruction etc. is placed in the margin, this marginalia is placed within square brackets [~]. The part of the main text against which it is located in the manuscript is indicated by the sign [~] at the beginning and end of that part.
<^^> OR {^text^} OR <v> OR {vtextv}	<p>Where the original author/scribe has changed the position of a small amount of text (a sentence or less) using an arrow, line or asterisk</p> <ol style="list-style-type: none"> 1. If a section is moved upward, this sign is placed at the original point (without the text) <^^> 2. This sign is placed at the destination point, enclosing the text {^text^} 3. If a section is moved downward, this sign is placed at the original point (without the text) <v> This sign is placed at the destination point, enclosing the text {vtextv} 4. If there are multiple cases of migration in a page, they have been numbered as <^1^>, <^2^>, <^3^> according to their sequence in the page.
^ OR v	<p>If the position of a large amount of text has been changed, the following sign is placed at the destination point</p> <ul style="list-style-type: none"> ▪ if the text has moved upward: ^ and ▪ if it has moved downward: v <p>In these cases, no sign is placed at the original location.</p>
└	A sign like └ or a long vertical stroke in the original manuscript to indicate a line break or paragraph break has been recognized by moving the following section of the text to the next paragraph. The sign └ appears in the transcription at the start of this new paragraph.

Fig. 5.17 Full list of transcription symbols

accommodate the complex structure of long works with many divisions. The files were then converted to the HTML format within the collation program Prabhed. In other words, the procedural markup, if required, *was worked into the target program Prabhed*: the operator loading files to that program did not have to worry about markup, any more than the creator of the original text file.

Using a word comparator with such files should be similarly trouble-free. Such a system frees the operator of the need to acquire a complex markup language. It will immeasurably help large-scale digitization of texts across the world, especially if the operator has to acquire other intricate skills—for instance, in handling a difficult keyboard layout.

How far can we go in simplifying or even eliminating text encoding, while meeting the needs of advanced text-computing programs? We would be happy if Bichitra set the digital humanities community thinking of the possibility.

However, there was another set of textual features where we made some compromises. These related chiefly to layout, and applied to both manuscripts and printed texts.

Most importantly, we ignored indentation—at the start of a paragraph, or within a verse stanza. To record indentation is perfectly possible. In fact, for our earlier collation software Pathantar (see Chap. 8), we had used a # sign to indicate one level of indentation, repeating it as required for deeper levels. When presenting the text file, the # would automatically be replaced by a certain number of spaces.

We could have used a similar device in the Bichitra files. We refrained only because we thought it would get out of hand. Tagore's poetry can have several depths of indentation (see Fig. 5.18). These intricately (even randomly) graded indent levels would have greatly increased the labour of transcription. It would also have made the text files unduly complex, and vastly complicated the collation results. We decided, not without heartburn, to enter all lines flush with the left margin. We also indicated all line breaks by the standardized symbol \perp , replacing various scribal practices like the Bengali word *phnak* (gap), the contractions 'para' or 'NP', or the symbol ¶.

We did, however, use # to indicate a short pause in the middle of a verse line, usually represented by a variable blank space (see Fig. 5.19). This symbol was used both in manuscripts and printed texts, though in the former it was sometimes hard to tell such a gap from the normal one between two words.

Speech headings in drama were simpler to handle. In printed play texts, to say nothing of manuscripts, these could be entered in all sorts of ways: on the same line as the speech or separately above it, centred or flush with the left, in various fonts, with varied punctuation or none. All these were printing or scribal conventions, with no implications for the contents of the text. There seemed no point in burdening the text files and, even more, the collation results with these inconsequential variants. We had no hesitation in reducing them to a single style:

[Heading] [space] [stop] [space]

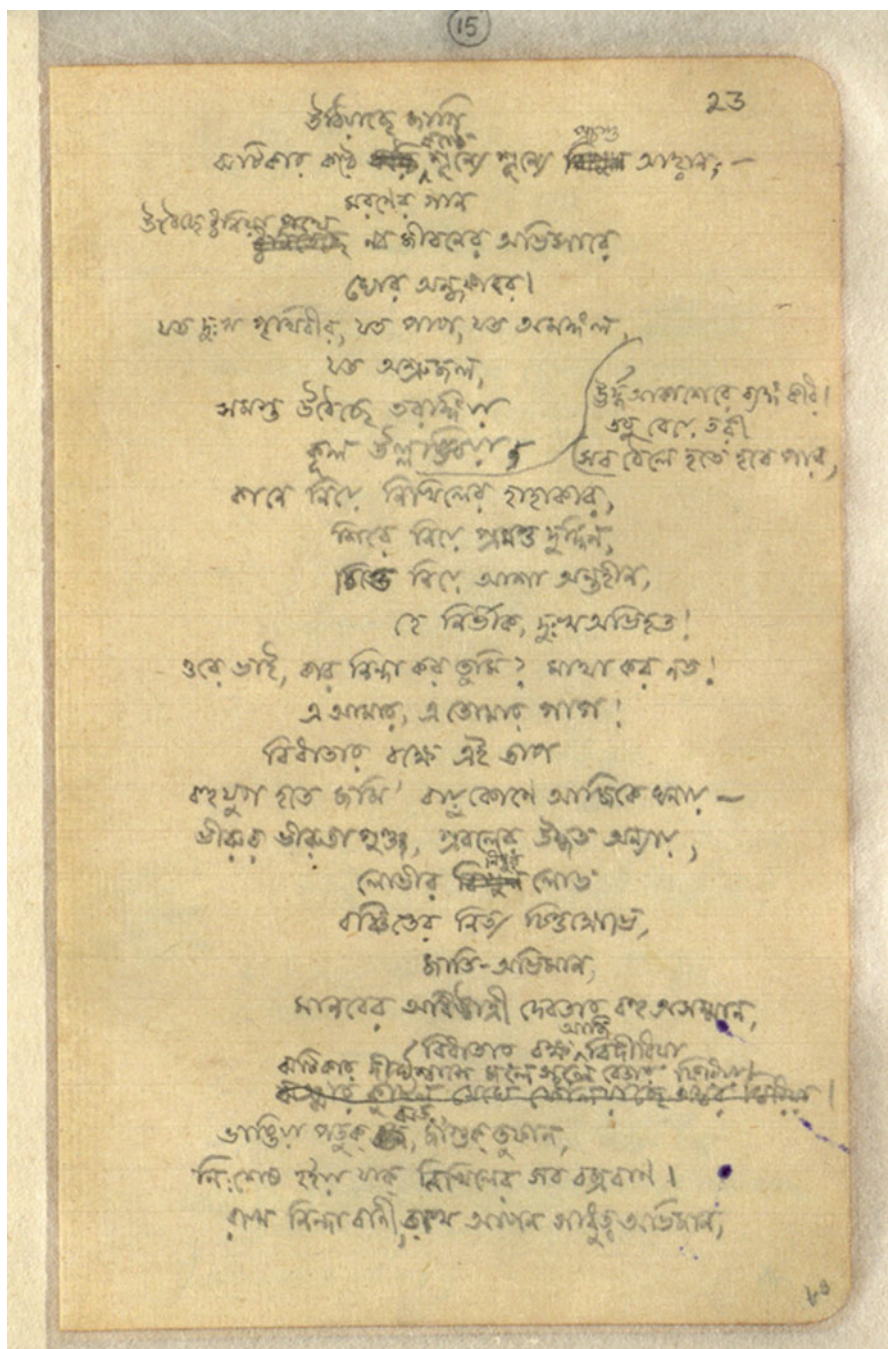


Fig. 5.18 Indentation at various depths in a poetical manuscript (MS RB 111 page 15; Bichitra image 22)

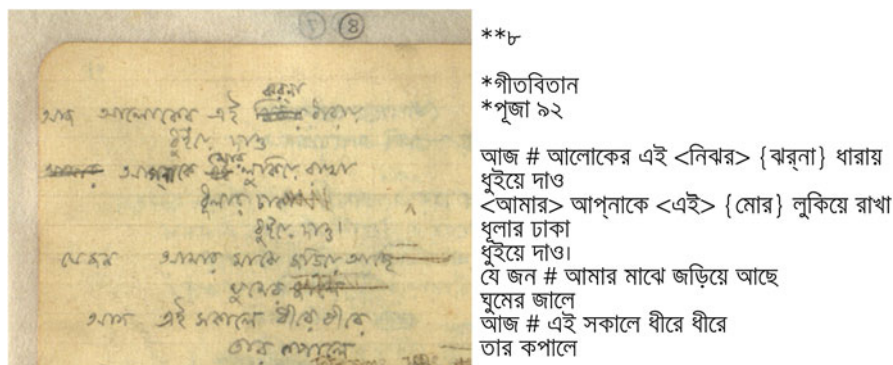


Fig. 5.19 Use of # in transcript to mark a gap or caesura after the first word in certain lines (MS RB 111 page 8; Bichitra image 17)

followed by the speech on the same line.

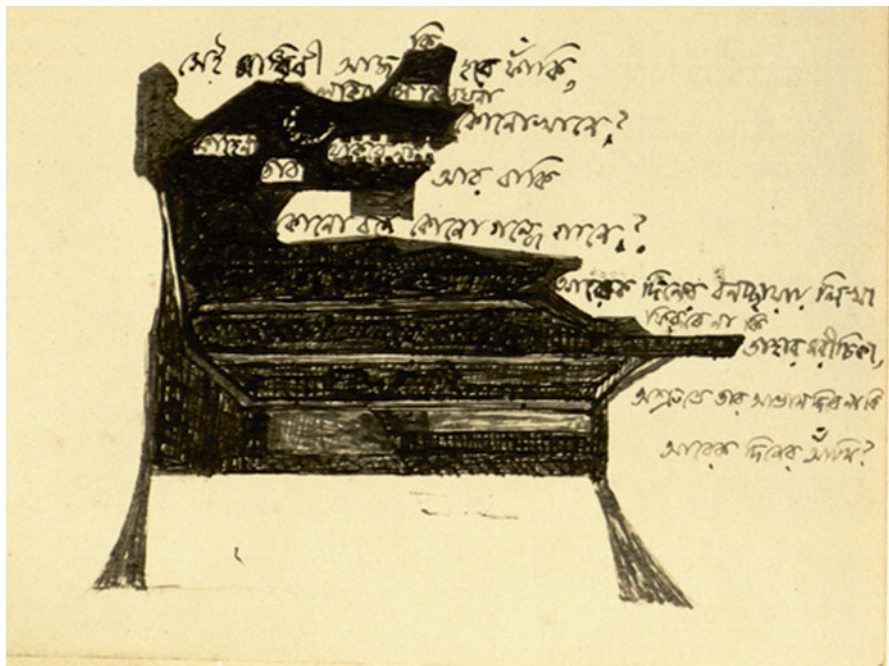
In manuscripts of plays, the speaker's names were sometimes omitted altogether. In such cases, we indicated each new speech by the symbol $\sqrt{\quad}$.

The Filter Software

Take a look at Fig. 5.20. This is the copy of a complicated manuscript page with a lot of transcriptional symbols, sometimes nesting the < > for deletions and {} for insertions inside each other. Such a transcription is almost impossible to read. Moreover, if fed into the collation program, the program would present perfectly correct results, but so intricately as to mystify the human viewer. For the user's benefit, and for collation or other processing, we needed to extract a clear comprehensible text from this welter.

Bhupati Ray, then a graduate student in Jadavpur's Computer Science Department, produced such a program, which we named the 'Filter Software'. This read the transcriber's symbols and, accordingly, removed all deleted text; placed translocated text in its final position; arranged the text as indicated by the other symbols; and also deleted the symbols themselves, leaving a clear text—the final version *emerging from that manuscript*. Needless to say, this may be quite different from the standard version or any other. Also, needless to say, the filter would preserve symbols occurring in the manuscript itself—we took care to ensure that none of these found their way to the transcriber's repertoire.

Once we had applied the filter software, we could claim to have plumbed the mysteries of the manuscript. It remained to place the resultant text file alongside those of other versions of the work—perhaps with implications for the bibliographical entry, but most importantly for the collation.



সেই <++++> {মাধুরী} আজ <++++> {কি} হবে ফাঁকি,
 <++++> {লুকিয়ে <++++> সে কি <++++> রয় না} কোনোখানে?
 <++++> {কাহিনী} <++++> তার <++++> {থাকবে না} আর বাকি
 <++++> কোনো বর্ণে কোনো গন্ধে গানে?
 <++++> {আরেক দিনের বনচ্ছায়ায় লিখা
 {ফিরবে না কি} তাহার মরীচিকা,
 অশ্রুতে তার আভাস দিবে নাকি
 আরেক দিনের আঁখি?}

Fig. 5.20 Complex transcript of manuscript with doodle (MS RB 048 page 14; Bichitra image 14)

Manuscript Doodles

Our transcription agenda did not engage closely with Tagore's celebrated manuscript doodles. These are unique creations that he evolved by joining up and elaborating the deletions in his drafts. There is nothing like them in the manuscripts of any other writer (see Chaudhuri 2010, 183–210). At first glance, they sometimes recall William Blake's illuminated manuscripts; but those were carefully laid-out ensembles of words and images, the latter deliberately designed to form their own spaces, not recycled scraps of text.

Barring small stray examples in earlier manuscripts, the doodles appear in full array in MS RB 102, among drafts of poems in the collection *Purabi* (see Fig. 5.21).

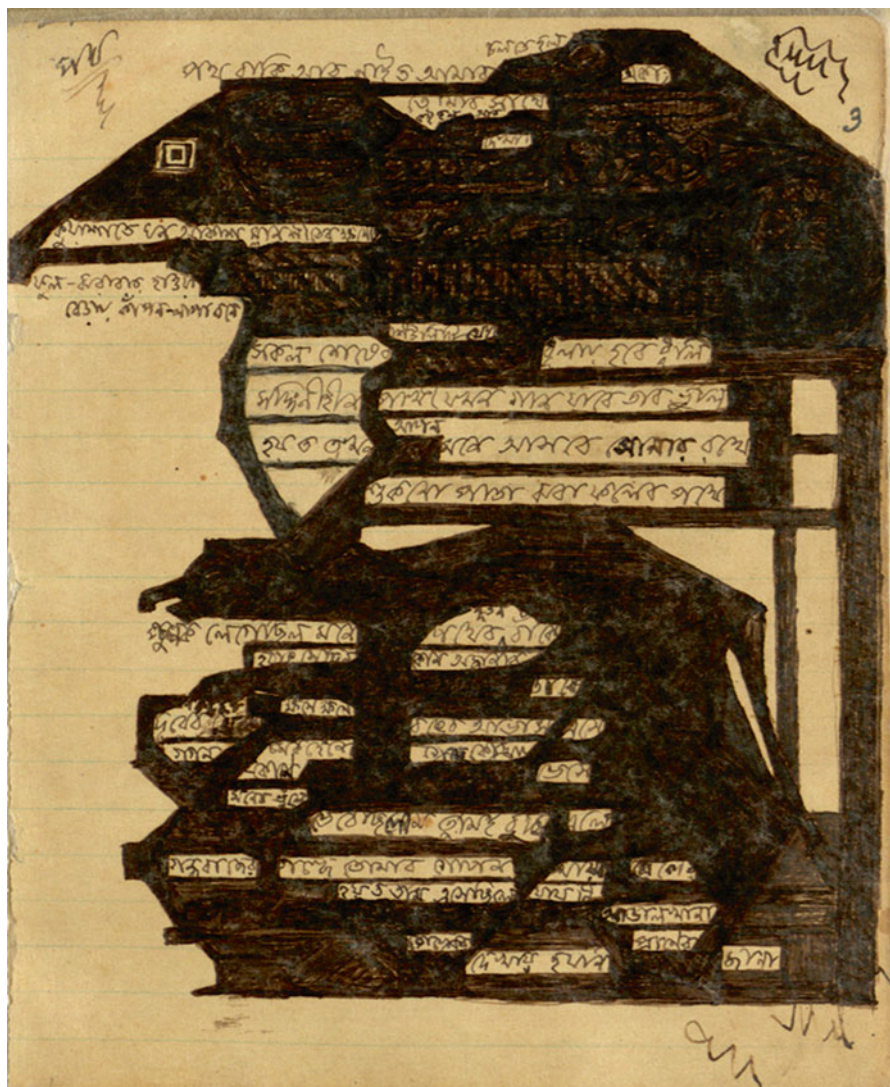


Fig. 5.21 Manuscript with doodle (MS RB 102 page 2; Bichitra image 4)

These are not doodles in the sense of idle scrawls, but elaborate artifices that seem to be the artist's true goal, with the words peeping out through lattices in the design. It was a disappointment but also a relief that for purposes of transcription, the doodles could be treated like any other deletion. If the text beneath them was decipherable, we transcribed it between the usual symbols <>; if not, we indicated it by +++. The Full Table of manuscripts notes (in the last column, 'Others') when a manuscript contains doodles.



Fig. 5.22 Manuscript with doodle partly obliterating some words (MS RB 048 page 16; Bichitra image 16)

A few doodles posed more substantial challenges. Sometimes the doodle could obliterate words the poet wished to leave standing. Again, in Fig. 5.22, the verses nestle along the meandering shape of the composite creature in the doodle. It is sometimes hard to tell what constitutes a single line, or where certain lines—notably one near the creature’s bent ‘knee’—are to be placed. For a similar problem in



Fig. 5.23 English manuscript with doodle and uncertain sequence of words (MS RB 008 page 128; Bichitra image 74)

tracing the sequence of words through the components of the doodle, see an English example in Fig. 5.23. Most remarkably, page 13 of the same manuscript RB 048 (see Fig. 5.24) is almost completely deleted by a great dark blob across the page, with lighter trails and scrawls. In the right margin are a few lines crammed into a meagre white space. They appropriately begin, ‘This ogre is called Oblivion.’



Fig. 5.24 ‘Doodle’ almost fully obliterating the text. The few remaining lines at the right begin ‘This ogre is called Oblivion.’ (MS RB 048 page 13; Bichitra image 13)

The doodle, having obliterated the text, has spawned new text in turn, a story about its own being. All the transcriber’s tricks cannot represent this sequence of text > image > text. For that, we need a new program for transcription in multimedia. We hope someone—perhaps ourselves—will make it one day.

Reference

- Chaudhuri, S. 2010. 'Writing Pictures, Drawing Words: The Manuscript Doodles of Rabindranath Tagore.' In *The Metaphysics of Text*, Sukanta Chaudhuri, 183-208. Cambridge: Cambridge University Press.

Sukanta Chaudhuri, Purbasha Auddy, and Debapriya Basu

Starting with the Basics

This section is about file management, but more basically about human management. When the core team of 30 people started working on Bichitra, they had no clear idea what the end product would look like. Sukanta, the project head, confesses he didn't either—which was probably a blessing, as he would not have taken on the task otherwise. Nothing on the scale and complexity of Bichitra had been undertaken before, and projects at all comparable in scope were operating on a much more relaxed (or even open-ended) schedule. We, on the contrary, had to complete the task in just over two years.

Having rushed in where angels would fear to tread, it was hardly surprising that we found the challenges mounting at every step. Thankfully, we soon acquired the knack of tackling them with something more than fatalism. Several minor problems would crop up every day, and major ones roughly once a week. We found we could solve them by putting our heads together and applying simple logic. In retrospect, some of the solutions might have been neater, but they worked nonetheless.

One thing, though, we realized from the outset. Given the vast amount of primary material, and the staggering number of files that would be generated, it was

S. Chaudhuri (✉)

Department of English, Jadavpur University, Kolkata, India

e-mail: schaudhuri@english.jdvu.ac.in

P. Auddy

School of Cultural Texts and Records, Jadavpur University, Kolkata, India

e-mail: pauddy@gmail.com

D. Basu

Department of Humanities and Social Sciences, Indian Institute of Technology,
Guwahati, India

e-mail: debapriya.06@gmail.com

vital from the start to keep absolutely clear tabs on all the material we collected, and all the work we put into them. It is important to grasp that not only is the end product of the Bichitra variorum accessed through computers; the back-end support system and indeed the backroom operations, the organization and management of the project, would have been impossible without computers. It called for a huge exercise in data and file management, including data relating to human output that would not, of course, go on the site or even lurk behind it.

It was also an exercise in *metadata* management. The file structure generated by the project and recorded on our spreadsheets was simultaneously creating the most elaborate Tagore bibliography ever compiled. To be sure, the directory tree in its original form (see ‘Spreadsheets as connectors’ below) could only be read by a computer, not by a human being. It would have to be translated into human terms and processed in various ways to create the hyperbibliography. But it provided the intrinsic design for such a construct.

Our most gigantic task was to create UTF-8 text files for every version of every work by Tagore, working from images of those versions. As explained in Chap. 3, OCR (Optical Character Recognition) is hardly available in Bengali. OCR is a process whereby the computer scans and ‘reads’ the image of a text and converts it, character by character, into a text file. We were lucky enough to get OCR-generated files of the standard edition of Tagore’s works. But these had to be modified for every variant version of each work by manually keying in the new text. Those other versions were often so different that we virtually had to create new files from scratch. This was the task that kept most of the team of 30 busy for 18 months and more, a few of them till almost the end of the 2-year project. They were set certain targets to meet: 20 pages (of a standard size) for prose and 30–35 pages for verse, variously adjusted to the nature of individual texts. Targets for manuscript transcription were necessarily lower. Also, of course, every transcription needed to be checked. We had thought this would take less time than the original transcription, but soon found this was not the case. We had moreover to keep a record of all the work done: how many image files we had received, how many had been transcribed as text files, and how many of those had been checked.

The team was divided, though flexibly, into four groups. The first team did the initial keying in of Bengali printed texts, which were then checked by the second team. A third team tackled the Bengali manuscripts, after an initial period of working with printed texts to get their hands in. The fourth team took charge of English printed texts and manuscripts (the latter often typescripts), dividing the work among themselves and checking one another’s output.

We realized early on that to err is human, and that technology is temperamental. It was vital to ensure a foolproof backup system, even to excess. Each operator saved their output on their own computer, as well as a shared flash drive which circulated between the transcription and checking teams. But most importantly, all the data was stored on an internal server, from where it could be centrally accessed for overall sorting, editing and processing. Each operator had their own account for accessing the server to copy necessary material. The images were stored on external hard drives and, of course, the internal server. Individual operators would make

copies of the image files they needed to work from. Moreover, all the computers were connected through a Windows sharing system, so that material could be accessed on any terminal as required.

Our task was complicated by the fact that (as explained in Chap. 4) we did not start with all the material to hand. Scans from a range of sources trickled in almost till the end of the project. There was no order to this flow: hence we could not process the material in a systematic way by genre, title or date as we would ideally have done. This inflow of material, then, was another thing to monitor—through a kind of stock register, as it were. But to prepare the register, we needed an initial working bibliography (bibliographical control, to use the technical term) of all the material we had to access.

To meet these needs, we maintained two MS Excel© spreadsheets right from the start. Unlike the spreadsheets described later on, they did not form part of the site's storage and retrieval system. They served simply as an in-house monitoring device for materials received and work done.

One spreadsheet was bibliographical in the broadest sense. Drawing on standard print bibliographies of Tagore's works (see below, 'Printed works'), we created a working list to serve as bibliographical control: needless to say, this needed extending and amending as we located more and more works and versions, and acquired more elaborate data about them. Against this, we recorded the items of which we had obtained images and/or text files.

Another spreadsheet kept a log of the work done: which items had been transcribed, and which of those checked as well. We used it to keep a discreet tab on the rate of progress: the date when each text had been allotted to a transcriber, the target date, and (depressingly at times) the actual date of completion.

Both spreadsheets were uploaded to Google Drive, with shared access only to the supervisors. Purbasha had the formidable task of presiding over these vital records of materials and work flow. We used to say she was the super-computer who kept track of what the mere computers were doing, and could pull out any data from the maze.

Marshalling the Material

Bichitra presents 139,157 pages of primary material: 47,520 of manuscript and 91,637 of print. This material can be divided in a different way, into the following categories:

- Manuscripts: 711 manuscript volumes
- Printed texts: Bengali:
 - approx. 361 books and 2426 journal items comprising
 - 4441 poems and songs (+ hundreds of very short pieces)
 - 84 dramatic texts
 - 165 novels and stories
 - 1191 essays and other non-fictional texts

- Printed texts: English:
 - 51 books and 433 journal items comprising
 - 1035 poems and songs (+ hundreds of very short pieces)
 - 17 dramatic texts
 - 5 short stories
 - 338 essays and other non-fictional texts

In other words, not only the number of pages but the number of *individual items* posed a huge challenge. They had to be classified and ordered so that both the human user and the computer could find their way around them and select the ones needed at any point of time. The computer's logic, needless to say, is utterly different from the human, so we needed a double retrieval system.

This indicates the most challenging aspect of the Bichitra bibliography. There are some learned Tagore bibliographies, their acme being the monumental 16th volume, edited by Sankha Ghosh, of the collected Bengali writings (*Rachanabali*) published by the Government of West Bengal (Ghosh 2001). But these are stand-alone printed volumes: readers consult them for information, then go away to locate the items elsewhere. The Bichitra bibliography, on the other hand, was at the same time an index and pathway to the contents of a website. Each entry had to be linked to its images, text file and collation files, in a different way to the search engine, and differently again to the Timeline. It had to open up to take in, say, all Tagore's plays, or all works published in the Bengali year 1323; but equally, to zoom in on a single poem or essay in all versions or a particular one. A click on the bibliographical entry had to take the user to any of these resources. At times it also had to link one text with another (as for collation) or trawl the entire corpus (for the search engine). Bichitra, we may claim, presents the world with the first true **Hyperbibliography** — at least the first on such a scale.

This makes Bichitra an *integrated* website: the data relating to all these works in all versions can be accessed in its totality, across the board. How much simpler our lives would have been if each work had been individually processed and recorded on a separate DVD! But that would have defeated the whole purpose of the site.

Data Collection: Catching the Fish

The intricate business of file management for the computer's access will be treated later in this chapter. Let's begin by seeing how the material is stored, recorded and sorted for human access.

The electronic database could only be created after a great deal of bibliographical research of the traditional 'pre-computer' kind. Unlike, say, in the case of Shakespeare, we did not have a ready-made store of data that only needed transfer to the electronic medium. Rather, we had to follow the advice given in a cookbook for settlers in new territories: 'First catch your fish.' The advantage of the electronic medium was that we could access nearly all the material in screen images and work our way quickly through them. To look it all up physically, perhaps across many libraries, would have taken far longer.

Needless to say, this does not lessen our debt to scholars like Prashantakumar Pal, Sisirkumar Das, Sankha Ghosh and Swapan Majumdar, or their predecessors like Charuchandra Bhattacharya, Pulinbihari Sen and Kanai Samanta. Our work would have been impossible without their labours: if we had to catch the fish, they provided us with rod and line. We admire all the more that they could achieve so much without the help of electronic resources. And last but far from least, Sankha Ghosh (Sankha-da or Sankha-babu, depending on how well one knew him) was a tower of strength as scholarly adviser to the Bichitra project itself.

To continue the angling metaphor, how did we deal with the fish all the way from riverbank through kitchen to table? A couple of basic points before we begin:

- Tagore's works are traditionally grouped in four categories: Poems and Songs, Drama, Fiction and Non-Fictional Prose. This familiar and convenient division was the natural choice for Bichitra as well. The site presents all material in this way. Nearly all the back-end spreadsheets and other files were so divided too, after a preliminary division into Bengali and English works.
- Bengali books and journals are variously dated by three eras: the Common or Christian (CE), the Bengali, and the Saka. These had to be reduced to a common base, especially for the Timeline. We opted for the Bengali, as this was the one most frequently found. There are formulae for converting Bengali years to CE (add 593, i.e., add 600 and subtract 7) and Saka to CE (add 78, or 100 minus 22). Thus broadly speaking, 2014 CE is year 1421 in the Bengali era. But the Bengali or Saka year does not begin in January; so for their last few months, you add 594 or 79 to get the CE equivalent. Often we do not know the precise month: another reason for sticking to Bengali dates.

The Material and Its Challenges

Printed Works

It was impossible to include every imprint of every work: in the case of popular titles like *Gitanjali*, they would run into hundreds. In most cases, it would also be unnecessary, as they are mere reprints of no independent textual value. We followed the usual bibliographer's principle of including only the versions that are textually significant, either because of known revision or, given the provenance, the likelihood of revision. So our bibliography includes:

- All manuscript versions.
- All early versions printed in journals (nearly always the first printed version): in case of serial publication, with details of the part that appeared in each number of the journal.
- The first print version in volume form, or included in a book-length publication.

- The first edition published by the Santiniketan Press or Visva-Bharati Publication Division—i.e., the institution founded and headed by Tagore himself, hence presumed to carry special textual authority.
- Any other version of known textual or bibliographical importance.
- All anthologies with which Tagore might have been associated.
- The Collected Works published by Visva-Bharati Publication Division ([Rabindra-rachanabali 1939](#)).
- Translations and adaptations by Tagore of his own work.

Translations and adaptations by other hands would have made for an impossibly large and open-ended list. It would also have mired us in copyright problems. On this principle, we even omitted English translations like *The Post Office* and *The King of the Dark Chamber*, published by Macmillan under Tagore's name but actually made by others.

For each version so included, we have provided the basic bibliographical data: title, date, publisher's name, details of journal publication and so on. Much of this data was available in Sankha Ghosh's compilation referred to above ([Ghosh 2001](#)), Swapan Majumdar's Tagore bibliography *Rabindra grantha-suchi* ([Majumdar 1988](#)), Prashantakumar Pal's detailed multi-volume life of Tagore, *Rabi-jibani* ([Pal 1982](#)), and various bibliographies compiled by Pulinbihari Sen ([Sen 2009](#)). But Majumdar's work comes only up to 1912 and Pal's to 1916. A lot of the data had to be extracted from the publications themselves. In fact, we checked all the data from first-hand sources, going through the contents, indexes and actual page-wise matter of the versions included. Even this arduous task was full of pitfalls: an item may be omitted from the Table of Contents or wrongly indexed, perhaps by the first line on top of a page rather than the first line of the poem itself.

With the English writings, bibliographical control was still more tenuous. Our starting point was Tagore's four-volume *English Writings*, originally edited by Sisir Kumar Das and, after his death, by Nityapriya Ghosh (Das and Ghosh [1994](#)); but this is far from complete. Very many items, chiefly essays and speeches, are available only in journals, and imperfectly indexed at that. Sometimes the only way to locate them was to sift through entire runs of likely journals like *The Modern Review*. There was also an unsuspected amount of manuscript material that had not been published at all.

Especially with English journals, a good number could not be obtained from our chief source of primary material, Rabindra-Bhavana. They had to be hunted through various other libraries and, once located, scanned by our own project staff. Other items, especially the text of Tagore's speeches, were brought out as pamphlets by Visva-Bharati or, more elusively, by the place where he delivered the speech.

Gathering the data was the first step. The next, involving much planning and labour, was to bring order to the range of Tagore's works in their many versions and formal categories. Some works are long book-length items: novels, full-length plays, travelogues or other long prose discourses. (He wrote no epic poems.) A much larger quantity—not just in number but in total bulk—comprises shorter pieces. Nearly all the poetry is lyric, or at most narrative or philosophic pieces of

moderate length. The fiction includes a lot of short stories besides novels, and the drama one-act or medium-sized as well as full-length plays. And, of course, very many items first appeared in a journal: not only short pieces but full-length novels in serial form. Much of the non-fiction consists of short essays, speeches or sermons; but there are also book-length discourses, and (much more than with poems) single short pieces printed as separate volumes or pamphlets. Later on, they might be placed in a larger collection.

The poems were often moved from one collection to another, sometimes more than once. The items now comprising the two-part collection of narrative poems *Katha o kahini (Legends and Tales)*, along with others placed in one or other part but later withdrawn, provide a specially intricate set of examples. Poems were also revised, sometimes drastically, when placed in anthologies like the classic *Chayanika* and *Sanchayita*. Chap. 2 describes the repeated revision of the poem 'Nirjharer svapnabhanga' ('The Spring Wakes from Its Dream').

Most elusively, short poems and songs could be included in a collection of verse, but also embedded in a long play, novel or prose discourse. An outstanding example is a set of poems in the famous novel *Shesher kabita (The Last Poem, 1929)*. These reappeared the next year, with minor changes, in the verse collection *Mahua*. Very many songs, popular as stand-alone items and so published in the standard song collection *Gitabitan*, are also included in a play, sometimes more than one. Such inclusion could be an afterthought for a song composed at a much earlier date. The musical drama *Shapmochan* consists almost wholly of such songs, whose selection differs radically from one version of the play to another.

Hardest to track are short pieces embedded in other short pieces. For instance, the original Bengali version of the song 'More life my lord, yet more' ('Prana bhariye trisha hariye') was set in a philosophical essay in the Bengali periodical *Tattvabodhini patrika* in 1912. A year later, it was published separately as a song in another periodical, *Prabasi*. These texts within a text can be really elusive: unless there is a reference in an earlier source, they can only be tracked down by combing every page of every work. Once located, they need a note in the bibliography stating that they are part of a novel, play, or essay. These are 'big' and 'short' pieces at the same time, stand-alone works needing separate entries but also components of bigger volumes, differently structured.

There were more unusual problems too. A major one concerned the eight-volume 1903 anthology *Kabya-grantha*, bringing together the contents of the 14 books of poems and songs Tagore had published by that date. The editor Mohitchandra Sen reorganized the nearly 1100 poems under 26 themes, cutting across the lines of the original volumes where they had appeared. Some titles of poems were changed and sections of text deleted. In a long preface, Mohitchandra justified this plan by arguing that his thematic arrangement would help the reader grasp the poet's intentions more easily. Most readers seem to have demurred: the arrangement was never repeated. In the process, *Kabya-grantha* created a standing challenge for later bibliographers. Moreover, Tagore wrote a set of new poems to introduce the theme of each section. Those introductory poems were brought together separately 11 years later, in the collection *Utsarga (Dedications, 1914)*.

To gather this data was an aspect of ‘traditional’ scholarship. Now came the challenge of putting it in electronic form. Only an electronic bibliography could map the intricate data, but to design such a tool was no easy task. It took a lot of planning to reduce the many-tiered disparities to a manageable system. Inevitably, there was an element of blind flying at the start. We would decide on a strategy, proceed smoothly for some time, then come across some new data that defied our categories. If we were lucky (we usually left space for manoeuvre) we could simply tweak the original design. Once or twice, we had to undo the work of days or weeks and start from scratch.

Manuscripts

Our manuscript material came from two sources. The bigger by far was Rabindra-Bhavana, Santiniketan. Much smaller in bulk but of great importance was a cache of English manuscripts in the Houghton Library, Harvard University. Except for a few celebrated instances earlier reproduced in facsimile (like the Harvard manuscript of the English *Gitanjali* which won Tagore the Nobel Prize), Bichitra opens up the wealth of Tagore’s manuscripts to the world for the first time.

When the site was nearing completion, Samantak offered a manuscript in the possession of his family: the only privately held manuscript included so far. We hope there will be others: Tagore often gifted short autograph poems to friends and admirers. More substantial manuscripts also found their way to private hands. In fact, we tried to track down some of them and failed. Many have sunk without trace (of the 16 alleged manuscripts of the novel *Gora*, not a single one made it to Rabindra-Bhavana). Others, more sadly, are traceable but inaccessible. Specially disappointing was the tacit refusal of a notable private collector to allow us access to a precious clutch of manuscripts in their family collection. We still have hopes of including them one day...

Once the manuscript (or rather its image) was obtained, we had to process it. A manuscript often contains the text of many works of different kinds, perhaps in both Bengali and English: only a consolidated list could convey its full contents, including important add-ons like Tagore’s marginal notes or his celebrated doodles. Library labels and catalogue entries often did not provide full or correct information: they had to be cross-checked with other sources like the bibliographies cited above, and finally with the manuscript itself. For Bengali, the most important source was Sanatkumar Bagchi’s handbook on the Tagore manuscripts in Rabindra-Bhavana (Bagchi 1989). The special problems with English manuscripts are separately treated below. Ultimately, we had to go page by page through every manuscript and build up our own bibliography—also, for English, a Master List. But we needed some bibliographical control to begin with, drawn from the above print sources. This chicken-and-egg puzzle called for much finesse in management.

The outcome of these researches was a spreadsheet serving as a comprehensive control sheet for manuscripts, the back-end file underlying the Index of Manuscripts on the webpage. This is one of 32 back-end files, about which much more below (see Fig. 6.1 for the item heads in this spreadsheet, with the corresponding features

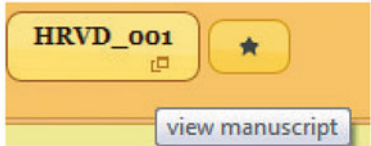
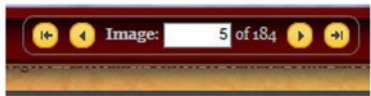
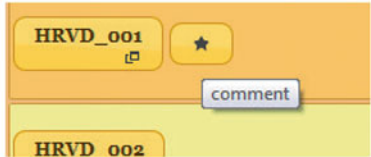



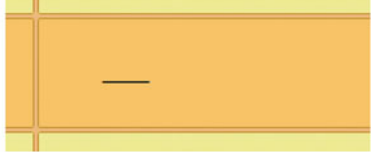
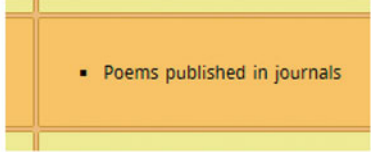
Backend	Interface
Folder name The shelf mark of the manuscript), linked to the image folder and transcription text file	
Number of images Not seen by the end-user, but required for image retrieval for the image viewer	
Remarks if any Not viewed in a separate column by the end-user, but accessed through a radio button	
Poems and Songs Titles only	
Drama Titles only	
Fiction Titles only	
Non-fiction Titles only	
Other material e.g., doodles or other information	

Fig. 6.1 Item heads in the ‘Control Spreadsheet’ for manuscripts, with corresponding onscreen display. The entries relate to the original Harvard ms. of the English *Gitanjali* (MS HL 1)

onscreen). Bichitra thus offers a fuller account of Tagore's manuscripts than either of the libraries housing them. Quite how varied and extensive the data is has emerged in Chap. 5.

The manuscripts also threw up many pieces we could not identify. We might even be left wondering whether we were looking at a single long piece or several short ones. Some items were identified by our expert adviser Sankha Ghosh. Others were traced later on by a word search of the complete works¹ SNLTR (2009), Tagoreweb (2010–2012): they proved to be fragments from somewhere in the middle of a known poem, or an early draft differing so widely from the final product that one would scarcely guess the link. Even so, a number of pieces remain unidentified. We would like to think these are our discoveries, expanding the Tagore canon; but there is a remote chance that one or two may be by other hands (see Chap. 5, 'The manuscript material'). Needless to say, we would be grateful for any help in identifying them.

We also met many challenges while checking each manuscript item against the list of printed works. One was the sheer extent of revision, and the variants produced as a result in both manuscript and print versions. Here we can only look at English examples in detail, though Bengali items offer the most intricate or interesting cases.

The poem beginning 'I know that at the dim end of some day' was published in the journal *Modern Review* in 1914, and then in the major collections *Fruit Gathering* (1916), *Gitanjali and Fruit Gathering* (1918), and *Collected Poems and Plays* (1936). It also occurs in five manuscripts, all opening with the same line as the printed version except MS RB EMSF_011. Here the poem occurs twice. The first instance (Image 2) is very faintly written in pencil with the first part missing. The second (Image 16), in ink, begins 'I know my days will end'. Assimilating these texts as earlier forms of the same poem took time and labour (see Fig. 6.2).

Again, with the poem 'I hid myself to evade you' published in *Lover's Gift and Crossing* (1918) and *Collected Poems and Plays* (1936), the version in manuscript MS RB 308A begins 'Yes, yes, strike me again, yet again'. MS HL 4 has an edited typescript beginning 'Yes, yes, strike me ~~more~~ again, yet ~~more~~ again', while MS RB 369(ii) begins 'Yes, yes, strike me more, yet more' (see Fig. 6.3).

Both these examples are English renderings of works originally written in Bengali. Lectures and essays that began life in English could be endlessly redrafted to suit various occasions, making it harder and harder to trace connections. There are as many as 20 manuscripts in the MS RB 304 series with material resembling the lectures ultimately printed as *Talks in China* (1924). Taking all sources into account, some of the *Talks* have over 30 versions. The MS RB 305 series of 16 manuscripts, containing material for *The Religion of Man* (1930), shows intensive cross-pollination between essays. Sections of two or more published essays can often be traced back to a single manuscript: for instance, a single piece in manuscript ('My Introduction', MS RB 319B) becomes three separate essays in the 1924 edition of *Talks in China*. The typescripts are richly loaded with material illustrating the evolution and transmission of Tagore's English essays. They are often heavily edited in

¹ We used Tagoreweb and NLTR at the start, and our own search engine once it was ready.

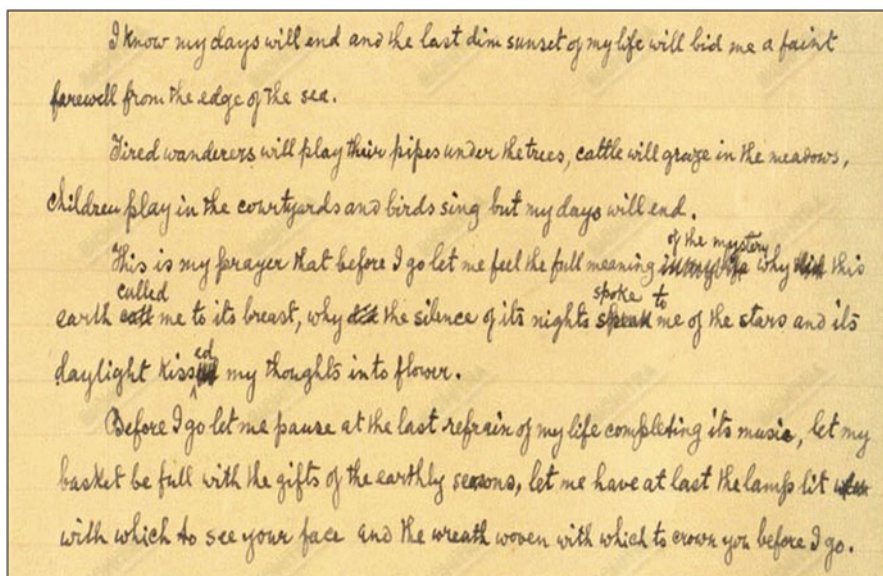
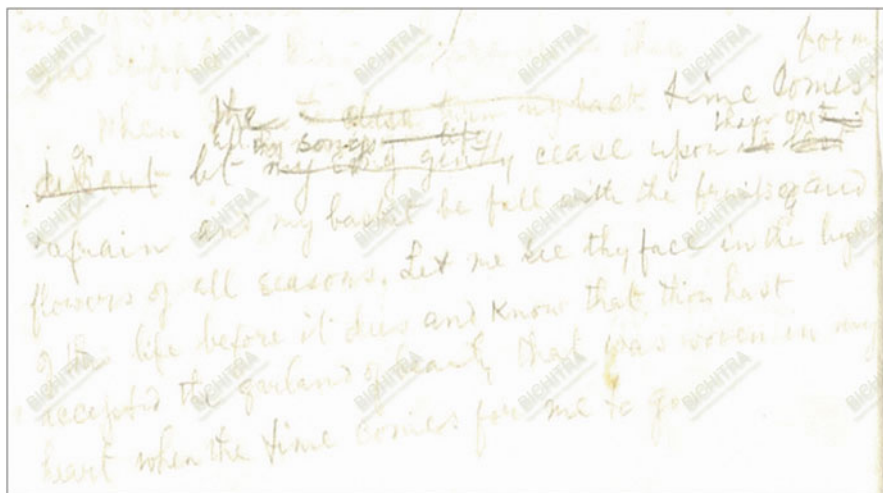


Fig. 6.2 Two versions of the same poem in the same manuscript: MS RB EMSF 11 page 1, Bichitra image 2, and page 16, Bichitra image 16

Tagore's hand, whole pages crossed out and long paragraphs inserted. Even so, the final manuscript version may bear little resemblance to the published text.

Even within a single essay, the typescripts show remarkable changes. The essay titled 'An Indian Folk Religion', published in *Creative Unity* (1922), survives in two Rabindra-Bhavana typescripts, MS RB 104A and MS RB 104B. In the former, it is conveniently titled 'The Folk Religion of India', but begins 'Men born and brought up on the upper slope of society ...'. The latter manuscript version has no

XXXIX.

Yes, yes, strike me again, yet again.
I hid myself and evaded you ever;
but now I am caught at last.
Strip me of all that I have.
Finish the game for good, - either
you win or I.
I have had my laughter and song in
wayside booths and courtly halls, - let
me see how you break my heart and make me
weep.

Crossing 25

39

Yes, yes, strike me ^{again} ~~more~~, yet ^{again} ~~more~~.
I hid myself and evaded you ever;
but now I am caught at last.
Strip me of all that I have.
Finish the game for good, -- either
you win or I.
I have had my laughter and song in
wayside booths and courtly halls, -- let
me see how you break my heart and make me
weep.

39

Yes, yes, strike me more, yet more.
I hid myself and evaded you ever; but
now I am caught at last.
Strip me of all that I have.
Finish the game for good, - either you
win or I.
I have had my laughter and song in
wayside booths and in courtly halls, - let
me see how you ~~can~~ break my heart and make
me weep.

Fig.6.3 Three versions of the poem 'I hid myself to evade you': MSS RB 308A page 33, Bichitra image 34; HL 4, page 39, Bichitra image 52; RB 369(ii) page 146, Bichitra image 25

title and begins ‘It was beginning to grow dark and the singing party sat on the grass under the open sky’. The opening words of the printed version, ‘In historical time the Buddha comes first of those who declared salvation to all men’, occurs several paragraphs later in both typescripts.

The English Corpus

As the above examples show, Tagore’s English writings present a special set of problems. Though substantial, they constitute only a fraction of his Bengali writings in bulk. Most of them are translations or adaptations from the Bengali, so that they already have a bibliographical reference point. But the relation between the Bengali and English texts can vary dramatically. Tagore might rework the original radically, or translate only a part—perhaps omitting the start—which hampers identification. He can fuse several Bengali originals in whole or part into a single English piece, as in the poem ‘The Sunset of the Century’ concluding his famous prose work *Nationalism*. A line loosely translated from Bengali might serve as the springboard for an independent English piece.

Moreover, Tagore revised his English writings over and over, often much more intensively than his Bengali works. Perhaps he felt less secure in what was, after all, an alien tongue; but that is not the whole story. He often produced radically different translations of the same Bengali text, as though revelling in a virtuoso series of variants. The short epigrammatic poems collected in *Stray Birds* (1916), *Fireflies* (1928) and the bilingual *Lekhan* (1927) are specially prone to multiple versions, totally different from one another and, at times, from those that made it to print.

With prose works, the problem is different but analogous. Tagore was a master of English prose, and wrote many original essays in that language: some of them were later recast in Bengali. But we find sections of an English work matching disjunct sections of Bengali, sometimes from more than one source. These can be even more elusive than matches in the shorter and more memorable verse.

In a word, Tagore’s English writings cannot be organized in a simple manner. Yes, there are some celebrated publications, chiefly collections of translated poems, essays (both translations and English originals) and aphoristic jottings. But looking at the units comprising these collections—the individual poems, essays and aphorisms—unveils much scope for confusion, as Tagore was fond of working his texts across genres. The most basic cross over, of course, lies in the fact that Tagore invariably translated his Bengali verse into English prose.

In a word, Tagore’s English writings offer textual and bibliographical problems out of all proportion to their published bulk. Their frequent status as translations or adaptations poses an extra challenge not present in the Bengali works. By bringing together all this material, including the original manuscripts from both major repositories—Rabindra-Bhavana and the Houghton Library—Bichitra makes it possible to explore this wealth for the first time.

But before presenting it to the world, we had ourselves to bring order to the scene. This was often bewildering, sometimes amusing but overall a very rewarding job.

The first task of the English bibliographical team was to identify the manuscripts with English material. As the detailed Index of Manuscripts shows, some manuscripts at Rabindra-Bhavana combine English and Bengali material, others contain English alone. The Harvard manuscripts are almost entirely in English, consisting as they do of drafts that Tagore sent his friend, the artist William Rothenstein. Harvard received them as part of the Rothenstein papers.

To list such manuscripts was easy if laborious, to identify the contents much harder. The first problem was to determine—if we could!—which English writings were actually by Tagore. Bichitra only includes pieces translated from Bengali by Tagore himself. Very often the handwriting confirms his authorship, checked against Das and Ghosh's *English Writings*. But there are also texts in hands other than Tagore's, or in typescript or even printed form, whose authorship we needed to check. Many members of Tagore's circle translated his works: they included his niece Indira Debi, his nephew Surendranath Tagore, and friends and associates like Edward Thompson, William Pearson, Krishna Kripalani and Kshitimohan Sen. It is not always easy to identify their handwritings. Tagore's own hand can differ widely; he would also dictate his work or have it transcribed by someone else.

We cannot even rely on attributions to Tagore made in print. Two of the best-known translations, *The Post Office* and *The King of the Dark Chamber*, still appear under his name though they were actually by others. (It took some research to discover that only one song in the latter play could be confidently attributed to Tagore.) Journal publications often cite a single name, Tagore's, leaving it uncertain whether he had composed only the original or also the translation. We had to consult other sources to confirm authorship: Pulinbihari Sen's work was specially helpful in this task (Sen 2009).

To take a complex example: in the collection *Broken Ties and Other Stories* (London: Macmillan, 1925), only the story 'Giribala' is indicated by Sisir Das as translated by Tagore himself. We had a journal version of the piece, but no manuscript which could confirm Tagore's role as translator. However, we found several manuscript and typescript versions of another piece which could be identified as the story 'Emancipation' in the same collection. The trouble was that the manuscripts marked this piece as part of a poetry collection, *Lover's Gift and Crossing* (London: Macmillan, 1918). Incidentally, this piece was developed from a poem titled 'Parishodh' ('Repayment') first published in the Bengali collection *Katha* in 1900. Much later, its story was worked into the musical drama *Shyama* (1939). Thus a Bengali poem was converted into an English short story, and also extended and reworked into a Bengali musical drama consisting entirely of songs. Such cases—and examples abound—obviously go far beyond 'translation' in the conventional sense.

These are instances of fascinating textual transmission executed by the author himself. Many more instances, often richer and more intricate ones, occur in Bengali alone. Sadly, it is impossible to describe them in a book written in English.

Folder and File Management

Retrieving Images and Texts

Both image files and text files increased in number with each passing day. It became an imperative need to organize them systematically, so that they could be retrieved without delay.

We began by dividing the material into Bengali and English, and then each division into the four standard forms or genres: Poems and Songs, Drama, Fiction and Non-fiction. For image files, there were two additional categories in Bengali: Works (the Collected Works or *Rachanabali*) and Collections or Anthologies. Each category in each language had a folder to itself with a subfolder for each work, further divided as described below. We used three-letter abbreviations in the Roman alphabet as filenames and folder names, taking care that no two works ended up with the same combination of letters. (See below, ‘Naming the Files’.)

Having devised a folder structure to meet our needs, we modified it suitably for text and images. The entire text of each version of a work constituted a single file. Hence all text files of the same title could be put in a single folder named after the work (see Fig. 6.4 for the folder structure for text files). This folder structure served to organize the text files in ready packages for collation. When text files of all the versions were ready, the title was ready for collation: the files could be retrieved from the folder and uploaded to the collation program.

For image files, we needed an extra level of sub-folders, as each image constituted a separate file. So the third level of sub-folders was followed by a fourth, where each separate version of the work constituted a sub-folder, and each image a

Fig. 6.4 Folder structure for organizing text files

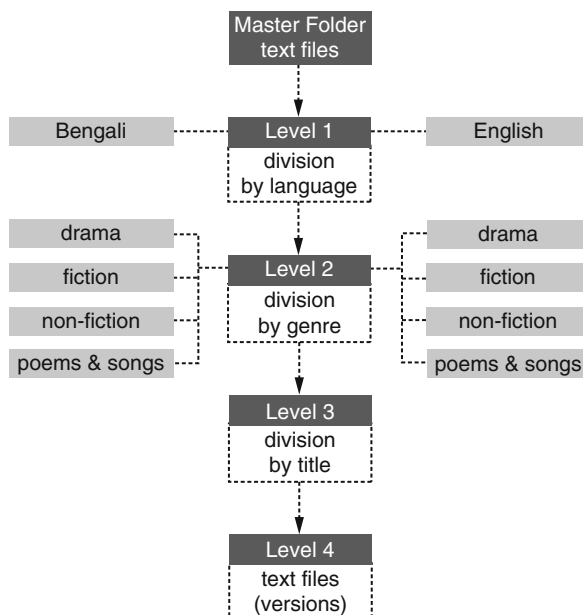


Fig. 6.5 Folder structure for organizing image files

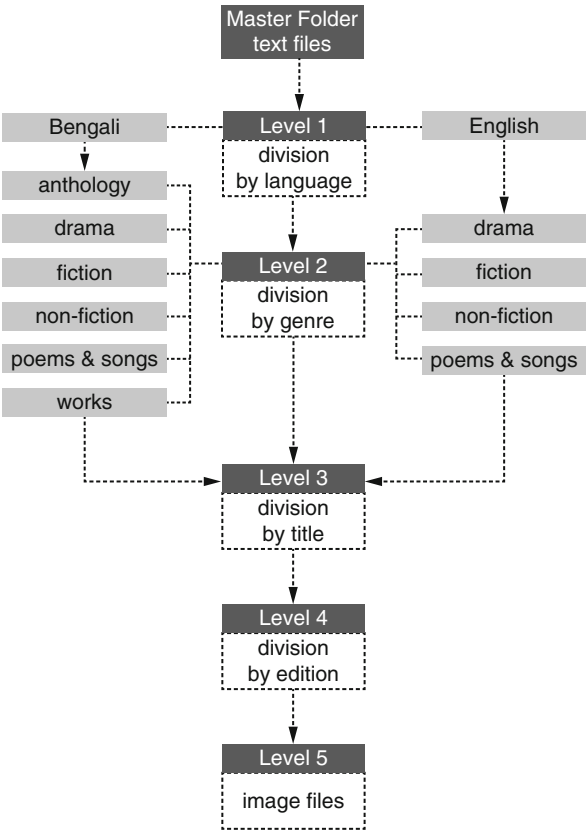


Table 6.1 Sample spreadsheet entry

Collection	Book	Folder name
Gitanjali	Gitanjali (Song Offerings) (London: The India Society, 1912)	e_p_git_1912_b_01

file within it. More often than not, manuscript pages were not neatly divided by content: separating the pages containing a particular work was simply not feasible. The entire manuscript had to be stored as one long unit, in a single sub-folder containing all the pages in a single sequence irrespective of content. Within this, each image was placed as a separate file. The structure took shape as in Fig. 6.5.

We thus have ten divisions at subfolder Level 2, incorporating the basic categories of material (six Bengali and four English). Their entries are laid out in ten spreadsheets, which serve as back-end files for linking titles of works to the coded filenames of image and text folders/files. They are among the 32 back-end spreadsheets supporting the entire website.

A sample spreadsheet entry looks as in Table 6.1. The method of naming files and folders is explained below under ‘Naming the files’.

These ten spreadsheets are linked to two others that lay out the data from the two Master Folders for Texts and Images (see Figs. 6.4 and 6.5). These two latter spreadsheets are accessed for actually fetching the text files and image files respectively.

Here, then, are 12 (10+2) of the 32 back-end spreadsheets on which the entire data retrieval system for Bichitra is based. These 12 spreadsheets are used for retrieval of image and text files; they underlie the ‘Browse Collection>Printed Books and Periodicals’ submenu. The other 20 back-end spreadsheets, underlying the Bibliography menu, are described below under ‘Organizing the bibliography’.

Naming the Files

At the School of Cultural Texts and Records, we have dealt with digital files for a long time, coding and naming them on certain principles. When we took on projects under the British Library’s Endangered Archives Programme, their guidelines greatly helped us to streamline our practice (British Library 2014, section 7). We fine-tuned the method we had already adopted of abbreviating the titles of works to a letter code using the Roman alphabet. This made it easier to call up or search for the files in a uniform way. We adapted our earlier practice to suit the more complex demands of the Bichitra project.

All titles in Bengali and English were given a three-letter code using letters from the title. This, however, was only the first step in generating the full filename. Each version of the work had to have a unique code name, which also had to indicate some other features to facilitate uploading and search by language, genre etc. So the code ‘git’ for *Gitanjali* had to be extended to include the following information:

- edition by date (e.g., 1912): git_1912. For manuscripts, manuscript number instead of date: thus git_H001, where H001 indicates it is manuscript no.1 in the Harvard collection.
- type of edition: git_1912_b, where b stands for ‘book’. The other possibilities are m (manuscript), p (periodical), c (collections or anthologies) and w (the collected works or *Rachanabali*).
- genre: p_git_1912_b, where p stands for ‘Poems and Songs’. The other possibilities are d (drama), f (fiction), and n (non-fiction). The genre code was placed even before the title code, as logically the genre to which a work belonged had to be determined before we could look for that title under the correct head.
- language: e_p_git_1912_b, where e stands for ‘English’ (and b for Bengali). This code had to come first of all, as all the material on the site is organized according to language.
- For short items (especially lyric poems and essays), another addition was necessary to indicate the specific item within the larger work or collection; thus
 e_p_git_001_H001_m
 for the first poem in Harvard manuscript no.1 or
 e_p_git_003_1912_b
 for the third poem in the English *Gitanjali* of 1912.

language	genre	title code	item number	year of publication	type of publication	publication number
e						
e	p					
e	p	git				
e	p	git	001			
e	p	git	001	1912		
e	p	git	001	1912	b	
e	p	git	001	1912	b	01

Fig. 6.6 The seven steps in generating the version filename.

Even for a long text like a play or novel, where there are no smaller items to be numbered in this way, the number 001 was inserted at this point, as all the filenames had to be of uniform structure.

- Yet another component to allow for multiple editions in the same year, or multiple versions of the same item in a single manuscript. If there had been two editions of the English *Gitanjali* in the same year, their first poems would have been coded respectively as

e_p_git_001_1912_b_01 and

e_p_git_001_1912_b_02.

If (as usually) there was only one edition or version in a particular year, 01 would still be added at the end, to ensure a uniform structure for all filenames.

Thus the full coded filename had seven components. The steps by which they were generated are explained in Fig. 6.6.

Unfortunately, we didn't have the foresight to develop this system fully from the start. By the time we finalized it, more than half the text files had been processed and needed to be renamed. To do so manually would have been a stupendous task, so we trawled the Internet for freeware that could be used to rename text files in large batches. We found several such, of which we chose Alex Fauland's AF5 renaming software to rename .txt files (Fauland 1999), and a Rename Master ([joe-joe no date](#)) to rename folders.

Image folders were named in almost exactly the same way. But here only the folders were codified, not the individual image files, which were named using an 8-digit numerical code (00000001, 00000002 and so on, including covers, title page, front matter, end papers etc.)

One final step was still required to incorporate the text file names, image folder names and collation results in the website, which used .php files as commonly in website applications. The PHP page parsers needed the exact paths of these files and folders. Since the structure was already well defined, the paths were easily understood by the human user; but they needed to be precisely defined for the PHP parser so that the webpage could load quickly. For example, the path of the text file

e_p_git_001_H001_m_01 (the first poem in the English *Gitanjali*, in the version found in MS HL 1) is

```
english/poems_and_songs/git/e_p_git_001/e_p_git_001_H001_m_01
```

Naming the files in this systematic way allowed smooth transfer of thousands of files to the server. All files with a particular letter at a particular point in the filename could be readily placed at a particular location on the server: e.g., all files starting with e to the English division, all those with p at the specified point to 'Journal (Periodical) Publication', and so on. Titles, of course, could be identified by their unique three-letter codes. This filename structure also allowed files of a particular type or content to be retrieved from the directory for onscreen display.

A separate set of text files (coded as w) of the versions in the Collected Works was kept aside for the search engine, which accessed only that version of the text. These text files for the search engine were renamed according to the same principles, and linked to the spreadsheets which contained metadata about those files. This metadata is displayed at the front end of the website: this is what end users get to see, rather than the filenames which would mean little to them.

The collation files used the basic three-letter code of the work being collated. This code would be incorporated in the name of the configuration file created by selecting the text files to be collated and specifying the parser. The same filename (with a different extension) would be adopted by the GCL (gross collation: see Chap. 8) file created after collation, as well as the final results folder. The vast number of files within that folder, containing the actual collation results for individual text blocks, would be automatically named by the software, indicating the position of the text block being collated. Thus the results file for section 2, segment 5 of the base text will be named 2.5.

Spreadsheets as Connectors

We were relieved to see that our organization of the images and text files enabled the project staff to find what they were looking for by their own intuition or common sense. The structure first conceived as a practical convenience to monitor the project finally generated the directory structure of the website (see Fig. 6.7 for a visualization). All the coded files are allocated to the appropriate folders within this structure.

It is this set of coded files that allows the Bichitra bibliography to function as a hyperbibliography, with links from the metadata entries to the actual images and text files. Of course, the coded files are meant for computer retrieval: the end users of the website do not have to browse the collection by these complicated codes. In the back-end spreadsheets described below under 'Organizing the bibliography', the codes are listed in technical columns which contain links to the actual image files and text file, as well as the collation files through a specified path. In the front-end 'Full Table' in the Bibliography menu, these links are converted into clickable symbols.

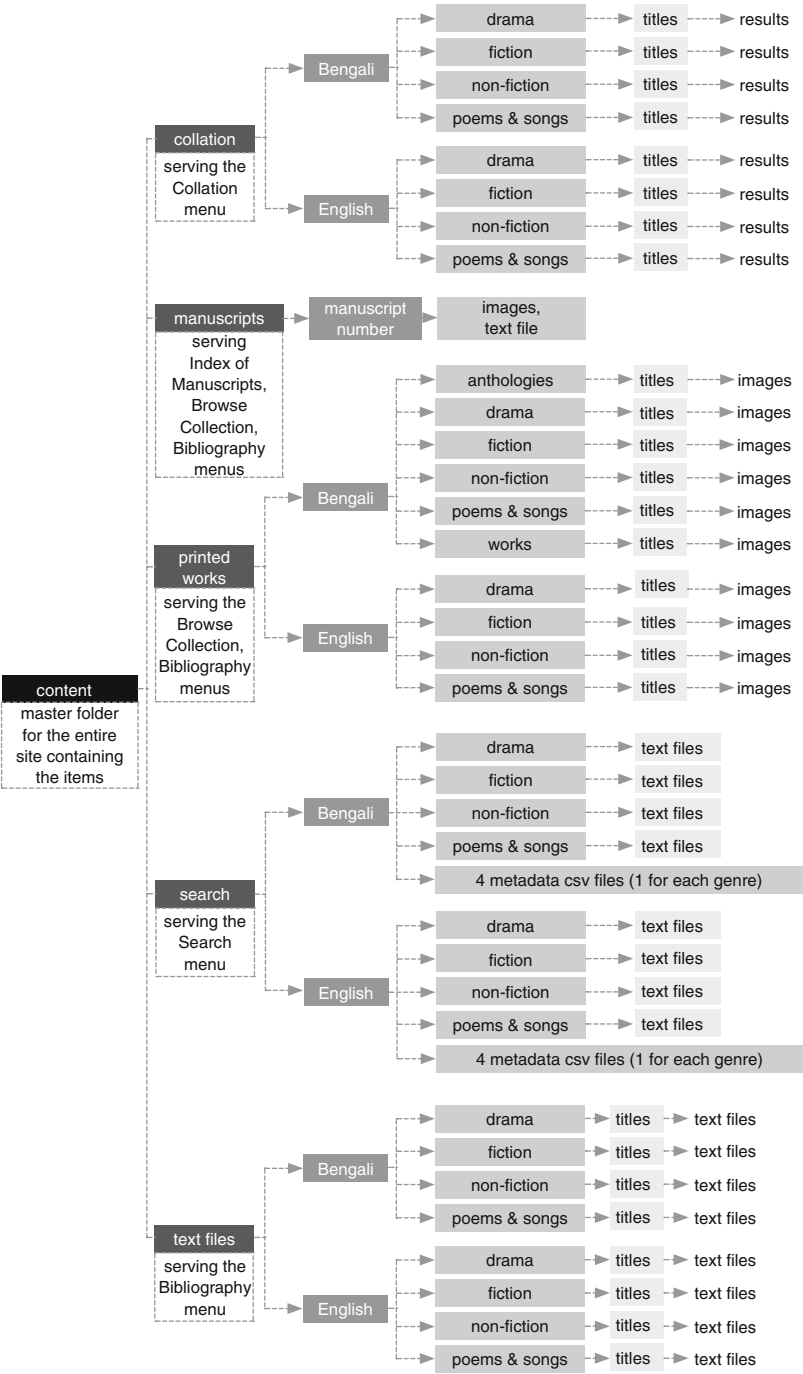


Fig. 6.7 The Bichitra directory tree

Organizing the Bibliography: A Spread of Spreadsheets

We have described the file management system for retrieving files for display: in other words, the system directed to the computer's use. We will now explain how we organized the material for human consultation and access in the front-end bibliography.

For this purpose, our first step was to create an MS Excel© spreadsheet with the headings shown in Table 6.2, to serve as a primary checklist.

We began by filling in data on the printed material in columns 3, 4 and 5. Each short poem or essay was entered separately. We then went page by page through each manuscript, noting the individual pieces and listing them in column 2. Those that could not be so related were entered as separate items. We were already on the way to the most detailed checklist ever made of all the material in Tagore's manuscripts and authoritative print editions.

From this initial list, we created a much more elaborate set of spreadsheets, uploaded as a Google document for the bibliography team to develop and the transcription team to use as a control sheet, checking off each item as it was transcribed. Here, each genre had its own file with many more column heads, which kept filling up as both teams proceeded with their work. For the Poetry/Song spreadsheet, the heads were as shown in Table 6.3.

Table 6.2 Primary checklist spreadsheet headings

Column no.	Column headings
1	Title/First Line
2	Manuscript/typescript
3	Journal/Periodical/Newspaper
4	Book
5	Anthologies/Collections
6	Original (for English works only)
7	Comments

Table 6.3 Poetry/song spreadsheet headings

Column no.	Column headings
1	Title
2	Collation file
3	First Line
4	Manuscript/Typescript, Text file thereof
5	Journal, Text file thereof, Image folder thereof
6	First Edition, Text file thereof, Image folder thereof
7	Other Edition(s), Text file thereof, Image folder thereof
8	First Anthology Inclusion, Text file thereof, Image folder thereof
9	Other Anthology Inclusion(s), Text file thereof, Image folder thereof
10	Inclusion in songbooks
11	Location in Collected Works (Rachanabali) (for Bengali works only)
12	Translations, Recastings etc. (for Bengali works only)
13	Original (for English works only)

Table 6.4 Bibliography full table headings

Column no.	Column headings
1	Title
2	First Line
3	Manuscript/Typescript
4	Journal
5	First Edition
6	Other Edition(s)
7	First Anthology Inclusion
8	Other Anthology Inclusion(s)
9	Original (for English works only)

Some of the columns did not apply to drama, fiction or non-fiction, whose spreadsheets therefore had fewer columns. There was no entry for the image folder under ‘Manuscript/Typescript’, as this data was listed in the separate spreadsheet for manuscripts described earlier, and the information was fetched from there.

These spreadsheets underlie the Full Table in the Bibliography menu accessed by end users. The Full Table displays only the column heads shown in Table 6.4.

The other components of the back-end spreadsheets, containing technical data, are translated into radio buttons. (See below for details.) Additional information is presented in pop-ups under the appropriate fields. Their presence is indicated by an asterisk [*] both in the back-end spreadsheet and in the front-end display. Other symbols like \$, #, + are used to indicate various functions like formatting, bullets, line breaks etc.

We needed to provide some additional lists as well. Many pieces, especially poems and essays, were published at different times under different titles. These items are listed in the main Bibliography by the title or number in the collection where they first appeared. To include the alternative titles in the main Bibliography would have further encumbered an already heavy and complex arrangement. We included this information in a table accessed separately through the drop-down menu. Another such table listed the contents of various collections of short stories: there were simply too many of these to include in the main bibliography. These tables (in the submenu ‘Additional Lists’) are stand-alone pages with no hyperlinks; but they can be searched for any version of the title or, for poems and songs, the first line, by using the Find command (CTRL+F or CMD+F). In fact, the Full Table in the Bibliography menu can also be searched for alternative titles in this way.

Ultimately, the back end of the Bibliography section came to comprise 20 elaborate spreadsheets (see Fig. 6.8). As described above under ‘Folder and File Management’, there are another 12 back-end spreadsheets supporting the ‘Browse Collection’ section.

The entire retrieval system of Bichitra rests on this raft of 32 back-end spreadsheets (see Fig. 6.9 for the full structure). They are parsed by customized parsers or small programming scripts (stored in the PHP files used to generate the webpages) that we created to extract ready meaning from the string of coded information. These back-end files are in constant operation behind the visible or front-end menu accessed by the user. We may say the back-end spreadsheets are like underground book stacks in large libraries, to which only the staff have access. The resources stored there are fetched as required for users in the public reading rooms.

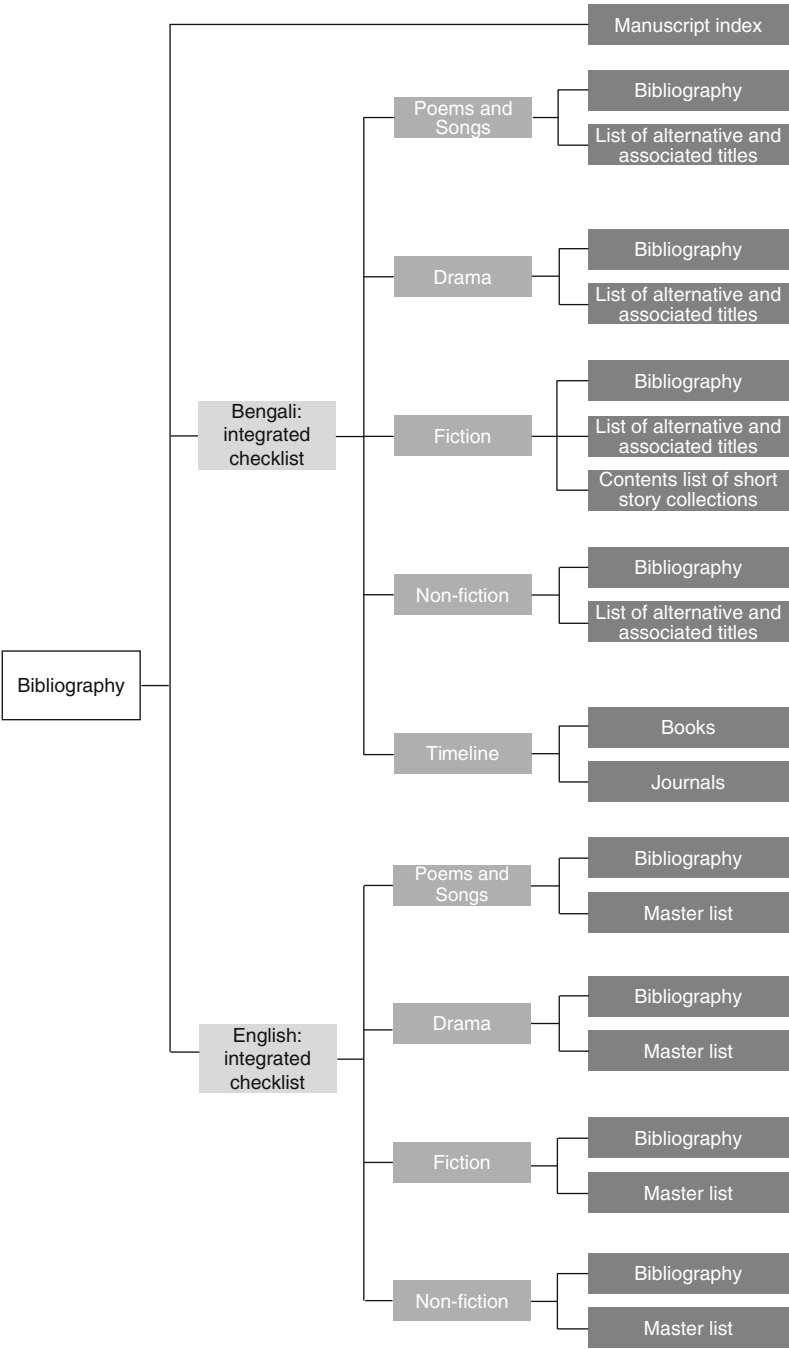


Fig. 6.8 Back-end spreadsheets for Bibliography menu. See also Fig. 6.9

Display name			
Bengali	1	Anthologies	for Browse Collection menu
	2	Drama (books+periodicals)	
	3	Fiction (books+periodicals)	
	4	Non-fiction (books+periodicals)	
	5	Poems and songs (books+periodicals)	
	6	Bengali works (rachanabali)	
English	7	Drama (books+periodicals)	
	8	Fiction (books+periodicals)	
	9	Non-fiction (books+periodicals)	
	10	Poems and songs (books+periodicals)	
Paths for fetching files			
	11	Image file	for Browse Collection, Bibliography menus
	12	Text file	
Checklists			
Bengali	13	Drama (with links: text file, image folder, collation result)	for Bibliography, Collation menus
	14	Fiction (with links: text file, image folder, collation result)	
	15	Non-fiction (with links: text file, image folder, collation result)	
	16	Poems and songs (with links: text file, image folder, collation result)	
	17	Alternative or Associated Titles-Bengali-Drama	for Bibliography menu
	18	Alternative or Associated Titles-Bengali-Essay	
	19	Alternative or Associated Titles-Bengali-Fiction	
	20	Alternative or Associated Titles-Bengali-Poems and Songs	
	21	Short stories-contents of collections	
	22	Chronology-book	
	23	Chronology-journal	
English	24	Drama (with links: text file, image folder, collation result)	for Bibliography, Collation menus
	25	Fiction (with links: text file, image folder, collation result)	
	26	Non-fiction (with links: text file, image folder, collation result)	
	27	Poems and songs (with links: text file, image folder, collation result)	
	28	Drama-master list	for Bibliography menu
	29	Fiction-master list	
	30	Non-fiction-master list	
	31	Poems and songs-master list	
Manuscripts	32	Manuscript Index	for Browse Collection, Bibliography menus

Fig. 6.9 The 32 back-end spreadsheets

Resources for the End User

The front-end resources in the Bibliography menu are easily accessed and understood, with further assistance from the User Guide. We are offering only brief descriptions below, with accounts of some interesting and challenging issues faced while making them.

Index of Manuscripts

The menu offers a ‘Full Table’ of all manuscripts, arranged by their shelfmark in the original collections to which they belong (Rabindra-Bhavana or Houghton Library, Harvard), with the (as yet) single private contribution at the end. The shelfmark in the first column provides a link to the image with concurrent transcript. The other columns list the contents by genre, while the last column indicates special features like the presence of Tagore’s celebrated doodles. As said above, this index is more comprehensive than any available in the libraries housing the manuscripts. It can be searched for particular entries using the CTRL+F or CMD+F function.

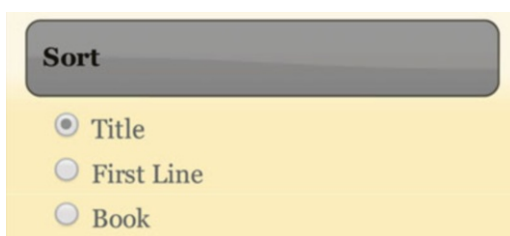
Integrated Checklist

The Full Table of Manuscripts is an invaluable bibliographical record. But few users are likely to know the shelfmarks of the manuscripts. They will want to search by title or first line. For this purpose, the tool of choice is the Integrated Checklist, covering both manuscripts and print publications. This is organized in two alternative ways, as an Alphabetical Index and a Full Table, to provide maximum information the way the user wants it.

Alphabetical Index

The Alphabetical Index is best suited to access individual works, grouped language-wise in the four standard divisions. The opening page (see Fig. 6.10) offers a choice of radio buttons, calling up the works by volume titles of collections (‘Book’), item title and (for poems and songs) first line.

Fig. 6.10 Alphabetical Index: Radio button options



The ‘Book’ option does not list all books, but only collections of short pieces. (Long book-length works like novels or long plays should be searched under ‘Title’ and not ‘Book’.) On choosing this option, clicking on a book title opens a table with full information about every piece in the collection. This table is organized like the ‘Full Table’ described below.






The ‘Title’ and ‘First Line’ buttons open a list of all items under that head, arranged alphabetically. Clicking on an item opens a pop-up window with full bibliographical data, arranged by formal categories (see Figs. 6.11 and 6.12). Clicking on any version opens up the image file of that version. For other material and facilities, like collation or a clear reading text, it is best to proceed via the Full Table.

The number of data fields or heads for these categories varies with the genre. The biggest number, 13, relates to poems and songs, as these were most often reprinted and anthologized, and embedded in longer texts as well as published separately. Songs called for extra entries about inclusion in *Gitabitan* and other song-books. Obviously, a novel or travelogue would not require these fields.

A minor problem concerned untitled numbered poems in a collection. These are entered under ‘titles’ like ‘Gitanjali 35’ or ‘Prantik 15’, while *Gitanjali* and *Prantik* can be accessed under ‘Book’. But ‘Gitanjali 35’ might have appeared with an actual title in a journal or anthology. These alternative titles can be located in the Full Table using the ‘Find’ command (CTRL+F or CMD+F), or in the List of Alternative Titles described above.

The use of numbered ‘titles’ threw up a bizarre alphabetical coincidence. As a rule, Tagore’s songs have no title, but are arranged numerically in various theme-based sections of Tagore’s collected songs, *Gitabitan*. As it happens, the three biggest sections all begin with the Bengali letter for *p*: *puja* (worship, i.e., spiritual songs), *prem* (love) and *prakriti* (nature). The songs in these three sections totalled nearly 1300, besides all other titles beginning with *p*. As a result, the *p* page in the Full Table was taking inordinately long to open. We finally split the *p* entries into four sections: one each for the three groups of songs, and one for all other titles (see Fig. 6.13).

Full Table

The other component of the Integrated Checklist is the Full Table of all works in a particular form or genre: Poems, Drama, Fiction or Non-Fiction. This contains all the items in that category, listed by title. (If you only know the first line, you can locate the entry in the ‘First Line’ column using CTRL+F or CMD+F.) As said above, the Full Table provides single-window access to all the material contained in Bichitra relating to a particular item. This includes, in the first instance, metadata arranged in columns (see Fig. 6.14). The icons against an entry open other windows with images of manuscripts  and printed works , clear reading text  and special information (‘Comments’) if any . The first column also carries the icon for the collation program .

The sprawling layout of the Full Table, too much even for a wide-screen terminal, called for some simple aids to the reader, like ensuring a static first column while the later columns scroll horizontally. This allows the user to check the data in, say, the 11th column against the title of the work. We also had to think of browsers using still narrower screens like a tablet or smartphone. That was one reason for

গীতাঞ্জলি-০০৫
✕

গণনাম	এস হে এস সঙ্গল বন
রচনাবলী	গীতাঞ্জলি, রবীন্দ্র-রচনাবলী, খণ্ড ১১ (বিশ্বভারতী, ১০৪৯) ঞ
গীতবিতান	৯৯, প্রকৃতি-বর্ষা, গীতবিতান (বিশ্বভারতী, ১০৮০) ঞ
পাঠ্যসিপি	RBVBMS_478 ঞ
পত্রিকা	আবাহন, মানসী (কার্তিক, ১০১৬) ঞ
গণনাম গ্রন্থ (কবিতা)	গীতাঞ্জলি (ইন্ডিয়ান পাবলিশিং হাউস, ১০১৭) ঞ
অন্যান্য গ্রন্থ (কবিতা)	গীতাঞ্জলি (বিশ্বভারতী, ১০০০) ঞ
গণনাম গ্রন্থ (গান)	নূতন গান, গান (ইন্ডিয়ান প্রেস, ১০১৬) ঞ
অন্যান্য গ্রন্থ (গান)	ধর্ম সংগীত (ইন্ডিয়ান পাবলিশিং হাউস, ১০২১) ঞ
	গীতিচর্চা (বিশ্বভারতী, ১০০২) ঞ
গণনাম সংকলন	গীতাঞ্জলি, কাব্যগ্রন্থ, খণ্ড ৮ (ইন্ডিয়ান প্রেস, ১০২০) ঞ
অন্যান্য সংকলন	গীতবিতান, খণ্ড ১ (বিশ্বভারতী, ১০০৮) ঞ
	প্রকৃতি-বর্ষা, গীতবিতান, খণ্ড ২ (বিশ্বভারতী, ১০৪৮) ঞ
অনুবাদ/অন্যান্য তথ্য	Come to me like summer cloud

Close

Fig.6.11 Alphabetical Index: pop-up window with bibliographical data fields (collections, manuscripts, journals, books, translations etc.) for a song in the Bengali *Gitanjali*

যোগাযোগ
✕

রচনাবলী	রবীন্দ্র-রচনাবলী, খণ্ড ৯ (বিশ্বভারতী, ১৩৪৮) ✎
পাঠ্যলিপি	RBVBMS_145(i) ✎
	RBVBMS_145(ii) ✎
	RBVBMS_145(iii) ✎
	RBVBMS_145(iv) ✎
	RBVBMS_145(v) ✎
	RBVBMS_145(vi) ✎
	RBVBMS_145(vii) ✎
	RBVBMS_145(viii) ✎
	RBVBMS_145(ix) ✎
	RBVBMS_145(x) ✎
পত্রিকা	তিন পুরুষ, দিৱস (১৩৩৫) ✎
প্রথম গ্রন্থ	প্রথম গ্রন্থ ✎
অন্যান্য গ্রন্থ	অন্যান্য গ্রন্থ ✎
প্রথম সংকলন	প্রথম সংকলন ✎
অন্যান্য সংকলন	অন্যান্য সংকলন ✎
অনুবাদ/অন্যান্য তথ্য	অনুবাদ/অন্যান্য তথ্য ✎

Close

Fig. 6.12 Alphabetical Index: pop-up window showing the serialized instalments of the novel *Jogajog* in the journal *Bichitra*

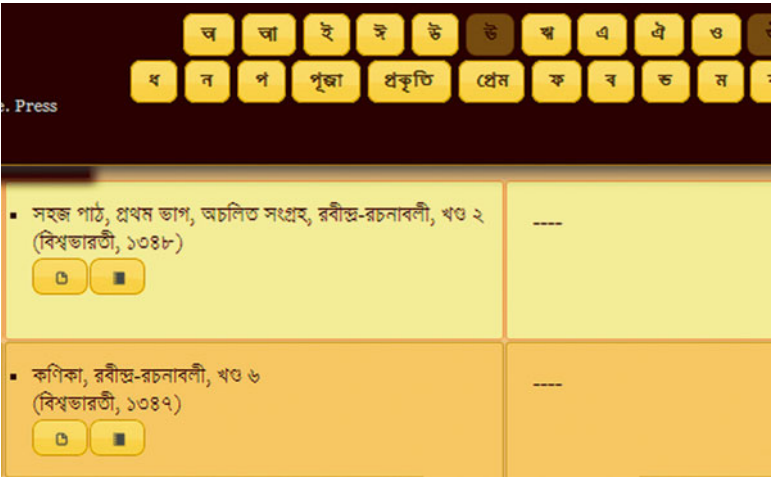


Fig. 6.13 Full Table: toolbar with the three additional buttons for p entries (words instead of letters)

reducing the number of columns in the Full Table as compared to the back-end spreadsheet, and instead providing access to some material through radio buttons.

The English Bibliography: From List to Master List

While going through the manuscripts and typescripts, we came across many interesting details adding value to a bare-bones listing, greatly benefiting the user. But the format of the bibliography had fairly rigid limits: it had to meet the encoding demands of the website software. We finally decided to include the extra details in a separate list, based on the bibliography but offering much more information. Given our constraints of time, it would have been too much to attempt such a list of the Bengali material. As a model of manageable size, we compiled an English Master List, prepared by Debapriya with help from Debapratim Chakrabarti.

Most of the additional data in the Master List falls into certain regular categories, but there are some interesting deviations. Whenever we came across little oddities, puzzles or interesting facts that might stimulate the researcher, we put it in the Master List. These included unpublished poems, needless to say; or marginal notes in manuscripts with a bearing on the authorship of a piece, or Das’s assertion in the *English Writings* that ‘All fruitless is the cry’ was the first English translation by Tagore. There are more complicated cases like the amalgamation of parts of five Bengali poems into the English ‘The Sunset of the Century’.

The major difference between the Bibliography and the Master List lies in the ‘Manuscript’ column and, at times, the ‘Comments’. They go into much greater detail about the intermediate variant stages of a text, as recorded in the manuscripts and typescripts. The Master List also includes some pieces not by Tagore, like English versions of the plays *The Post Office* and *The King of the Dark Chamber*, or

Title	First Line	Manuscript/ypscript	Journal	First Edition	Other Edition(s)	First Anthology Inclusion
A little flower blooms	• A little flower blooms	—	• The Modern Review, November 1913 ★ □ □ ■	—	—	—
A Peace Hymn from the Atharva Veda	• Peaceful be all motives	—	• The Modern Review, 1915 □ □ ■	—	—	—
A Poxy (Love's Gift 31) st	• My flowers were like milk and honey and wine	□ □ □ • RRBMS_077 □ • RRBMS_305A □	—	• Love's Gift and Crossing (London: Macmillan, 1918) □ □ ■	—	—
A weary pilgrim, I travel	• A weary pilgrim, I travel	• RRBMS_005 □	• The Modern Review, August 1929 • Vine-Bharati Quarterly, New Series Vol. X, part II, August-October 1944	—	—	—
A' troupe of young things' sing (Song from the Cycle of Spring) st	• Again and again we had said	□ □ • EMSE_007 □	• The Modern Review, February 1916 ★ □ □ ■	—	—	—
Ah, let it sound in my heart	• Ah, let it sound in my heart	• RRBMS_305A □ • RRBMS_306(ii) □ • RRBMS_404	• The Vine-Bharati Quarterly, Vol. 44, nos. 3 and 4	—	—	—

Fig. 6.14 Full Table: screenshot with column headings and icons

Table 6.5 A sample manuscript column entry in the English Master List

Title/first line	Manuscript/typescript
I hid myself to evade you	RBVBMS_059(i) ts IMG 28. RBVBMS_059(ii) ts IMG 28. RBVBMS_060 IMG 14 'I hide myself to evade you'. RBVBMS_308A ts IMG 34 'Yes, yes, strike me again, yet again.' RBVBMS_309A ts IMG 26 'I h[a]{i}d myself'. RBVBMS_369(ii) ts IMG 25, IMG 26 'Yes, yes, strike me more, yet more'. RBVBMS_446 ts IMG 62 [1st line as above. Contains ts and ms notes on provenance on IMG 1. Original versions of the Bengali translations by Tagore, done in Chicago, which were later shortened by Tagore for publication. Contains Index of poems from IMG 2-4]. HRVD_004 ts IMG 52 'Yes, yes, strike me [more] {again}, yet [more] {again}.' [IMG 105: 'Gitanjali - part II with corrections'].

doubtfully by him, like some translations from the medieval Hindustani mystic poet Kabir. As a rule, Bichitra omits such works.

The Master List is divided by genre, and repeats the seven column headings of the first preliminary spreadsheet described under ‘Organizing the Bibliography’ (see Table 6.2). But given the crucial importance for the Master List of manuscript readings, it gives the exact image reference, and cites the opening words of every variant version. Comments in square brackets provide editorial notes about special features if any. This is particularly useful where the first line of a printed version differs markedly from the manuscript reading. Thus the ‘Manuscript’ column for the printed poem beginning ‘I hid myself to evade you’ looks as shown in Table 6.5.

To help access, poems have been entered twice in the Master List: by first line with full details, plus a cross reference under the title. Similarly, essays are fully documented as separate items, and cross-referenced under the collection to which they belong. Poems that are closely related, or that spring from the same Bengali original, have been cross-referenced too.

The genetic history of Tagore’s writings often takes an intricate, unconventional course. His translations and other English writings pose special bibliographic problems which have received little attention so far. It seemed worthwhile to record these complexities in the website: the Master List, devoid of hyperlinks, provided the best compromise for conveying the information without greatly complicating the site structure. With the Bengali works, eminent Tagore scholars have devoted their lives to tracing the genetics of the text. Here the challenges are truly formidable, as regards both data collection and metadata management. The English Master List provides the model for a far bigger exercise with the Bengali works. We hope the latter task will be taken up soon. Given the funds and opportunity, we would love to do it ourselves.

Timeline

This is a postscript about a postscript. After the site had been officially completed and launched, users alerted us to the need for a clearer projection of Tagore’s chronology without having to dig out individual dates from the bibliography. We

therefore compiled the dates into an integrated Timeline, and Ritwick, our long-suffering web designer and webmaster, was persuaded to add it to the site he thought he had completed.

There are in fact two Timelines: one by year for volume-form publications, the other by month for journal publications. (The timeline of composition would have been impossible to compile.) The first starts at 1878, when the 17-year-old Tagore published his first book-length work, a long poem called *Kabi-kahini* (*The Story of a Poet*). (It had appeared in the journal *Bharati* a few months previously.) The journal Timeline goes back to 1874, to a piece in the *Tattvabodhini patrika* credited to 'a 12-year-old boy'. Both Timelines extend long beyond the poet's death in 1941, indeed to the present day, to take in works published posthumously for the first time.

Within each Timeline, entries are grouped as usual by the four main genres, Poems and Songs, Drama, Fiction and Non-Fiction. One can click on a particular (Bengali) month or year from the opening menu to open a pop-up window showing all publications for that date, or scroll horizontally from one year to another by clicking on the buttons flanking the pop-up. To search by title, one can open a drop-down menu at the top of the page and choose first the genre, then any title belonging to that genre.

The Timelines have their own back-end spreadsheets, separate ones for books and journals. They are two of the 20 basic back-end spreadsheets underlying the Bibliography, extracting their data and sorting it by month and year (see Figs. 6.15 and 6.16).



Fig. 6.15 Timeline for books, arranged by Bengali years. Drop-down search menu at top



Fig. 6.16 Timeline for journal publications, arranged by Bengali months. Drop-down search menu at top

References

- Bagchi, Sanatkumar. 1989. *Rabindranather pandulipi: samiksha o bishleshan*. Kolkata: Pustak Bipani.
- British Library. *Endangered Archives Programme: Guidelines for photographing and scanning archival material*. http://eap.bl.uk/downloads/guidelines_copying.pdf. Accessed 12 June 2014.
- Das, Sisir Kumar ed. 1994–2007. *The English Writings of Rabindranath Tagore*. Vols.1–3. Vol. 4 ed. Nityapriya Ghosh. New Delhi: Sahitya Akademi.
- Fauland, Alex. 1999–2013. Tools and Utilities: A.F.5 Rename your files. http://www.fauland.com/af5_dl.htm. Accessed 22 December 2014.
- Ghosh, Sankha comp. 1408/2001. Guide to publications (Granthaparichay). In *Rabindra-rachanabali*. Vol.16. Kolkata: Government of West Bengal.
- Joejoe. No date. Joejoe's Freeware Utilities: Rename Master. <http://www.joejoesoft.com/cms/showpage.php?cid=108>. Accessed 22 December 2014.
- Majumdar, Swapna. 1395/1988. *Rabindra granthasuchi*. Vol.1 pt.1. Kolkata: National Library.

- Pal, Prashantakumar. 1389/1982. *Rabijibani*. Vol. 1. Kolkata: Papyrus. 1391/1984 – [1410]/2003. Kolkata: Ananda Publishers.
- Rabindra-rachanabali*. 1346/1939–. 32+2 vols. published so far. Kolkata: Visva-Bharati Publishing Division.
- Sen, Pulinbihari. 1416/2009. *Pulinbihari: janmashatabarshik shraddharghya* (includes various bibliographies compiled by Pulinbihari Sen). Kolkata: Visva-Bharati Publishing Division for Rabindra-Bhavana.
- SNLTR. 2009. Society for Natural Language Technology Research. www.rabindra-rachanabali.nltr.org. Accessed 22 December 2014.
- Tagoreweb. 2010–2012. <http://www.tagoreweb.in>. Accessed 22 December 2014.

Sukanta Chaudhuri, Dibyajyoti Ghosh, Prakash Koli Moi,
and Arabinda Moni

One of the major advantages of an electronic text is that it can accommodate a search function. Given the vast corpus of Rabindranath's writings, a search function is specially necessary. But instead of a simple search engine that just yielded the names of the files (i.e., the titles of the works) in which the search term appears, we wanted a search program closer in user experience to Internet search engines like Google, Yahoo or Bing. In major search engines like these, clicking on a search result directs the user to that result itself, which can be a webpage or a file. In the context of textual computing, such a program would be called a hyperconcordance.

In literary scholarship, a concordance is a complete list of all the words found in an author, a work or other literary corpus, with the immediate textual context in which a word occurs. It also tallies the total number of occurrences, thereby indicating what words are important in the corpus. In pre-computer times, concordances were laboriously compiled of a few canonical works (first and foremost the Bible) or major literary figures like Shakespeare. In the mid-nineteenth century, Mary Cowden Clarke took 16 years to compile a Shakespeare concordance (Clarke 1845). Today, the availability of electronic text files has transformed the scene. In 2003–2004 a young Master's student, while serving as a Marine in Kuwait, took less than a semester to create a digital Shakespeare concordance for his Master's project (Johnson 2003). On an electronic database, the results of a comprehensive set of simple searches constitute a concordance. If these results are hyperlinked to the source texts and metadata, we have a hyperconcordance. This is the case with the Bichitra search function.

S. Chaudhuri (✉) • D. Ghosh

Department of English, Jadavpur University, Kolkata, India

e-mail: schaudhuri@english.jdvu.ac.in; ghosh.dibyajyoti@gmail.com

P.K. Moi • A. Moni

Department of Computer Science and Technology, Jadavpur University, Kolkata, India

e-mail: prakashkolimoi@gmail.com; arabindamoni@gmail.com

But if this was our goal and the Internet search engine our model, why did we create a separate engine, instead of applying a generic, freely available search algorithm like those provided by Google, which can be incorporated in any website? The answer is that the latter course would have interfered with some of our other purposes. One was structural: some of our text files (which absolutely required to be kept as such to suit the general design of the whole project) would have to be converted into web pages. More crucially for the end user, we could not have shown the total number of occurrences of a search term; nor could we have incorporated both the genre-wise division and, within it, the subdivision into titles.

As in most other cases, we found that ready-made solutions would not meet our needs. We had to go the hard way and work out our own strategy.

Basic Strategy and Rationale

By one of the many lucky coincidences that favoured the Bichitra project, Arabinda and Prakash, then undergraduates in Jadavpur's Computer Science Department, had created a simpler search program of the kind we had in mind as part of their course work, guided by their professor Chandan Mazumdar. Sukanta learnt of this from one of his confabulations with Chandan, and thought we could make a joint venture of it. Arabinda and Prakash thus took up the task of creating a hyperconcordance. They developed the engine in consultation with members of the regular Bichitra team like Dibyajyoti and Purbasha, under Sukanta's overall guidance.

The sum total of Rabindranath's writings contained in the website amounts to 33,784 text files, with each individual version of each individual title, be it a poem, play, essay or work of fiction, constituting one or more such files. (For the search engine, long works had to be divided into several files—e.g., the chapters of a novel or scenes of a play.) But we felt that to include all versions or editions of a work in the search would cause a huge amount of unnecessary duplication, as a word would be repeated for its occurrence in each separate version. This would greatly slow down the search process, and not materially benefit the users, who would have to plough through a great deal of superfluous material. So we confined the search to a single version of each work. For Bengali, this was usually the *Rachanabali* or Collected Works brought out by Tagore's own university, Visva-Bharati, though the songs were accessed from a more authoritative posthumous edition (1380/1974) of *Gitabitan*, the collected songs. For the English works, we used the first extant printed edition. The *English Writings* (Das 1994–2007) published by Sahitya Akademi, Delhi, though sizeable, has too many gaps. We thereby reduced the number of files to roughly 8500, or just over 25 % of the total number.

We adopted another strategy to streamline operations. A search function uses a good part of the computer's processing resources. Hence we needed to have a much reduced real-time search operation, which would not involve looking through all 8500 files. In that case, if ten users tried to access the search engine at the same time, it would involve a very high waiting time for each of them. Obviating this would call for a very high-end server, which was not only beyond our means but not justifiable by the total requirements of the site.

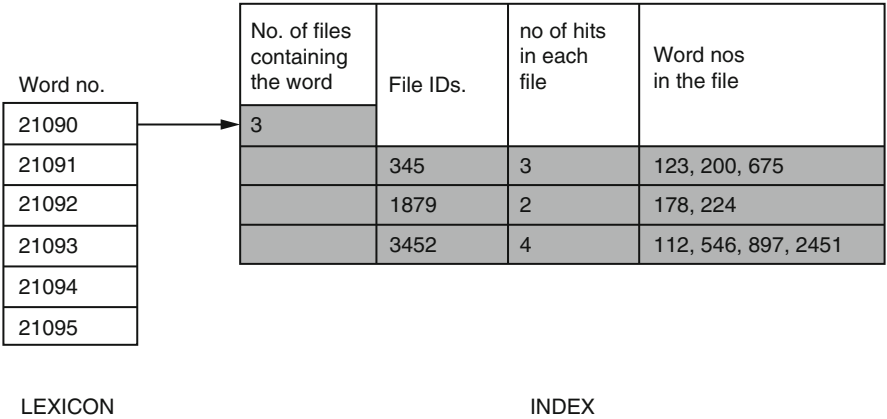


Fig. 7.1 Structure of search engine lexicon and index

Instead, we decided to create an index which would be searched rather than the actual text files. The index was created offline and uploaded to the website, instead of being re-created every time an end user performed a search query: this too reduced search time, as also the risk of malfunction or delay owing to server problems. The index assigned an ID to every unique word in every document, and stored the number of times each word occurred in the entire database. It also stored the address of the text files in which the word occurred, and the position of the word in these text files (see Fig. 7.1).

The end user needed a little more help as well. When dealing with 33,784 text files (not to mention nearly 140,000 image files and millions of collation files), file names cannot be descriptive. The core filename had to be short, and formed according to a code intelligible to the computer. To the human user, they would appear to be gibberish. In order to describe the text file to the end user (and also to help the human project workers dealing with those files), a metadata sheet was created by Purbasha for the Bengali files, and by Debapriya for the English files. The process is fully described in Chap. 6.

Showing the Output

The user enters the search term (one or more words) in the appropriate box. Bengali terms can be entered in either Bengali or Roman letters following Avro keyboard practice: the latter will automatically be transformed into the Bengali word on pressing the Enter key.

A tip: To retain the Roman spelling, press CTRL+M or CMD+M before keying in. This is needed to look for words in the Roman alphabet embedded in the Bengali text.

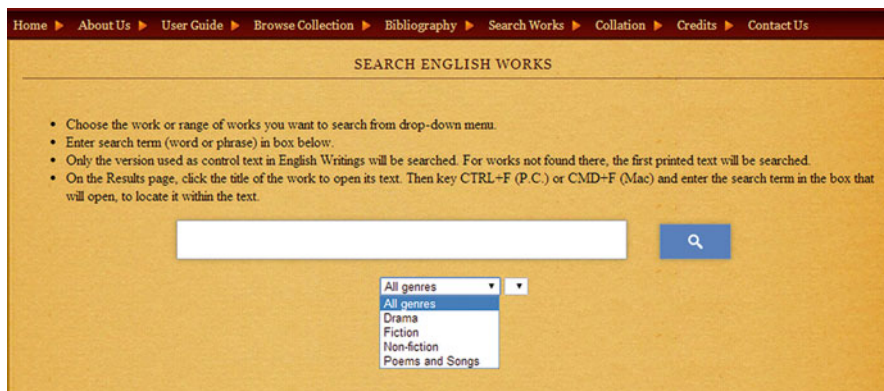


Fig. 7.2 Search: opening page

One can do specific genre-wise, collection-wise and title-wise searches. A drop-down menu allows a choice of ‘All Genres’, all works in a particular genre (poems, plays, fiction or non-fiction) or a single title within a genre (see Fig. 7.2). This genre-wise and title-wise search is made possible by the categorizing of file-name codes and creation of separate metadata lists.

Drawing on the information in the metadata sheet, the search output display cites

- The total number of occurrences of the search term.
 - The title of the work(s) where the search term appears.
 - If the source text is a short poem, song or essay, the collection to which it belongs.
 - If the source text is a poem or song, its first line.
 - If the source text is a novel or play, the chapter number or act and scene number.
- The text was split into separate files for this purpose.
- Bibliographical details of the source.

(See Fig. 7.3.) Moreover, the results display cites the entire line containing the query term, so that users can judge from the context which results are most relevant to their needs. Within this context, the query term is displayed in bold type. Results are usually displayed ten to a page, though the number may vary depending on the size of the entries.

For a fuller context, the user can click on the hyperlinked bibliographical data to open the text file with the full text of the work. By then applying the standard ‘Find’ command (CTRL+F in Windows and Linux, CMD+F in Mac), one can locate the word(s) in the text—an indispensable help when it occurs in a long text (see Fig. 7.4). For the ‘Find’ command, the search function is taken over completely by the browser, hence Roman letters cannot be transliterated into Bengali: to find a Bengali term, one must enter it in Bengali letters.

To ensure search engine optimization (SEO), the search results need to be prioritized, placing the most salient ones first in the list. With single-word searches using

SEARCH ENGLISH WORKS

Search

Go To Page

All genres

13 Results found in 0.022 sec

[Visva-Bharati Quarterly | A Vision of India's History | When individual communities, who come to dwell in the same neighbourhood | Visva-Bharati Quarterly, 1923](#)

Number of occurrences = 1

Hindu India has been placed by its birthright Where the harmony between the component differences....

[Gitanjali | Gitanjali 35 | Where the mind is without fear | Gitanjali \(Song Offerings\) \(London: The India Society, 1912\)](#)

Number of occurrences = 2

35 WHERE the mind is without fear and the head.... dreary desert sand of dead habit Where the mind is led forward by thee into....

[Red Oleanders | Red Oleanders \(London: Macmillan, 1925\)](#)

Number of occurrences = 12

have heard it and will bear it in mind He goes Chandra Ah now didn.... mad girl when I am with you my mind scorns to be cautious Nandini No.... Really So recklessly that you don't even mind confessing it Bishu Sir Governor it.... s attraction like the tidal wave tears away mind from its anchorage of books Antiquarian.... disease not of the body but of the mind Governor What s the remedy Doctor.... it never had before to have to never mind you may go I'll think.... can be had to draw the car Never mind we can press the diggers into.... not to speak of extras Your Lordship s mind is like that of the gods.... No for whatever his blood may be his mind in a sense is really pious.... Governor That may be Fortunately for us our mind knows not its own secret Come....

Fig. 7.3 Search: results page, showing all results for 'where the mind' in all Tagore's works

SEARCH ENGLISH WORKS

35 1 of 2

WHERE the mind is without fear and the head is held high;
Where knowledge is free;
Where the world has not been broken up into fragments by narrow domestic walls;
Where words come out from the depth of truth;
Where tireless striving stretches its arms towards perfection;
Where the clear stream of reason has not lost its way into the dreary desert sand of dead habit;
Where the mind is led forward by thee into ever-widening thought and action—
Into that heaven of freedom, my Father, let my country awake.

Fig. 7.4 Full text of the poem 'Where the mind is without fear' opening from Search results page, with search term (in box, top right) highlighted in the text

the Bichitra search engine, the work with the maximum number of occurrences comes first on the list, and so in descending order. For a string of words, the instances where all the words occur (not necessarily as the same string) are placed first; the ones with fewer words follow in descending order of relevance (see Fig. 7.3).

Inflexions and Variant Spellings: Longing for a Lexicon

There is one problem we could not overcome, given the still imperfect resources for computing in Bengali. Unlike standard search algorithms like, say, Google's, we could not provide for searching correct alternatives to misspelt words, or undertaking semantic searches. This is basically owing to the absence of lexicons in Bengali (as in most other Indic languages, indeed most other languages). Hence if the user misspells the search term, the Bichitra search engine (like all others in Bengali) will only search under that wrong spelling and return a zero result: there is no lexicon to search for possible correct alternatives to the misspelt term. This is specially problematic where the 'misspelling' is a viable alternative spelling, of the kind abounding in Bengali. Tagore's spelling varies from edition to edition, work to work, and from one period of his life to another. The only way of ensuring a full search is by searching for all variant forms of the word.

There is, however, a safeguard when searching for a combination of words. If one word is spelt wrongly and hence unrecognized, the search engine will search for the other words in turn. This ensures a strong chance of locating the combination.

Another problem area is where the search term occurs in inflected form, with a case ending or plural ending, and perhaps some change to the stem. As Bengali is a highly inflected language, this is often the case. In the absence of a lexicon listing variant spellings and grammatical forms, the search engine will only look for the form entered by the searcher. Again, a full search would require all the possibilities to be entered one by one.

This is one of many arguments for the imperative need of a Bengali lexicon for word-processing and text-processing programs. That must be a separate major project: it could not be an ancillary to the task of creating Bichitra.

References

- Clarke, Mary Cowden. 1845. *The Complete Concordance to Shakespeare*. London.
- Johnson, Eric M. 2003. Open Source Shakespeare: An Experiment in Literary Technology. George Mason University. http://www.opensourceshakespeare.org/info/paper_toc.php. Accessed 22 December 2014.
- Das, Sisir Kumar ed. 1994–2007. *The English Writings of Rabindranath Tagore*, vols.1–3. Vol.4 ed. Nityapriya Ghosh. New Delhi: Sahitya Akademi.

Sukanta Chaudhuri, Spandana Bhowmik,
and Sunanda Bose

The head of the Bichitra project recalls how, at the very start of his very first computer class, the instructor told them: ‘Remember the computer is an idiot. It doesn’t understand anything. It can only do what you tell it to do.’

This is even more relevant in 2015, when the computer seems so much more intelligent. Its alleged powers were put to a severe test when we began working on collation programs—that is, programs to compare variant versions of a work—in the School of Cultural Texts and Records some 8 years ago. How would the simple zero and one of the binary system mesh with the subtle variations in literary texts?

Metaphorically speaking, we had to start even before zero, as we were working with a non-Latin font. Those who do not can have no idea of the additional challenges this poses. The problems concerning fonts and keyboards have been described in Chap. 3. Also, the English or Roman alphabet has the benefit of many analytic tools, developed over the years and available on the Internet. With Bengali, as with almost all non-Latin alphabets, if there are any such tools at all, they are hard to find and often imperfect. But we set ourselves a bigger challenge than demanded by our immediate need. We wanted to develop programs that would work with UTF-8 files in any language.

The Tagore corpus comprises two languages, Bengali and English. We had to keep all other possible languages in mind. Our collation program **Prabhed** was by

S. Chaudhuri (✉)

Department of English, Jadavpur University, Kolkata, India

e-mail: schaudhuri@english.jdvu.ac.in

S. Bhowmik

India Foundation for the Arts, Bengaluru, India

e-mail: spandana139@gmail.com

S. Bose

School of Mobile Computing and Communication, Jadavpur University, Kolkata, India

e-mail: neel.basu.z@gmail.com

far the most complex of all the programs developed in-house for the Bichitra project. It is perhaps the first text-processing software of any kind to have been written in the first instance for a non-Latin font and later extended to the Roman.

It was not, however, our first sally into collation programs. We may start by recounting our earlier ventures in this regard. We will not even try to talk about other programs like the excellent and indispensable Juxta, created by the University of Virginia and generously offered by them for all to use (as we hope to offer ours once we have solved some residual problems); the TEI-inspired Versioning Machine; or the immensely elaborate and versatile Collate, now housed at the University of Saskatchewan, perhaps the only collation program of comparable scope to Prabhed though with a different set of functions. But Prabhed is the only one that offers full multi-level collation of various strata of the text.

First, the prehistory.

The Road to Prabhed

Tafat (Version 1)

‘Tafat’ means ‘difference’. Tafat 1.0 was developed and gifted to the School by Siddhartha Chaudhuri, then finishing his time at the Indian Institute of Technology, Kanpur. This could handle Unicode texts in three alphabets, Bengali, Devanagari (used *inter alia* for Hindi), and Roman, saving the results as plain text or linked HTML files. It showed up, in colour-coded form, all additions and deletions; all replacements of one word by another; and transpositions or changes in position. Its one major shortcoming was that (like all collation programs of the day except Collate) it could only present results for two versions at a time.

Using this version, we began to collate the variant Bengali texts of the play *Bisarjan* (*Sacrifice*). We deliberately chose this work in view of its many widely variant versions, ranging in date from 1890 to 1939, the longest ten times the size of the shortest. We took eight versions into account. We noted two problems in course of our work. Firstly, transpositions over more than a certain span could not always be recorded correctly. Secondly, Tafat 1.0 could not tell between sentence endings, which divided units of prose, and line endings or verse endings, which provided a more relevant division between units of verse.

Pathantar

Tafat 1.0 was a great start, but it clearly needed a lot of fine-tuning. We also wanted collation results for more than two versions at a time. We deliberately say ‘wanted collation results’, not ‘wanted to collate’. To compare more than two versions at a time calls for a completely different line of computer logic: it seems safe to say that no one has yet cracked the problem, at least not to the point of creating a viable program. Basically, then, we had to compare two texts at a time—each of them against all the others in turn—and merge the results so expeditiously that the user

would effectively obtain collation results for n texts. This is the principle behind all the collation programs we have developed so far and, to the best of our knowledge, all other working programs that can present collation results across multiple texts.

We began to develop a more elaborate collation program, on a completely different basis from Tafat, even while working on *Bisarjan*. The new program was called 'Pathantar' (Textual Variants), and was developed by Sujoy Sengupta of a small but innovative Kolkata firm, Synapse Technologies. Pathantar stored the full collation results, after comparing all the versions, in two files: a preliminary HTML file and an XML file, the first showing the interface displaying the results and the second serving as a repository for the variant readings.

Pathantar was more informative than Tafat. Like Tafat, it prepared a linked HTML file from the base text. By clicking on hyperlinks in the base text, it not only showed the variants in other versions in a separate frame, but displayed the whole sentence (for prose) or line (for verse) in which the variant occurred. If one text lacked some word(s) found in another version, it would show a hyperlinked legend 'GAP' at that point. It could also show gaps or caesurae within a line, as well as tabs and punctuation marks: the last could be optionally displayed or suppressed.

Pathantar could distinguish between prose and verse, and collate the two differently. When loading the text files into the program, one could choose the sentinel symbol or delimiter: the full stop or equivalent for prose, the line break (new line character) for verse. But Pathantar could not automatically distinguish prose from verse. If a single section of text (say, a scene in a play or a chapter of a novel) contained both verse and prose, the verse and prose segments had to be manually divided, and each part treated as a separate text block.

Within the line or sentence, Pathantar began by comparing the first words of the compared versions. If these did not match, it compared the last words. On that basis, it determined whether the lines/sentences were the same. A line/sentence appearing in one but not in another version, or vice versa, was indicated by the legends 'Extra Line' or 'New Line'.

Pathantar also allowed one to adjust the 'transposition boundary' within a long work, to determine how far forward the program would search to find a match. The default boundary was four prose sentences or verse lines.

We completed the *Bisarjan* project using Pathantar instead of Tafat. But before publishing it in CD form, we had to devise a more user-friendly interface than the 'raw' output file. We got in touch with our colleagues at the School of Education Technology at Jadavpur University, whose young instructor Arunashis Acharya prepared an attractive platform with a four-window display (see Fig. 8.1). The base text was displayed in the top left window, and the same text as a result-linked HTML file in the top right. On clicking any link in the latter, the lower right window displayed the variant readings in the other versions: not just the salient words but the whole line or sentence, each version separately colour-coded. The lower left window was reserved for displaying a reference document—that is to say, any one of the other versions, selected from a drop-down menu, that we may wish to look at and compare with the base document. Later, we adopted much the same four-window interface for displaying the fine collation or word-by-word comparison results in Prabhed, the collation program used in Bichitra.



Fig. 8.1 Pathantar: four-window display

Behind this display, the output folders containing the results were stacked along with the set of text files used to arrive at the results. Arunashis created an ‘input interface’ whereby the relevant results and text files could be drawn as necessary to show in the four-window display. Behind these again was a manually created table of the parts into which the text had been divided (see Fig. 8.2). The cells in this table did not always correspond to acts and scenes: as explained above, the verse and prose parts within each scene had to be placed in separate text blocks. The output folders were linked to this table, and in turn fed the four-window display.

Preparing this table manually was a long and arduous task: a single prose scene with interspersed songs might require a dozen or more divisions. Such a manual operation was obviously impossible for the vast corpus of all Tagore’s works. We explain below how we set about solving this problem when we embarked on Bichitra.

Bisarjan consists chiefly of verse. So when developing Pathantar, our focus lay almost unthinkingly on verse, not prose. With *Bisarjan*, Pathantar worked well enough once we had divided the text manually into separate prose and verse sections. In fact, if collating short verse texts, Pathantar is better in some ways than Prabhed, the program used for Bichitra. But on applying Pathantar to prose, we faced some major problems.

First of all, when comparing long sentences, simply checking for matches in the first and last words is often inadequate. Even if we take the first few or last few words, a couple of stray words inserted among them in one version will confuse the process. A synthetic or inflected language like Bengali faces another big problem in this respect. A mere change of inflection will make the same word register differently. This can sometimes happen even if the grammatical function remains the same, rather like alternative prepositions in English: ‘He arrived at London’ and ‘He arrived in London’.

১২৯৭	১৩০৩	১৩০৬	১৩২২	১৩৩৩	১৩৩৮	১৩৪৬	১৩৪৩
উৎসর্গ		উৎসর্গ (মুদ্রণে পাত্রগণ-এর পরে)	উৎসর্গ	উৎসর্গ	উৎসর্গ	উৎসর্গ	
পাত্রগণ	পাত্রগণ	পাত্রগণ (মুদ্রণে উৎসর্গ-এর পরে)	পাত্রগণ	পাত্রগণ	পাত্রগণ	পাত্রগণ	পাত্রগণ
১/১ক তুর্ক থেকে 'গোবিন্দ (নেপথ্যে চাহিয়া) জয়সিংহ'				১/১ক তুর্ক থেকে 'গোবিন্দ (নেপথ্যে চাহিয়া) জয়সিংহ'			
১/১খ 'জয়সিংহের প্রবেশ। জয়। কি আদেশ' থেকে 'হুজ্জ আসি একবার। অপর্ণা জয়সিংহে প্রস্থান'	১/১খ 'গোবিন্দমাণিক্য অপর্ণা ও জয়সিংহের প্রবেশ। জয়। কি আদেশ মহারাজ।' থেকে 'কিরাব কেমনে?'	১/১খ 'গোবিন্দমাণিক্য অপর্ণা ও জয়সিংহের প্রবেশ। জয়। কি আদেশ মহারাজ।' থেকে 'কিরাব কেমনে?'	১/১খ 'গোবিন্দমাণিক্য, অপর্ণা ও জয়সিংহের প্রবেশ। জয়সিংহে। কি আদেশ মহারাজ।' থেকে 'কিরাব কেমনে?'	১/১খ 'জয়সিংহের প্রবেশ। জয়সিংহে। কী আদেশ' থেকে 'হুজ্জ আসি একবার। অপর্ণা জয়সিংহে প্রস্থান'	১/১খ 'গোবিন্দমাণিক্য, অপর্ণা ও জয়সিংহের প্রবেশ। জয়। কি আদেশ মহারাজ।' থেকে 'কিরাব কেমনে?'	১/১খ 'গোবিন্দমাণিক্য, অপর্ণা ও জয়সিংহের প্রবেশ। জয়সিংহে। কি আদেশ মহারাজ।' থেকে 'কিরাব কেমনে?'	
১/১গ 'গোবিন্দ। এখনো এলনা' থেকে 'রক্ত মুছে ফেলি + SD'				১/১গ 'গোবিন্দ। এখনো এলো না' থেকে 'রক্ত মুছে ফেলি + SD'			
১/১ঘ 'অপর্ণা জয়সিংহের প্রবেশ। অপর্ণা। মা তাহার নিরেছেন' থেকে 'কে বলিয়া দিবে মোরে'	১/১গ 'অপর্ণা। মা তাহারে নিরেছেন' থেকে 'কে বলিয়া দিবে মোরে'	১/১গ 'অপর্ণা। মা তাহারে নিরেছেন' থেকে 'কে বলিয়া দিবে মোরে'	১/১গ 'অপর্ণা। মা তাহারে নিরেছেন' থেকে 'কে বলিয়া দিবে মোরে'	১/১ঘ 'অপর্ণা ও জয়সিংহের প্রবেশ। মা তাহারে নিরেছেন' থেকে 'কে বলিয়া দিবে মোরে'	১/১গ 'অপর্ণা। মা, তাহারে নিরেছেন' থেকে 'কে বলিয়া দিবে মোরে'	১/১গ 'অপর্ণা। মা তাহারে নিরেছেন' থেকে 'কে বলিয়া দিবে মোরে'	
১/১ঙ 'জয়। মাতারে কোরোনা দোষী' থেকে 'দণ্ড লব নিজ শিরে + SD'							
	১/১ঘ 'অপর্ণা। এই যে সোপান বয়ে' থেকে 'করিয়ছি পণ + SD'	১/১ঘ 'অপর্ণা। এই যে সোপান বয়ে' থেকে 'করিয়ছি পণ + SD'	১/১ঘ 'অপর্ণা। এই যে সোপান বয়ে' থেকে 'করিয়ছি পণ + SD'	১/১ঙ 'অপর্ণা। এই যে সোপান বয়ে' থেকে 'করিয়ছি পণ + SD'	১/১ঘ 'অপর্ণা। এই যে সোপান বয়ে' থেকে 'করিয়ছি পণ + SD'	১/১ঘ 'অপর্ণা। এই যে সোপান বয়ে' থেকে 'করিয়ছি পণ + SD'	
১/১চ 'হাসি। এইবার সব' থেকে 'আজ মন্দিরেতে নয়। + SD'				১/১চ 'হাসি। এইবার সব' থেকে 'আজ মন্দিরেতে নয়। + SD'			
১/১ছ 'ভাবতীর প্রবেশ। গণবতী। মার কাছে কি করেছি' থেকে 'পূজার সময় হল। + SD'	১/১ক তুর্ক থেকে 'পূজার সময় হল। + SD'	১/১ক তুর্ক থেকে 'পূজার সময় হল। + SD'	১/১ক তুর্ক থেকে 'পূজার সময় হল। + SD'	১/২ক তুর্ক থেকে 'পূজার সময় হল। + SD'	১/১ক তুর্ক থেকে 'পূজার সময় হল। + SD'	১/১ক তুর্ক থেকে 'পূজার সময় হল। + SD'	

Fig. 8.2 Manually created table of text divisions for Pathantar

These problems could be winkled out by fine-tuning the program, but thereby greatly complicating the code. On the other hand, going by anything other than an exact match would create a set of new problems, as totally different words would then appear to match.

Setting the transposition boundary (i.e., the length of text that the program would search to find a match) was another problem. If you set the boundary too closely, you might miss a genuine match if the relevant text has moved by a good distance; if too broadly, you are likely to get false matches, especially where the text repeats itself at intervals as in the refrains of poems and songs, or in common stage directions like ‘Enter’ and ‘Exit’. Such repetition is a general problem, which we have not quite eliminated even in the later and different package, Prabhed, used in Bichitra. We believe we might solve it in a new program now under development.

As for the more general problem of transposition boundaries, it was insoluble beyond a soon-reached point in single-tier collation programs like Pathantar (and most others). Translocations beyond a particular point could only be detected by manual checking. To do so on any scale would obviate the very advantage of computer collation. To do so through the huge corpus of Tagore’s works would take years and be extremely error-prone.

We prepared an experimental CD of *Bisarjan* and a more finished DVD of Tagore’s collection of lyrics *Sonar tari* (*The Golden Boat*) using Pathantar. The *Sonar tari* DVD required a table of all the items, with icons providing links to various resources like image, clear text and collation (see Fig. 8.3). This provided an early prototype for the full bibliographical table that we finally prepared for Bichitra.

But we were already thinking ahead on very new lines.















































































































































































































































































	সংখ্যা	পরিচালনা পট	শোনার তরী ১০০০	কাব্যগ্রন্থাবলী ১০০০	কবিতা-এক ১০১০	চরিত্রিকা ১০১৬	কাব্যগ্রন্থ ১০২২
সোনার তরী	০১	  	  	  		  	  
বিদ্যবতী (কল্পকথা)	০২	  	  	  			  
শৈশব সন্ধ্যা	০৩	  	  	  			  
জাহাজ হেসে ও জাহাজ মেয়ে (কল্পকথা)	০৪	  	  	  			  
মিথিতা	০৫		  	  	  		  
সুপ্তেখিতা	০৬		  	  	  		  
বোম্বার এবং আমবা	০৭	  	  	  			  
সোনার কীল	০৮	  	  	  			  
বর্ষা যামল	০৯	  	  				  
হিং টিং ছট (কল্পকথা)	১০	  	  	  			  
পরাণ-পায়র	১১ক	  	  	  	  	  	  
	১১খ	  	  				  
	১১গ	  	  	  	  	  	  
বৈষ্ণব-কবিতা	১২	  	  	  		  	  
দুই পক্ষী	১৩	  	  	  			  
অকালের টান	১৪	  	  	  	  		  
পানভঙ্গ	১৫	  	  	  	  		  
যেতে নাই দিব	১৬	  	  	  		  	  
সমুদ্রের প্রতি	১৭	  	  	  		  	  
প্রতীক্ষা	১৮		  	  			  
মাসল-সুন্দরী	১৯		  	  			  

Fig. 8.3 Manually created table for the poetical collection *Sonar tari*. Each column represents a version of the text. The three icons carry links for text, image and collation

Gross and Fine Collation: Early Thoughts

This was when we first started thinking of collating at two levels: **gross collation**, or comparison of relatively large blocks of text, and **fine collation**, or detailed comparison of individual words and punctuation. A gross collator would computationally generate the full structural design of the work, comparing its various versions. As said above, the laborious task of manually compiling a structural table for a single play, *Bisarjan*, could not possibly be extended to Tagore's complete works. It was imperative to delegate this work to the computer by means of a suitable program.

Among many other advantages, a gross collator would solve the problem of translocation across large spans of text. We commissioned a 'gross collator add-on' for Pathantar, by comparing the gross word-match percentage in the compared passages. For reasons we never quite fathomed, this could not be made to work.

Yet up to that point, Pathantar was the best thing we had produced, and it worked splendidly for fine collation, especially of verse. We did not want to throw it overboard. When the gross collator add-on failed, we thought of preparing a separate gross collator with a different logic, and *manually* loading its output onto Pathantar for fine collation. We had not yet dared to think of a seamless, fully computerized process embracing both gross and fine collation. Nor had we quite realized how unviable a manual transfer from one to the other would be.

The first attempt at an independent gross collator was made by Arunashis. He created a PHP-based program and tried it out on relatively small test sets, breaking up a text file into text blocks ('chapters'), and the latter into paragraphs. Already, the gross collating exercise was splitting into two levels. The structure of Prabhed, the program finally used for Bichitra, was rather similar, but its working principle very different.

Prabhed: At Last

We might have continued with such leisurely experiments had it not been for the impetus (and no less the funds) afforded by the Bichitra project. Sunanda Bose, commonly known as Neel, joined the band while we were sorting out our ideas, and soon became one of the key players. Like many programmers of greater experience, he set out by thinking that it would be easy to put together a collation program by applying the 'diff' utility. Luckily for us, Neel had little experience at the time: he had just completed his training and not yet found a job. (He did soon enough, in Kolkata's flourishing IT sector, but left to become a research fellow at Jadavpur.) The software program that finally emerged, 'Prabhed' (Difference), was the creation of Arunashis, Spandana and Neel, with some textual and structural contribution from Sukanta.

By this time, we had concluded that our demands could only be met by integrating gross and fine collation in an outwardly seamless process, even if the two tasks were addressed by different programs. The gross collation, moreover, had to be at

two levels. The whole exercise, then, would incorporate three levels: as we termed them,

- **Sections:** the chapters of a novel or long prose tract, the scenes of a play, or the cantos of a long poem.
- **Segments:** the paragraphs within a prose chapter, speeches within a scene, stanzas within a canto; or a short poem, story or essay in its entirety.
- **Words.**

(See Fig. 8.4a, b)

There was no collation program in the world that could do as much, and there still is no other. The exceptional range and complexity of Tagore's writings gave us

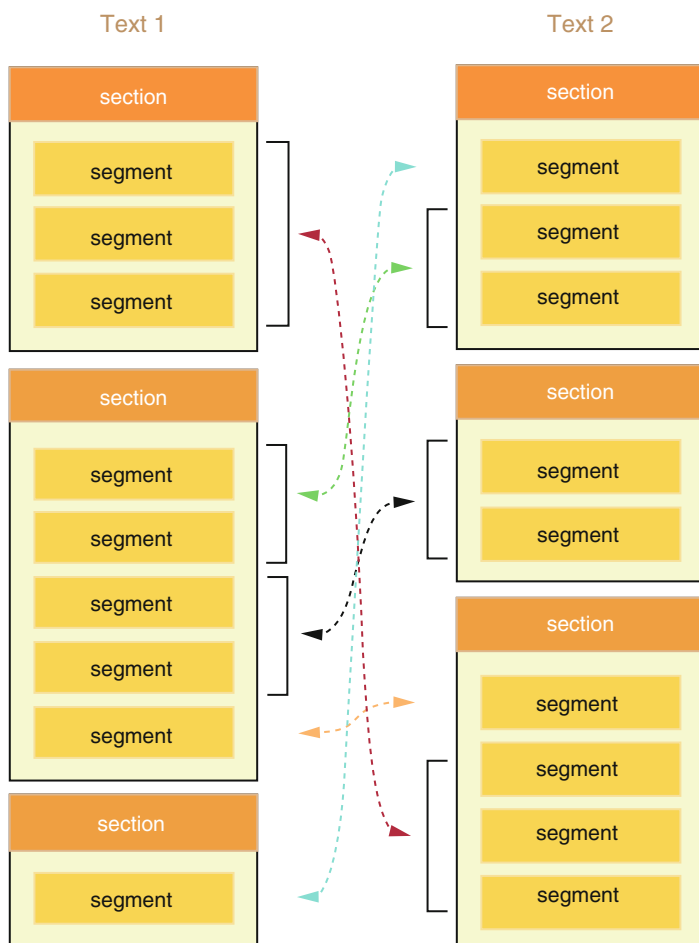


Fig. 8.4 (a) Prabhed, gross collation: Diagram indicating the sections and segments of a work in changed positions in two versions. (b) Prabhed/Tafat, fine collation: four-window display

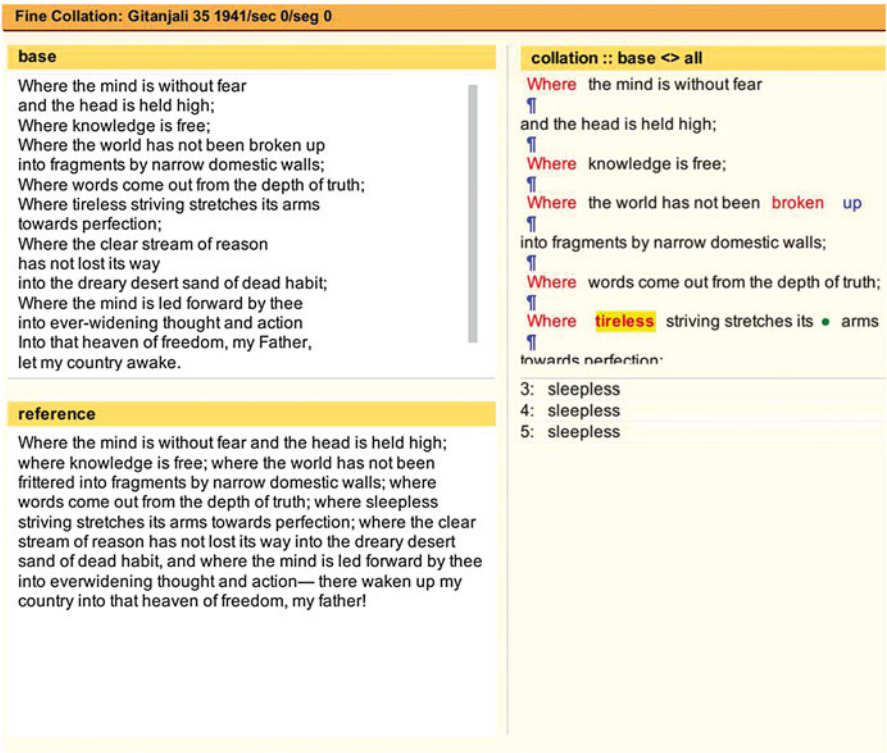


Fig. 8.4 (continued)

the best possible scope for creating one. Rather more stressfully, the 2-year deadline (reduced to one or less by the time we got to this point) wonderfully concentrated our minds.

Like Arunashis but unlike the rest of the Bichitra team, Neel had come to textual computing from the computing rather than the textual end. When it was all over, he recorded his initial feelings as we embarked on the exercise: ‘The biggest, and unavoidable, problems were the author and the language. The first had never conceived of his works being collated by computer, so he had thoughtlessly cluttered the texts with needless spelling changes and textual revisions. And the grammar and syntax of the Bengali or English language doesn’t match that of computer language. But if we turn for help to computational linguistics and natural language processing, we clutter the field impossibly.’

How did we solve these challenges? At the risk of putting the cart before the horse, let’s begin with a guided tour of Prabhed as it appears on screen. After all, we ride in the cart.

Viewing Prabhed

Gross collation: Section Level

By clicking on the ‘Collation’ icon in the Full Bibliography (Fig. 8.5), you may be surprised to see a display of colourful bands spread across the page (see Fig. 8.6). Each colour indicates the full text of one version, and the dark-to-light divisions indicate the sections (chapters, scenes, cantos etc.) within it, in proportion to their



Fig. 8.5 The collation button in the Full Table



Fig. 8.6 Prabhed: opening display at section level for the play *Achalayatan*. Each colour-coded band indicates a version of the text. Each differently shaded block within it indicates a scene (section)

relative size within the work.¹ Select any section in any version as your base text by clicking on it. The matching sections in the other bands will be highlighted and underlined in red. The selection panel at the bottom will stabilize, with the base text as the first entry to the left and the others following. Figures in the selection panel indicate the percentage of correspondence of each version with the base text (see Fig. 8.7). We will call this percentage *match percentage*. There is also another link in the selection panel (usually appearing as a tiny hollow square) which opens a text file of that section in that particular version (see Fig. 8.8).

That's not all. In the same selection panel, click on any version other than the base text, and a vertical panel will open to the right of the screen, showing a segment-by-segment comparison (paragraphs, speeches, stanzas) of that section between the base text and that version (see Fig. 8.9). The segments in the base text are indicated by little coloured rectangles in the left-hand margin, those in the reference text (the version being compared) in the right. Corresponding segments are joined up by grey lines. Segments present only in one version appear only in one or the other margin.

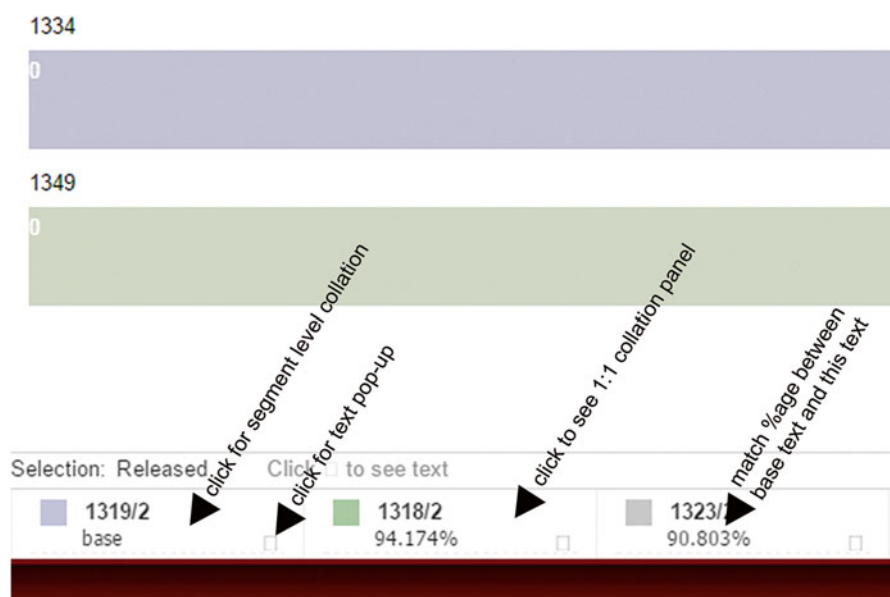


Fig. 8.7 Prabhed: close-up of selection panel at bottom of section-level display for the play *Achalayatan*

¹This is not true of the relative size of the various versions in their totality. All the bands span the full screen, so that a short version occupies as much total space as a longer one.



Fig. 8.10 Prabhed: segment-level view for the play *Achalayatan*, with panels opened for the relevant text blocks. Each band indicates a scene (segment) of the text, each small coloured part of it a speech

Gross Collation: Segment Level

Here, at first, you will only see a single band, representing the section (chapter, scene or canto) you chose from the earlier screen (see Fig. 8.10). This band is divided into many small parts, each standing for a single segment (paragraph, speech or stanza) within that section. With short poems, stories or essays, there may be no effective difference between section and segment level: the whole work may consist of a single ‘stanza’ or ‘chapter’.

Click on any small part. (These are wide in proportion to their size, and the narrowest are scarcely visible. But there is a numerator to the top right that you can click to move from segment to segment.) On clicking, a number of other bands appear below, representing the corresponding sections in the other versions: the matching segments in each are highlighted and underlined in red. Again, the selection panel at the bottom shows the percentage of correspondence, and one can click a tiny square to open the relevant text.

But unlike the former screen at section level, here one cannot open a column showing 1:1 correspondence between matching parts from two versions. The parts are too small for that. Any detailed comparison must be a fine collation at word level. This fine collation is presented in the next screen, opened by clicking the name or coloured square of the base text in the selection panel at the bottom.

Fine Collation

The fine collation display resembles the one used earlier in Pathantar, except that now only the variant words show up in the Results window, not the whole line or sentence where it occurs. The display has four windows. The top left contains the base text, the

links to the other versions being placed in a menu to the left of it. The top right contains the same base text, but this time hyperlinked and four-colour-coded:

- Black where the text is the same in all versions.
- Red where there is some corresponding material in all the versions, but not always the same.
- Blue where the base text has material missing in one or more other versions.
- Green (showing up as a dot) where the base text lacks material found in one or more other versions.

By clicking on a red, blue or green link in the top right window, the lower right window displays the variants. The lower left window can display any one variant version chosen from the menu (see Figs. 8.11 and 8.12).

Planning and Making Prabhed

To recapitulate, the big jump between Pathantar and Prabhed was to incorporate gross collation—that is to say, the comparison of blocks of text (**sections**), which may be broken down into sub-blocks (**segments**). A section could be a chapter of a long prose work, a scene in a play, or a canto in a long poem like an epic. A segment could be a paragraph within the chapter, a speech within the scene, or a stanza within the canto. A short piece like an essay or a lyric poem may not need such three-level collation, though they often contain some minor division (say, the title or an introductory section). Here, the gross collation can be quickly clicked through to arrive at the fine collation.



Fig. 8.11 Prabhed/Tafat: four-window fine collation display. A red word (highlighted) has been clicked in the top right window to show a variant reading in other versions at the bottom right



Fig. 8.12 Prabhed/Tafat: four-window fine collation display. On the left: A blue word (high-lighted) has been clicked in the top right window; versions with a gap at that point are indicated at the bottom right. On the right: A green dot (highlighted) has been clicked in the top right window; other versions have additional text at that point as shown at the bottom right

Parsing

The first need was to enable the program to recognize the text blocks and sub-blocks, or sections and segments, and store them separately for each version. For this, we needed parsers to divide the text file into one or more sections, each section into one or more segments, and each segment into one or more lines. (Bichitra does not store results at line level, but recognition of the line level is crucial all the same.)

What is parsing? In our context, it is the process of analyzing on a computer a string of characters constituting a text, following the rules of formal grammar. It is how we teach the computer to recognize when we have reached the end of a chapter, canto, stanza, paragraph, word, or whatever. We can then ‘slice’ or divide the text according to these grammatical divisions. We had to design our parser with the following factors in mind:

- How could we mark the section and segment divisions in a way easily handled by the human transcribers, but readable by the software?
- Could we use the same parser for prose and verse?

The first problem called for immediate attention: the transcription guidelines had to be in place from Day One. The army of workers transcribing texts had to be told what to do. We could not expect them to be trained in XML tagging or TEI markup. If we got someone else to insert the tagging later on, that would mean more time and expense, and an extra round of checking. But even more crucially, the tags would be

read by the collation program and included in the results—unless a special safeguard was factored into the code, making an elaborate program still more complicated.

The second problem was equally crucial, and had a bearing on the first. The definition of section, segment and line had to be content-sensitive, distinguishing between prose and verse. In verse, the line division is obviously important. For prose, we can take the sentence as the ‘line’: the actual line division on the page is accidental. We needed content-specific parsing rules to distinguish between these two situations.

We resolved both problems by a childishly simple but entirely effective means. We used the Enter key—or more exactly, the ‘newline character’ (↵) operated by the key—as the separator. In prose, ↵ appears at the end of a paragraph, but in verse at the end of each line, though the grammatical sentence may stretch out for several lines. We incorporated these provisions in two different parsers, one for prose and one for verse: the **standard** and the **linefeed parsers** as we named them.

The standard parser recognizes the line as a string of words ending in a sentinel symbol (full stop, exclamation mark, question mark). It was used for prose. But the linefeed parser, used for verse, treats the newline character (↵, operated by the Enter key) as the sentinel symbol, and ignores punctuation marks. Using this distinction, we devised a simple set of rules for marking sections and segments as follows.

In the standard parser for prose:

- The text between two sentinel symbols constitutes a ‘line’ (in this context, a sentence). *The transcriber was to type on as normally from one sentence to the next.*
- A single newline character (so that the following text starts on a new line) is a segment separator—i.e., marking the start of a new paragraph in a prose text or a new speech in a prose play. *The transcriber was to use the Enter key to start the new paragraph on a new line.*
- Two newline characters (creating a blank line between two text blocks) is a section separator—i.e., marking the start of a new chapter of a prose work, or a new act or scene in a prose play. *The transcriber was to leave a blank line between the two sections.*

In the linefeed parser for verse:

- A single newline character indicates the start of a new verse line. *The transcriber was to use the Enter key to start each new line.*
- Two newline characters (creating a blank line between two text blocks) is a segment separator—i.e., marking the start of a new stanza in a poem or a new speech in a verse play. *The transcriber was to leave a blank line between the two segments.*
- Three newline characters (creating two blank lines between two text blocks) is a section separator—i.e., marking the start of a new canto, or a new act or scene in a verse play. *The transcriber was to leave two blank lines between the two sections.*

This process, so simple to master, resulted in a logical model of the text: that is to say, it allowed Prabhed to identify specific chunks of characters (representing blocks of text) as sections and segments. Having prepared the field in this way, we could now set about the actual task of collation.

We have seen how with our earlier software Pathantar, we used some approximations (or heuristics, to use the technical term), judging the match between two lines or sentences by comparing their first and last words. This could yield misleading results all too often. We were gradually convinced that it is unwise to use approximations at all in collation software, given the extraordinary range and unpredictability of linguistic phenomena. We would simply create an extra tier of problems, compounding the instability of the text itself with the fresh instabilities introduced by heuristic methods. In Prabhed, we set out to create software that would compare everything with everything.

1:n and n:n Collation

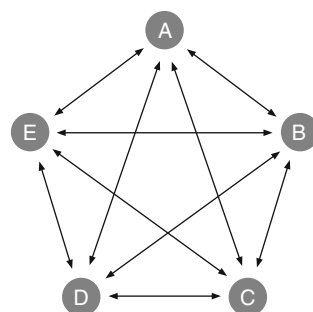
As noted while discussing Pathantar, all working collation programs we know of actually compare two versions at a time. A few elaborate ones combine these 1:1 collations to produce a final output comparing several versions. This was our goal with Prabhed as well.

Prabhed starts by comparing the same base text against a number of different versions, one by one. It then combines the results in a single file. This gives a 1:n collation. It then takes each of the other versions in turn as the base text, and compares it with all the other versions, one by one. In other words, if there are five versions, A, B, C, D and E, it will compare

A with B, C, D and E
 B with A, C, D and E
 C with A, B, D and E
 D with A, B, C and E
 E with A, B, C and D

(See Fig. 8.13.) Note that to collate A with B is not the same as collating B with A. If the position of the two texts is reversed, the collation results will be recorded

Fig. 8.13 Five-point star and pentagon: diagram of full collation of a text in five versions. A text in six versions would have a six-point star and hexagon, and so on



from the opposite end, as it were: the match percentages will be different, sometimes glaringly so. See more on this under ‘Split and merge’ below.

By combining all the results, we will effectively have carried out an $n:n$ collation, though Prabhed displays the results on a $1:n$ basis. Match percentages can only be calculated with reference to a single base text taken as 100: that is tantamount to a $1:n$ collation. By the same logic, when proceeding from one level to the next—section to segment, segment to word—we have to select a single version as the base, and measure the others against that. Again, we are presenting only a $1:n$ collation. And the vertical column that opens up to the right of the Sections page offers only a $1:1$ collation.

At different points for different purposes, Prabhed thus presents $1:1$ and $1:n$ results, adding up to an $n:n$ collation. But the root process is always $1:1$. It seems correct to say that a genuine $n:n$ collation engine, collating all with all versions in a single process, has not yet been operationalized. It calls for a completely different computational logic that no one has yet harnessed for the purpose.

Three/Four-Tier Collation: Up and Down the Ladder

As seen by the viewer, Prabhed operates from the top down. We start with the biggest text blocks, namely sections, and proceed downward through progressively smaller blocks: segment, line (though this is not used in Bichitra) and word. But the actual operation of Prabhed is in the opposite direction, from the bottom up.

Obviously, we can only measure the similarity between sections by aggregating the similarities between the segments constituting them. By the same logic, segment collation must draw on line-level collation, and line-level collation on a word-level exercise. And to compare words, we must necessarily compare them character by character.

Hence the collation process carried out by Prabhed starts with the most basic components of the text, namely characters. It works upwards through the levels of word, line, segment and section to cover the whole document. It then reports those results in the opposite order as making most sense to the human user, who must know the level or nature of the text blocks being compared to assess the differences between them (see Fig. 8.14).

The best thing about the bottom-up rationale is that it ensures a comprehensive collation of literally every particle of text. There is nothing heuristic: no guesswork, no approximations, no resort to shortcuts or samples.

How Like Is Like?

How will the collator determine whether two words are the same? The problem is more acute in a synthetic language like Bengali, where a much higher proportion of words are inflected than, say, in English. So a slight difference in the ending might mask the presence of the same word in two forms. But an equally slight difference

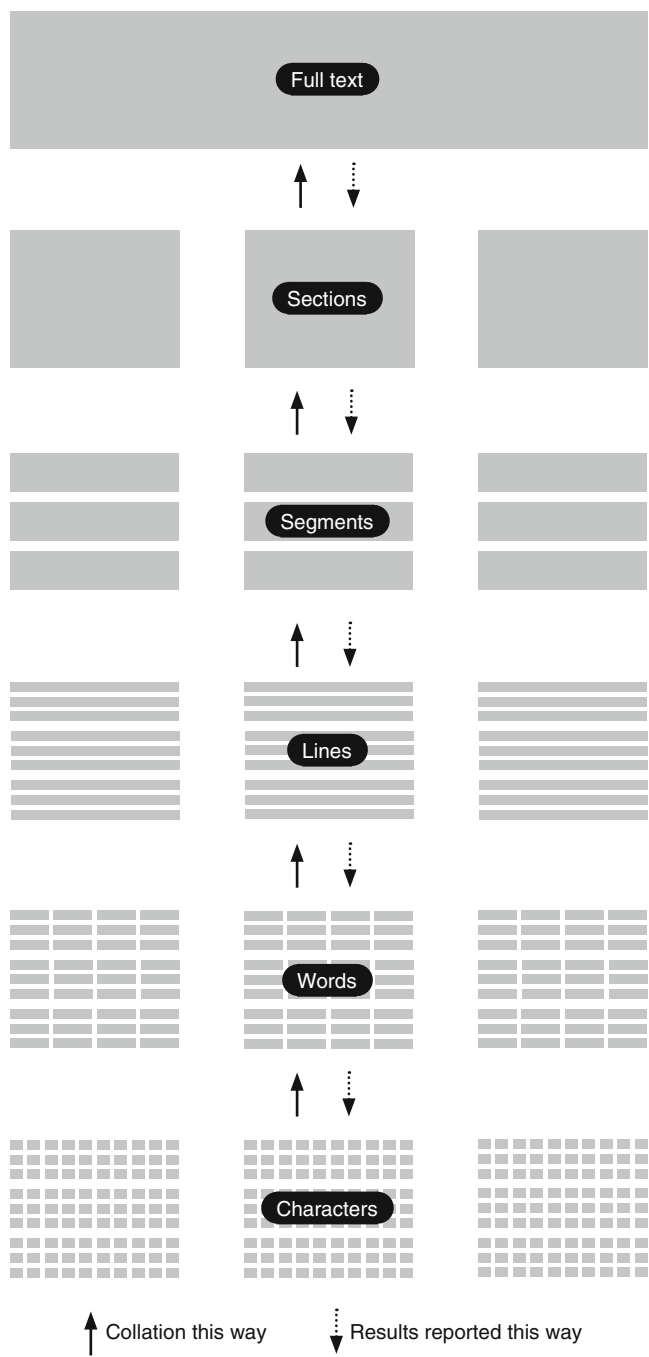


Fig. 8.14 The 'collation ladder': character to section and down again

might also distinguish two entirely separate words. To ensure as balanced a count as possible, we adopted the following scale of correspondence:

- In a word of four or fewer characters, all of them must match.
- In a word of more than four characters, the program would allow one difference for every four subsequent characters. Thus a word of 5–8 characters could differ by one character; a word of 9–12 characters by two; and so on.

It should be remembered that the Bengali script makes extensive use of vowel tags and conjunct letters. Hence what appears visually as a single glyph may represent a combination of up to four consonant and vowel characters.

Match Percentages and the Similarity Matrix

It is virtually never the case that two text blocks are absolutely the same. In that case, how much variation do we allow in declaring a match between two blocks? 1 %? 10 %? 25 %? Any such distinction is bound to be arbitrary. So rather than ask ‘Do these two blocks match?’ expecting a simple Yes/No answer, it is better to measure *the degree to which they are similar*. These are the match percentage points seen in the selection panel at the bottom of the page. 100 % would mean an exact match, 0 % the absence of any similarity whatsoever.

Our basic objective was to collate comparable text blocks from different versions of a work. For this, we needed to create a set of tables, each with one of the compared versions as base text, its text blocks laid out in rows. The blocks from the other version being compared with it (the reference text) would be laid out in columns. The cell where a row met a column would contain a percentage value to indicate the degree of similarity between the two blocks. The resultant table or spreadsheet is called a **similarity matrix**.

The collation pages in Bichitra include a ‘grid view’ (see link to the top right of the segment collation page) with these basic results in ‘raw’ form (see Fig. 8.15). The same results are presented in an attractive, user-friendly way in the coloured displays, in band or riband format for gross collation and four-window format for fine collation. Once the results are obtained, designing such an interface calls for imagination, but it is a secondary task.

The crucial question is, how to arrive at the results? We have seen how the collation proceeds through a nested structure of textual components: document, section, segment, line, word. To repeat, the match percentage of a pair of sections (say, chapters in a novel) obviously depends on the match percentage of the segments constituting those sections (i.e., the paragraphs in those chapters). So first of all, we need a segment-level matrix, which will compare each segment in a section of the base text with every single segment in the corresponding section of the compared or reference text, and lay out the results in a grid. Ultimately, Prabhed compares every segment in each version with every segment in every other version, not only the apparently matching segments as heuristically determined. A similar exercise is

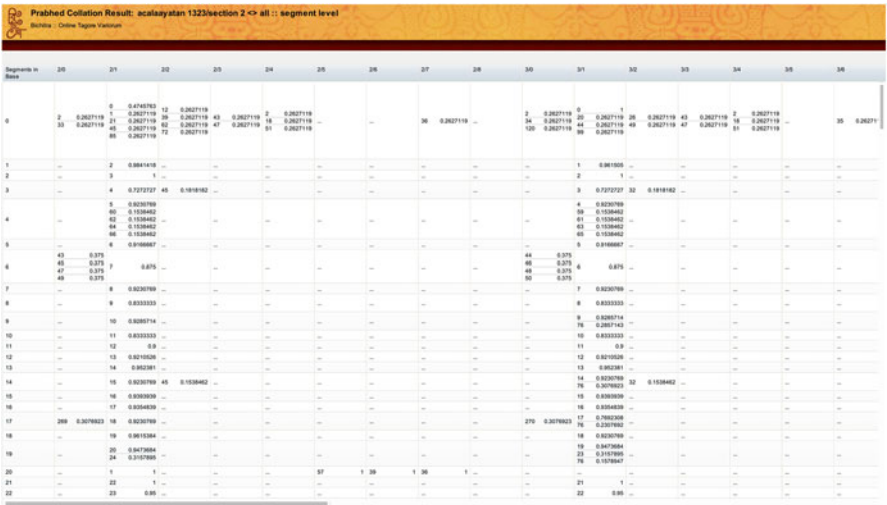


Fig.8.15 Prabhed: grid view at segment level

carried out at section level, based on the segment-level results. As Neel remarked, Prabhed is the world’s most patient and painstaking clerk.

Packing and Unpacking: Margin

Once the ‘clerk’ has completed his accounting, we have an unbiased repository of match percentage values, calculated against all possible sets of homogenous text blocks. These full and unfiltered collation results covering all the versions are stored in a single .gcl file (GCL = Gross CoLlation). We termed this process **packing**. However, to keep the .gcl file within manageable size and processing time, the file does not store collation data at line and word levels, though of course that data is implicit in the results at segment level.

With the .gcl file to hand, we can apply appropriate filtering rules to extract the matches we require: the texts need not be collated again and again. The user may try different filtering values till he gets satisfactory results. We termed this process of filtering and result extraction **unpacking**. The chief purpose of unpacking is to extract from the huge mass of data, comparing every section and segment to every other, only those match percentages that seem significant—i.e., those above a certain threshold, say 60 % or 70 %. For Bengali—at least Tagore’s Bengali as incorporated in Bichitra—we found the optimal threshold value to be 60 %, and set this as the default. All results below the threshold are eliminated to keep the final quantum of files within manageable size (see Fig. 8.16).

These significant results are extracted by applying the filter of a threshold value that we called the **margin**. Where we place this margin or cut-off point depends on the nature of the text. (Prabhed has provision for such adjustment.) Where two

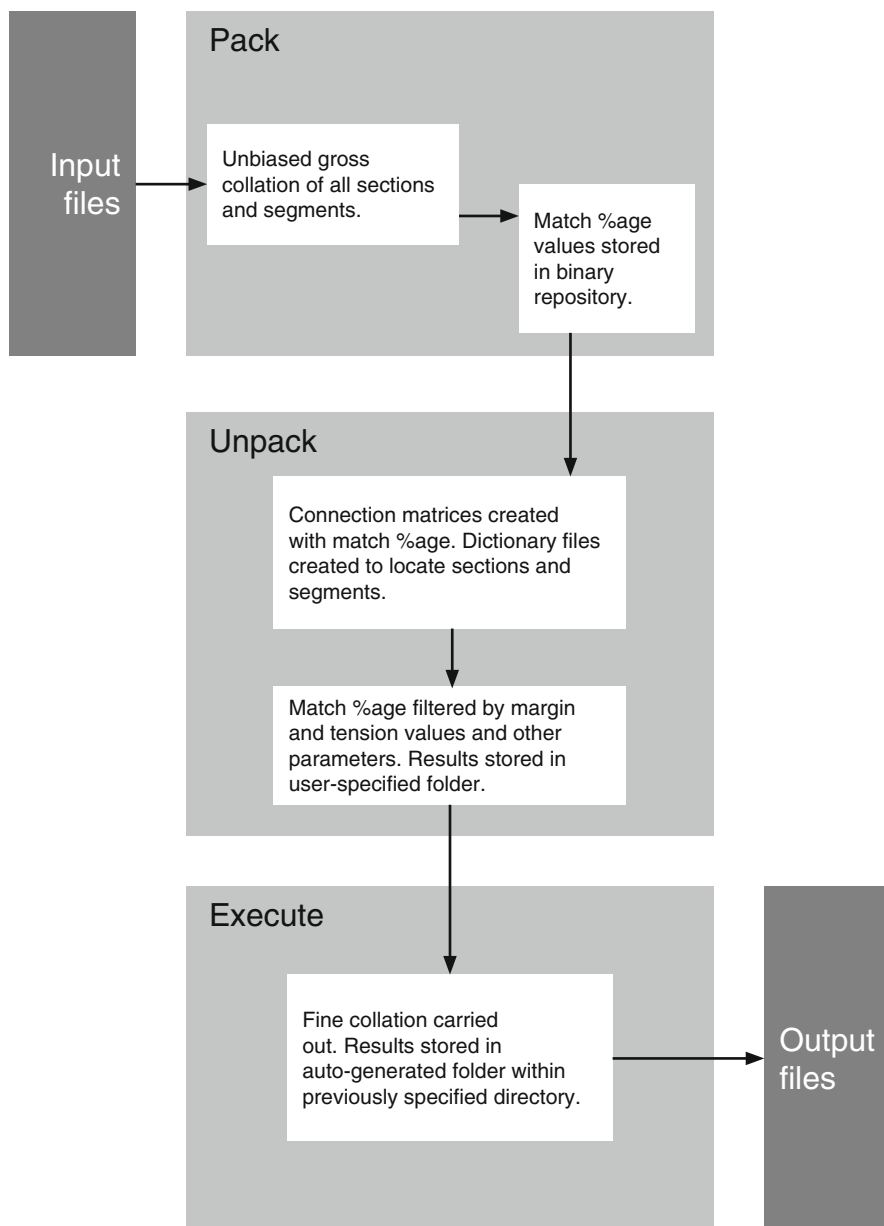


Fig. 8.16 Packing and unpacking: workflow diagram

versions are hugely different, we may need to keep the margin low if we are to salvage any similarities at all. To some extent, the margin also needs adjusting to the nature of the language. An analytic language like English needs a high margin, as any English text is full of ‘little words’ like *of, on, in, and, or, a, an, the*, and so on—chiefly articles, prepositions and conjunctions. These alone can account for a fair degree of similarity between totally dissimilar passages. A relatively synthetic language like Bengali, where the function of articles and prepositions is largely served by suffixes rather than separate words, seems to require a lower threshold: but on the other hand, we have to allow for the fact that the same word may not be recognized as such owing to a suffix at the end or, still more trickily, an inflexion that modifies the stem. To allow for such cases, we need to load a lexicon, which does not as yet exist for Bengali. (See the conclusion of Chap. 3.) The need is still greater for highly inflected languages like Sanskrit or Latin.

The percentage of significantly similar segments, acquired by the above means, determines the match percentage of two compared sections. These match percentages appear in the selection panel at the bottom of the page. So also, the percentage of significantly similar sections determines the match percentage of two compared documents, though Prabhed does not actually show the percentage at this level.

During unpacking, the program ‘slices’ the text into sections and segments to store as separate plain text files, which can be opened by clicking on the little box in the selection panel. Also during unpacking, Prabhed organizes its results according to its own directory structure, keeping the visualization in mind, so that the necessary collation result files as well as the ‘sliced’ text files can be called up as required. The match percentage data is stored in spreadsheets, and also in JSON files. The latter, moreover, contain the paths of all relevant files, which are drawn on as required for the four-window fine collation display.

At this point, we can view the first two levels of collation results on a browser. Now we can

- Check the degree of similarity between sections and segments, and extract the corresponding percentage values.
- View plain text files of the sections and segments.
- Trace the appearance, disappearance or transposition of a text block (section or segment).

All these results are organized within an appropriate directory structure, so that the program and indeed the website can easily fetch the required files from the huge corpus.

The Weighted Mean: Split and Merge: Tension

A section contains not only many segments, but segments of different sizes. When two or more corresponding sections are collated, the corresponding segments within them may yield various match percentages. Let us take a simple case of three

sections, A, B and C, with 20 segments in each: a^1 to a^{20} , b^1 to b^{20} , c^1 to c^{20} . We may find that a^1 , b^1 and c^1 match quite closely ($a^1:b^1$ match percentage 90 %, $a^1:c^1$ 95 %). But the $a^2:b^2$ match percentage is only 53 %, and $a^2:c^2$ is 58 %. The remaining segments, too, match in widely varying degrees. (As a general rule, the bigger the text blocks being compared, the lower their match percentage.)

This is of little consequence if the segments are of roughly the same length. But suppose the section in question is a chapter of a novel, and the segments are paragraphs within it. The first paragraph (a^1 , b^1 , c^1) is a single short sentence of ten words, the second (a^2 , b^2 , c^2) a long stretch of 87 words, and the other 18 of varying length. If we give equal value to every paragraph, a high match percentage in paragraph 1 will play a disproportionately high part in determining the similarity between the chapters as a whole: they will seem more alike than they really are. The opposite may also be the case: a difference of three words in a one-sentence paragraph of six words will produce a low match percentage of 50 %, pulling down the overall match percentage between the compared chapters—i.e., making them appear less alike than they are.

To prevent these skewed results, Prabhed uses a **weighted mean**. It considers not only the match percentage between compared text units, but also the relative size of those units, or their proportionate space within the larger unit of which they are a part. In collating text blocks and selecting comparable ones, Prabhed takes the weighted mean into account in calculating the match percentages.

Another set of problems arises where a single text block in one version is split into two or more parts (not necessarily contiguous or of equal size) in another. These are cases of **split and merge**: i.e., the text is split when viewed from one direction but merged when viewed from the other. In such a case, one side of the equation may be called a subset of the other. In such a case, if we take the longer block as the base text and compare it in turn with each short, partial block in the reference text, we will get a series of low match percentages, as most of the longer block will be missing from each of the shorter ones. But if we reverse the direction and look for the shorter blocks in the longer, we will find high match percentages, as each shorter block will be substantially present in the longer. So in this case, an adjustment needs to be made between the small-to-big and big-to-small ratios (see Fig. 8.17).

To meet this need, we introduced a new parameter. If two compared passages crossed the ‘margin’ or threshold of match percentage in one direction but not the other, we applied a certain value (lower than the ‘margin’) that we called **tension** in the opposite direction. If the latter collation crossed this lower threshold of ‘tension’, we accepted the compared text blocks as similar. In such a case, the resemblance is very likely owing to splitting and merging.

Translocation

As explained above, one of the chief problems with our earlier software Pathantar was dealing with transposition or translocation of relatively large chunks of text: for

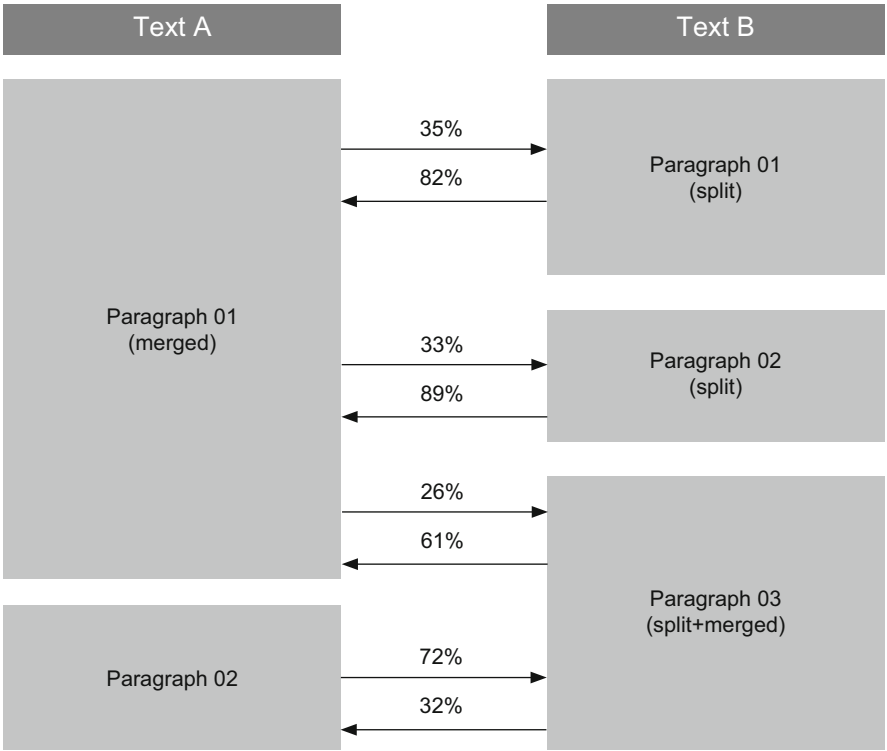
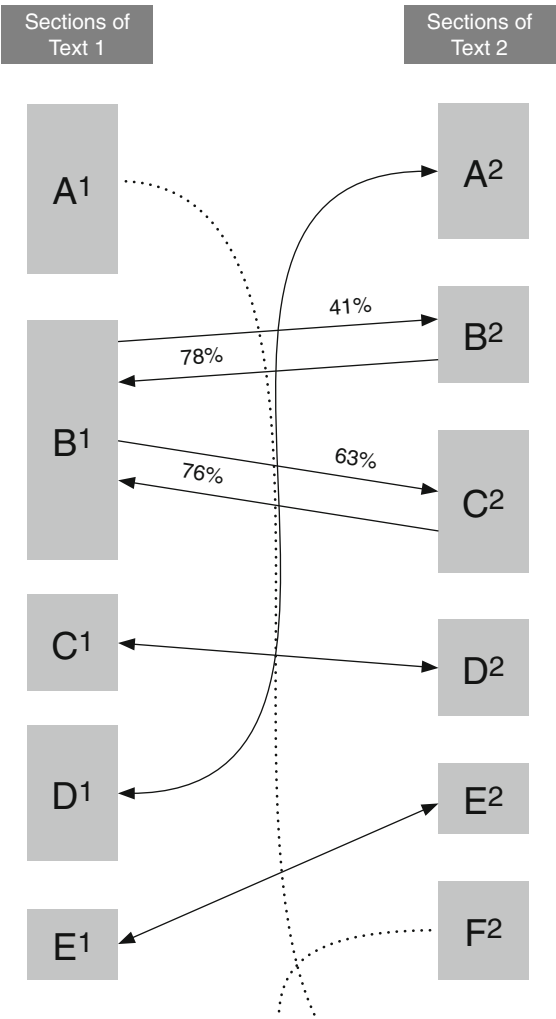


Fig. 8.17 Split and merged text blocks: diagram showing various possibilities

instance, if the third chapter in one version became the fifth in another, or a single chapter in one version was split in two or more, perhaps with the parts distributed among other chapters along with other text. The farther the distance between the respective locations, the greater the problem of collation. In fact, we do not know of any other collation program that can handle this problem successfully.

The great advantage of Prabhed in this regard is that it does not take heed of transposition boundaries or indeed any boundaries. It compares every section, segment and line with every other in the compared version(s), and identifies similar passages by match percentage irrespective of their locations. It can therefore spot the ‘migration’ of a section (say, chapter) from one position to another, even from the beginning to the end of the work. What if a section is split in two or more parts, perhaps placed at wide intervals? In such a case, all matching parts will be shown if they meet the ‘margin and tension’ test: that is to say, the sections containing small parts of the text will also be shown as matching to that extent (see Fig. 8.18). And when collating at segment (say, paragraph) level, even segments that have been transferred to some other section (not shown as a match at section level) will be shown as matches if they meet the requirements.

Fig. 8.18 Translocation of text blocks. B1 is a ‘split-and-merge’ case vis-à-vis B2 and C2



In a word, Prabhed can spot translocated text at whatever distance, as well as the parts of a split text block, however far and wide they may be scattered. It can also spot multiple occurrences in Version B to match a single occurrence in Version A. We believe this to be a function unique to Prabhed.

Multiple Repetition

However, there is one situation where this unique feature of Prabhed can backfire to the user’s disadvantage: that is where a text block is repeated, perhaps several times, in both versions. A typical example would be a standard stage direction like ‘Exit’ in a play text. In such a case, if left unadjusted, Prabhed will show multiple matches

in Version B for every occurrence of the text block in Version A. The match percentage will be very high, perhaps even 100 %. Such matches are responsible for many of the ‘railway junction’ effects in the 1:1 sectional comparison table opening to the right of the section-level collation page (see Fig. 8.19). The proliferation of such matches can create a kind of spam in the collation results, obscuring more important similarities.

These are not false matches: the text is genuinely being repeated. To deal with this problem, we introduced a new parameter. If a text block in Version A shows a high match percentage with several blocks in Version B, every match after the first in each section of Version B containing such matches is treated as a false match and excluded from the results set. The default match percentage for this ‘Discard repeated’ parameter is 85 %, but it can be manually set to any value.

This is not a happy solution. If the text block is repeated in Version A as well as Version B— a^1 , a^2 , a^3 etc. against b^1 , b^2 , b^3 etc.—it will exclude every genuine match after the first: i.e., $a^2 < > b^2$, $a^3 < > b^3$ etc. It represents the kind of approximation we have generally tried to avoid in Prabhed. Time did not permit a better solution during the making of Bichitra, but we hope to find one in the days to come.

Gross to Fine Collation: Tafat 2.0

Once the similar text blocks have been identified, and appropriate filtering of the results carried out, we will have created a list of ‘comparable sets’. These comparable sets are sent to the character-level collator or fine collator. For this purpose, we used a second version of Tafat (**Tafat 2.0**) that Siddhartha had created by this time.

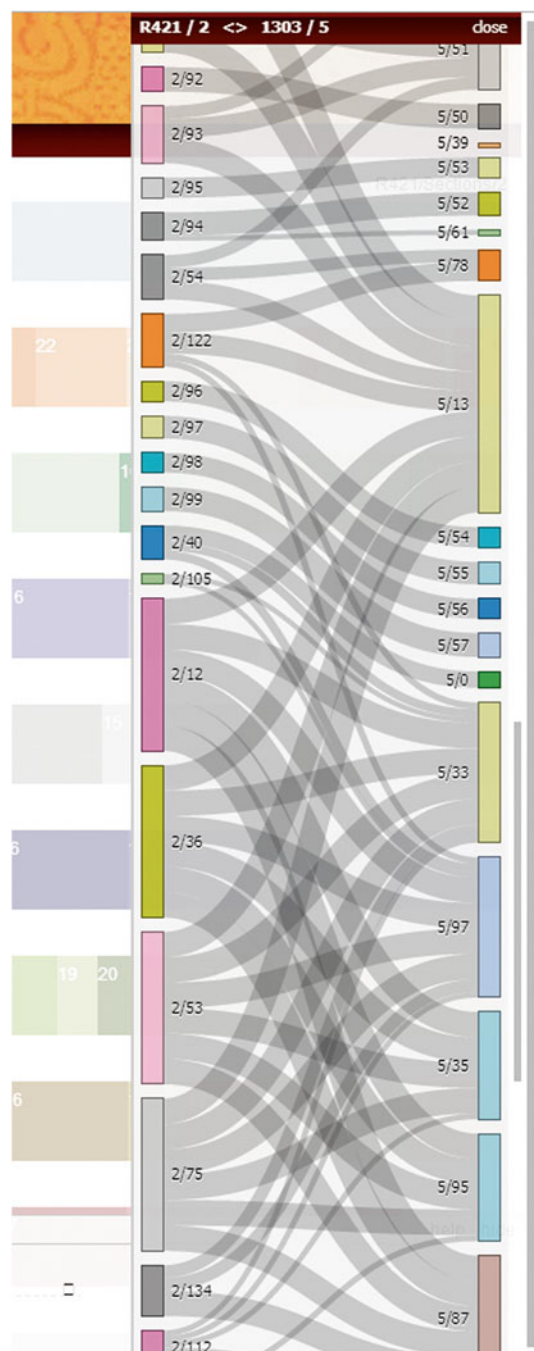
Tafat 2.0 made two quantum leaps beyond our earlier fine collation programs:

- As compared to Pathantar, its collation took account of characters rather than text blocks.
- As compared to Tafat 1.0, it could present results in 1: n form, which could then be compiled into an $n:n$ format.

Tafat 2.0 takes a set of comparable text blocks as plain text Unicode files and generates an HTML response. By incorporating this in a user-friendly interface, especially through the four-window display we had found so serviceable, we arrived at a solution to meet our needs. This involved some adaptation of Siddhartha’s original Tafat 2.0 code, as the original form of the HTML output was not geared to the four-window display. The basic collation program, however, was left untouched.

To fine-tune the gross collator and integrate it with the fine collator, Sukanta, Spandana and Neel went through a gruelling routine for months. To begin with, the gross collation results of Prabhed were incorporated in two output spreadsheets, one before and one after applying the threshold parameters. A new version of this gross collator would be created virtually every day, or rather night, after Neel returned from the lab where he then worked. We conferred on the phone till late into the night, scrutinizing the two spreadsheets against the actual pattern of variants. In

Fig. 8.19 Prabhed: 1:1 collation panel with 'railway junction' effect, showing multiple matches between segments in two versions



other words, we were carrying out the bizarre task of manually checking a computer's output. When we were satisfied with the two spreadsheets—i.e., when they were correctly recording the variants—we froze that threshold and sent the resultant set of segments for fine collation.

There was one more problem to solve in Tafat 2.0. We have already described the occurrence of 'split and merge' in the context of gross collation, which Prabhed can deal with efficiently. But unlike Prabhed, Tafat proceeds by collating comparable text blocks—identified as comparable by a character-by-character collation, starting at the beginning of the base text block. Hence if a large text block (say, a paragraph) in one version is split into two or three in another, it has problems matching the second and following parts of the latter to the whole of the former: instead, it shows those parts as deleted or inserted (as the case may be) in the latter. We largely solved this problem by the handy means of 'sewing' together the split parts to form a total block comparable to the single block in the other version. The parts to be sewn are identified by the operation of 'tension', defined above, as the split parts of the corresponding merged block in the other version. But we must admit that we cannot always control the results. This is one of the residual problems with the Bichitra collation engine, which we need to solve in the future.

32-Bit Versus 64-Bit: The 3 GB Barrier

This section is a warning against complacency. Once we felt confident about Prabhed's output and our control over it, we started testing with large text sets: four or five different versions of long plays or novels, each version of at least 20,000 words, divided into many sections and segments. We found Prabhed was taking a great deal of time to process the files and prepare the .gcl file. Prabhed was only doing its duty: faithful clerk that it is, it was comparing and calculating every possible combination of two words between every two versions of the text set.

The first solution we devised was to make the program multi-threaded, so that if the host computer had a multiple-core processor, the workload would be distributed between the cores rather than concentrated in one single core. This made the program perceptibly faster, how much so depending on the number of cores. We chose two eight-core units and a few quad-core ones for collation. If a computer had a total of n cores, Prabhed had instructions to use $(n-1)$ cores, leaving one core for the native programs running at the same time. We were ready to resume celebrations when the biggest problem surfaced. The program started crashing while processing larger documents. It would start normally, but stop somewhere in the middle of the job without prior notice.

After much study and repeated experiments, we located the problem. Prabhed copies all texts to the RAM (Random Access Memory) or temporary memory of the computer. It then identifies and marks the sections and segments, 'slices' or divides them, compares the text blocks and stores the results. The longer the text files, the more involved and numerous are the resultant interconnections between different tiers of text blocks; hence more and more RAM is needed to handle it.

This is where our program was facing what is commonly known as the ‘3 GB barrier’ in computing. This is a limitation of some 32-bit operating systems running on x86 microprocessors. It prevents the operating systems from using more than about 3 GB of temporary memory, even if the machine has, say, 8 GB or 16 GB of RAM. In fact, even if the OS (Operating System) configuration is 64-bit, the program needs to be compiled in 64-bit too. Thenceforth our task became twofold. We prepared both 32-bit and 64-bit versions of the gross collator, and installed each further development of the program on both versions. Once the 64-bit version was up and running, Prabhed successfully collated five files together, each of approximately 100,000 words, without any glitch. It still took a lot of time, but that was a problem we could live with.

Dynamic Collation Versus Uploaded Result

How and where would we put the collation engine to work? Even before we had created Prabhed, we were debating this general issue of basic importance. Should we keep the whole process dynamic or not? That is to say, should the collation be performed on-site in real time, on the go, according to the user’s demand? That would call for uploading only the text files, the collation software, and a framework to visualize the result. Users would select online the versions they wished to compare, issue the collation command and obtain the result. Or should we upload the text files and a comprehensive set of fixed, ready results, which the user accessed as required? The first alternative would have vastly reduced our labour, but it had some inherent problems.

To begin with, the response time would vary widely—unacceptably so with long texts. A collation of the novel *Gora*, for instance, would take more time than any user would agree to wait. Further, many collation requests might be made at the same time, many or all of them involving complicated text sets. In that case, each request (especially if relating to long texts) might take inordinate time to process.

Also, we aimed to make our resources available even to users in remote places with basic computers and slow Internet connections. We could not allow such a crucial function as collation to be so reliant on the server load capacity, network speed, or type and number of requests at a given time. We therefore chose the more toilsome but safer option of collating the files offline, and posting the results on the website.

Postscript: Initial Character Match

Because Prabhed compares each character in each version with each character in every other version, the collation takes time. Simply ‘packing’ and creating .gcl files of the four versions of the novel *Gora*, Tagore’s longest work, took nearly five hours even with multi-thread processing using the 64-bit version of the program. We have

subsequently thought of a way this time can be substantially reduced. Sadly, enlightenment dawned too late to help with Bichitra.

It is really a language-specific modification rather than a programmatic one, but it will apply to many important world languages. By a happy accident, words in Bengali, English and many other languages hardly ever match if their first characters are different: the range of spelling variants generally leaves the first character untouched. In the consonant+vowel conjunct glyphs common in Bengali, the vowel might change in spelling the same word, but the consonant almost never. The only significant instances are some words beginning with the *s* sound, which has three letters to represent it in Bengali; but it would be an extraordinary coincidence to find more than one, or at most two, such variants even in a text block of some size.

We have prepared a new version, Prabhed 1.1, which modifies the collation logic so as to eliminate any word whose first consonant does not agree, irrespective of the general ‘one deviation in four’ parameter described above. By this simple modification, the time taken in gross collation has been reduced by 30 % or more.

This is a very simple improvement. We are thinking more deeply about Prabhed, more radical ways to enhance its working and extend its functions: for instance, to integrate the gross and fine collations in a single seamless program, eliminating Tafat 2.0. We are even mulling bigger breakthroughs to a general program that can compare variations in, and variant structures of, all kinds of complex entities from verbal texts to DNA molecules and the structure of human communities. We are hardly in a position to talk about these new dreams as yet. But we hope we can keep up our knack of turning dreams to realities.

Sukanta Chaudhuri, Ritwick Pal, and Purbasha Auddy

The Challenge

Bichitra aims to provide single-window access to the full range of primary material on Tagore, as well as detailed analytic information through facilities like the hyper-bibliography, hyperconcordance and collation engine. In other words, it serves two crucial functions:

- **Ready access to primary material in digital form.** By a click of the mouse, anyone anywhere in the world can virtually (pun intended) obtain all the material for which they would earlier have to visit Santiniketan or Harvard and go through the formalities of access.
- **Value addition.** They can also obtain manuscript transcripts, clear reading texts of all manuscript and print versions, and full bibliographical information; carry out detailed searches; access the source text behind the search terms; and obtain the results of a more intensive three-tier collation than possible with any earlier collation program.

47,520 manuscript images, 91,637 images of printed texts, along with a huge dataset of collation results, a search engine and a clutch of back-end spreadsheets: this

S. Chaudhuri (✉)

Department of English, Jadavpur University, Kolkata, India

e-mail: schaudhuri@english.jdvu.ac.in

R. Pal

Pixel Poetics, Chandannagar, India

e-mail: ritwick@pixelpoetics.com

P. Auddy

School of Cultural Texts and Records, Jadavpur University, Kolkata, India

e-mail: pauddy@gmail.com

huge and diverse store of material had to be arranged and interpreted within the frame of the browser window. The task of preparing the store rested with a large team as described in earlier chapters. We had to ensure that the fruits of their labour reached the users in the most efficient and user-friendly form. This goal is what sustained us in designing and executing the world's largest integrated literary website.

To compound the challenge, the site apparatus had to be in three languages: Bengali, English and Hindi. Bengali, of course, is the language of the bulk of Tagore's works. English was essential so that readers could access the English works without recourse to any other language. Hindi, India's primary official language, seemed a natural addition for a project sponsored by the Government of India and dedicated to the nation. Besides, there might be many users venturing on the Bengali texts with a basic knowledge of the language, who would appreciate clear instructions in a language they knew better, like English or Hindi.

We selected the website designer through a process of public tender. Bids came from one or two biggish players in Kolkata's flourishing IT industry; also from small setups, run from home or a rented room by young technologists. Not entirely to our surprise, we found the latter more attuned to our needs. The large firms dealt chiefly in financial and technological rather than textual projects. From their bid documents, it was evident that they would not look beyond marginal adaptation of their usual practices. (In one instance, a two-kilo load of bid documents contained just one page addressing our specific needs.) The small operators, on the contrary, had given serious thought to our needs and taken the trouble to design a customized structure and appropriate web interface. Some of them also seemed to have some appreciation of textual and literary issues. After hesitating between two such firms—neither, incidentally, from the big IT hubs—we finally chose Pixel Poetics, a small firm run by Ritwick Pal, a young entrepreneur from Chandannagar, a former French colony in the outer suburbs of Kolkata. Their design consultant, Pinaki De, also created the Bichitra logo.

A full analysis of the website structure would involve a level of technicality beyond the scope of this book. The following account gives the salient points.

The Strategy

The Interface

We may start with the interface (see Fig. 9.1). Though this was the last item to be designed, it would be the first to meet the user's eye. It had to be both visually appealing and user-friendly.

The **colour scheme** of low-tone ochre and brown is seen in many old buildings in the original 'ashram' or inner enclave of Santiniketan, the site of Tagore's university. In Indian tradition, moreover, ochre is the colour of renunciation. The colours were intended to make an academic and spiritual impact. The home page offers glimpses of Tagore's unique manuscript doodles as a signature style associated with the poet. The doodles also feature in the slideshow on the home page, along with a rare photograph of Tagore (from a private collection) and his likeness in two

বিচিত্রা: বৈদ্যতিন রবীন্দ্র-রচনাসম্ভার
 Bichitra: Online Tagore Variorum :: School of Cultural Texts and Records
 বিচিত্রা: ইলেকট্রনিক রবীন্দ্ররচনাসম্ভার :: স্কুল অফ কালচারল টেক্সটস এন্ড রেকর্ডস

Home ► About Us ► User Guide ► Browse Collection ► Bibliography ► Search Works ► Collation ► Credits ► Contact Us

A PROJECT OF THE SCHOOL OF CULTURAL TEXTS AND RECORDS, JADAVPUR UNIVERSITY, KOLKATA
 IN COLLABORATION WITH RABINDRA-BHAVANA, VISVA-BHARATI
 SPONSORED BY THE MINISTRY OF CULTURE, GOVERNMENT OF INDIA

MANUSCRIPTS
 47300 scanned pages of Tagore manuscripts in Bengali and English with transcription
 Manuscripts Guide

Manuscript-wise index

PERIODICALS & BOOKS
 91637 scanned pages of printed Tagore material in Bengali and English

New Additions

- Timeline: Books
- Timeline: Journals
- Alternative Titles
- Table of contents of short story volumes
- Manuscript-only items added to bibliography

**"The light of thy music illumines the world.
 The life breath of thy music runs from sky to sky."**

The Bichitra Tagore Online Variorum is the work of the School of Cultural Texts and Records, Jadavpur University. The Ministry of Culture, Government of India sponsored the project on the occasion of Tagore's 150th birth anniversary. It is a fully integrated knowledge site comprising all Tagore's literary works in Bengali and English, but excluding most letters, speeches, textbooks and translations (except Tagore's translations from his own Bengali).

Read More

PARTNERS:
 Bangiya Sahitya Parishad ■ C. DAC ■ Calcutta University, Central Library ■ Centre for Studies in Social Sciences, Calcutta ■ Houghton Library, Harvard University ■ Indian National Library, Kolkata ■ Jadavpur University: Central Library, Centre for Distributed Computing & School of Education Technology ■ National Informatics Centre ■ Raja Rammohan Roy Library Foundation ■ Senate House Library, University of London

Website updated on: 17/06/2014
 189624 visits
 All rights reserved
 School of Cultural Texts and Records, Jadavpur University | 2014

Fig. 9.1 Bichitra home page

sculptures, one in the garden of Shakespeare's birthplace at Stratford-upon-Avon, England, and the other at Borobudur, Indonesia. The two sculptures symbolize Tagore's fame and reception in East and West. The home page also contains two lines of verse from Tagore's own translation of one of his most famous songs:

The light of thy music illumines the world.
 The life breath of thy music runs from sky to sky.

They seem appropriate to a website disseminating his writings to the whole world through cyberspace.

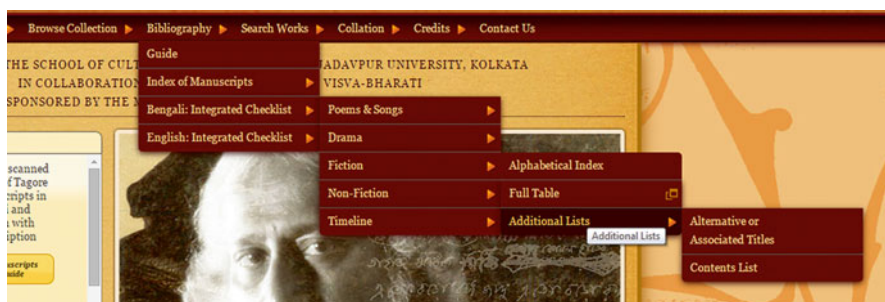


Fig. 9.2 Menus and sub-menus

The site contains the following basic operational features:

- The links are placed in the usual navigation-bar style of any website, with optimized categorization for each section in a drop-down menu with sub-menus (see Fig. 9.2). The home page also contains some important links in the illustration panels on the left.
- In each page, there is provision for choosing the interface language (Bengali, English or Hindi).
- All Bengali material uses the UTF font ‘Siyam Rupali’ created by Omicron Lab and placed under Creative Commons. Users need not have the font installed on their own computers, though to do so (or in fact download the entire Avro keyboard package) might prevent glitches in the correct display of Bengali text, or the correct entry of Bengali words in the search window. A jQuery© plugin for AVRO phonetic is used to allow phonetic Bengali typing. The Hindi interface uses the Gurumaa Hindi UTF font (under general public license).
- There are different ways of searching for specific entries in a section:
 - Subcategory filters. This function is incorporated in the Alphabetical Index of the Bibliography (see Fig. 9.3). Initially divided by genre (Poems & Songs, Drama, Fiction and Non-Fiction), the Alphabetical Index then filters by ‘Title’ and ‘Book’, as also ‘First Line’ for Poems & Songs. Likewise, the Title-wise Index of Manuscripts can be filtered by genre. This feature helps to narrow down the search field and thus reduce search time.
 - An alphabetical key search function, incorporated in the Alphabetical Index as well as the ‘Search Options’ panel opening at the top of the ‘Full Table’ pages on clicking CTRL+F or CMD+F (see Fig. 9.4). This feature allows access to the image or text of any version of a work with a maximum of three clicks.
 - A search engine incorporating both title and text search. A Bengali search term can be entered in Bengali characters or, alternatively, in Roman characters that convert to Bengali on applying the space bar. For searching the full text opening from the results page, one needs to apply the ‘Find’ command (CTRL+F or CMD+F).
 - The ‘Find’ command can also be applied to other pages like the Full Table in the Bibliography menu.

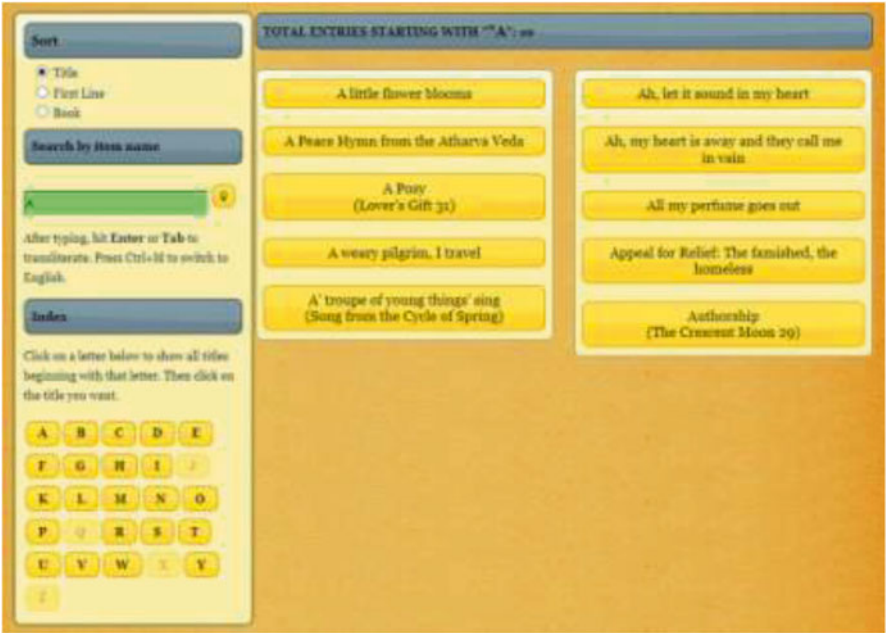


Fig. 9.3 Alphabetical Index: section of opening page for ‘English: Poems and Songs’ showing radio buttons, search term entry box and alphabet search key panel

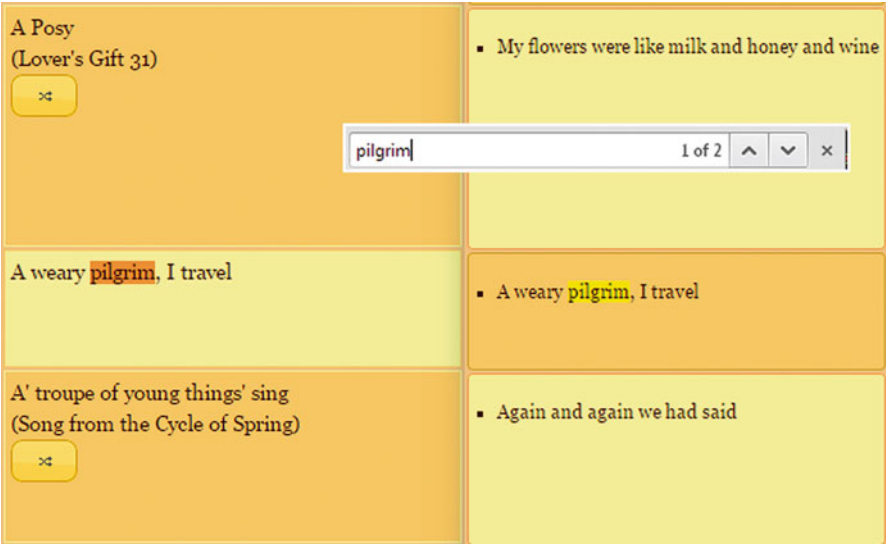


Fig. 9.4 Full Table: search options panel (actually at top right of page, dragged down in this sectional view)

- There is a sortable and searchable tabular structure for large datasets like the Index of Manuscripts and the Full Table in the Bibliography menu (see Fig. 9.5). The data is presented like a spreadsheet, with hyperlinked icons in the appropriate cells to open new windows or pop-ups with more data.
- The image display section is designed to move from one page image to another without reloading or changing the rest of the display layout: only the image of the actual page will be changed. Manuscript images and transcripts can be viewed side by side for easy comparison, though unfortunately there is no provision for simultaneous scrolling. The transcription column can also be hidden by clicking a radio button at the top right, so that the image shows across the full width of the page. Needless to say, there are simple zoom and rotate features for easy viewing (see Fig. 9.6 with command bar showing these features).
- There is a detailed User Manual as well as a shorter Quick Guide to the various functionalities.

Data Feed

One of the easiest ways to maintain a structured dataset in Windows is by using MS Excel©. So we organized the raw data in 32 Excel© spreadsheets, as explained in detail in Chap. 6. An Excel© parser was used to parse the spreadsheets and upload the appropriate data (image files, text files or collation files) to the website database by fetching it from a corresponding set of files backing up the website. The two sets of files were linked by the headings ('Journal', 'Text File' etc.) in the top row of the spreadsheet. These headings apply to all the data in the rows below, constituting a set of pre-formatted indices (see Fig. 9.7).

Table 9.1 is an example of an Excel© spreadsheet format, with sample entries from the bibliography backend sheet for English non-fiction.

Server Model, Scripting and Coding

Server Model

We have used the popular LAMP server model, an open-source web development platform so called from the initial letters of its four components, **Linux**, **Apache**, **MySQL** and **PHP**.

Linux: Linux provides the open-source operating system for the site.

Apache: The Apache Web server is a public-domain open-source Web server in accord with the current standards of HTTP (Hypertext Transfer Protocol), the protocol underlying the World Wide Web.¹ The source code of Apache is freely

¹HTTP defines how messages are formatted and transmitted, and what actions Web servers and browsers should take in response to various commands. For example, when you enter <http://bichitra.jdvu.ac.in/> in your browser, this actually sends an HTTP command to the Web server directing it to fetch and transmit the requested Web page.

<div><div>Fiction</div><div>Drama</div><div>Non-Fiction</div></div> <div><div>D</div><div>A</div><div>B</div><div>C</div><div>D</div><div>E</div><div>F</div><div>G</div><div>H</div><div>I</div><div>J</div><div>K</div><div>L</div><div>M</div><div>N</div><div>O</div><div>P</div></div>				
Death threatens	▪ Death threatens, I will take thy dear ones.	—	▪ The Modern Review, November 1913	—
Defamation (The Crescent Moon 10)	▪ Why are those tears in your eyes	▪ REVEMS_369(4)	—	▪ The Crescent Moon (London: Macmillan, [Novem
		▪ HRVD_005		
Deshabandhu Chittaranjan Das	▪ Thy motherland spreads the veil	—	▪ Visva-Bharati News, Vol. IV, No. 1, July 1935: p. 1	—
Do not call me back	▪ Do not call me back	▪ REVEMS_210	▪ The Visva-Bharati Quarterly New Series, Vol. XIV, Part III, November 1948-January 1949	—
Do not insult thyself	▪ Do not insult thyself	—	▪ Visva-Bharati News, May 1933	—
			▪ The Modern Review, July 1939	
Do not tease my soul	▪ Do not tease my soul	▪ REVEMS_005	▪ The Modern Review, January 1932	—
		▪ REVEMS_123		

Fig.9.5 Full Table: section of page showing toolbar, links and icons

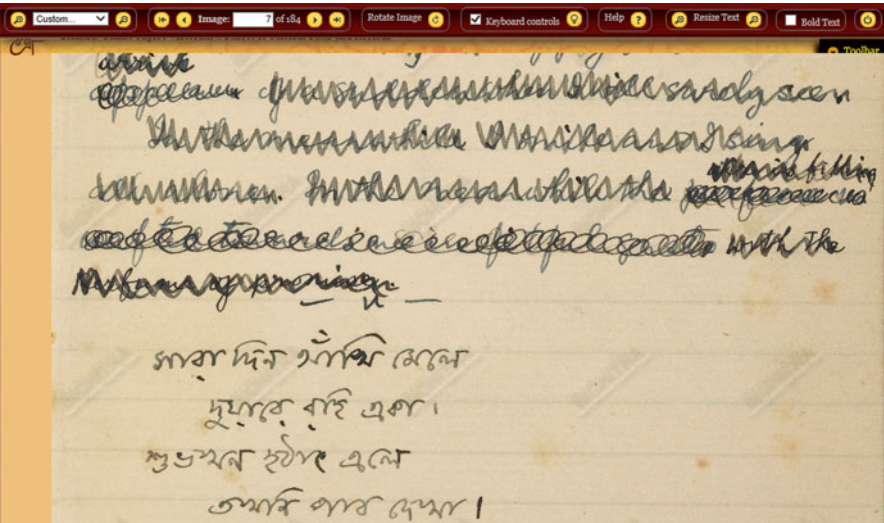


Fig.9.6 Image display page (enlarged section) with command bar

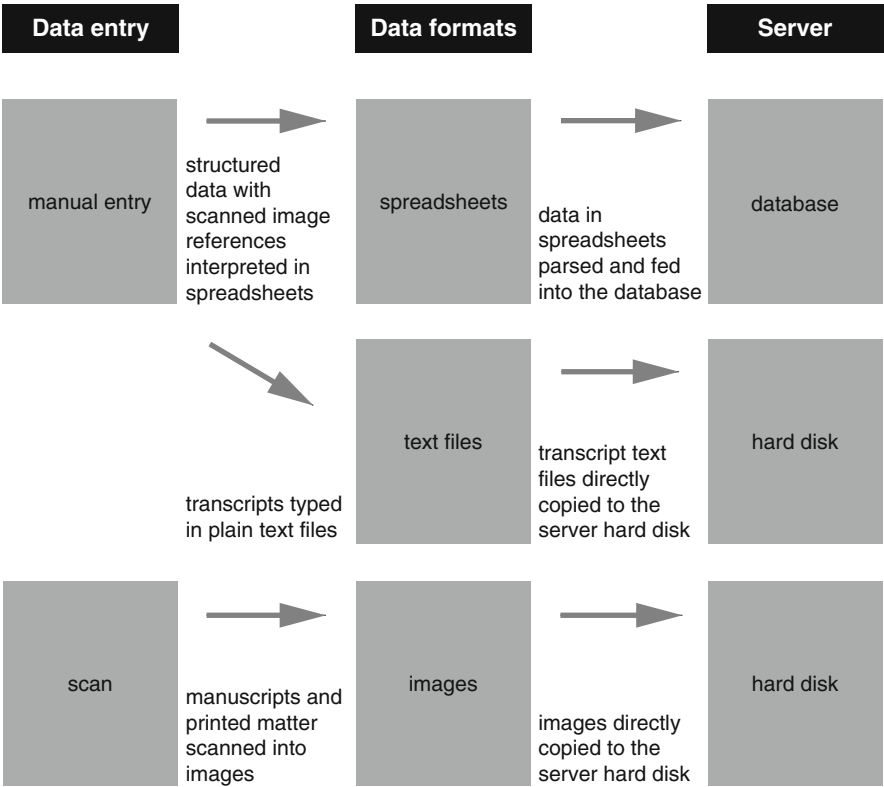


Fig.9.7 Entry and handling of data by server

Table 9.1 MS excel© spreadsheet format (English non-fiction)

Title: first words	Collation file	Manuscript/ typescript	Text file	Journal	Text file	Image folder
AUTHOR'S PREFACE: PERHAPS it is well for me to explain that the subject matter of the papers published in this book	english/non_fiction/ sad/ e_n_sad_000	-	-	-	-	-
The Relation of the Individual to the Universe: The civilization of ancient Greece was nurtured within city walls	english/non_fiction/ sad/ e_n_sad_001	HRVD_023*Titled 'The Universe as Alive' and 'World Consciousness'	e_n_sad_001_ H023 _m_01	The Modern Review, July 1913*Titled, 'The Relation of the Universe and the Individual'	e_n_sad_001_1913_j_01	e_n_ sad_001_1913_j_01

available, so that anyone can adapt the server for specific needs. There is also a large public library of Apache add-ons.

Database (MySQL): Setting up a website calls for a large body of concepts and techniques for managing the data, called the database management system (DBMS). The basic goal of a DBMS is to store and retrieve information efficiently and conveniently. It must also protect the information against system crashes and attempts at unauthorized access.

MySQL is an open-source relational database management system (RDBMS) based on Structured Query Language (SQL). A query language is a language in which the user requests information from the database. SQL (Structured Query Language) is the most widely used such language.

Server-side scripting (PHP): PHP (a recursive acronym for Hypertext Preprocessor) is a widely used open-source general-purpose scripting language, specially suited for web development and easily embedded into HTML. The PHP script is embedded within a web page along with its HTML. Before the page is sent to the user, the Web server calls PHP to interpret and perform the operations called for in the PHP script. After the PHP code is interpreted and executed, the Web server sends the resulting output to its client, usually as part of the generated web page. Thus the PHP code can generate a web page's HTML code, an image, or some other data.

The LAMP server model is easy and flexible to use, with robust scope for efficient data storage and representation. PHP handles the logical as well as the structural data representation, and connects with the MySQL database through the 'mysql' extension available for PHP. Formatted requests are formed through PHP and sent to the database to fetch the desired results. For example, a query like 'Fetch manuscript name from the manuscript table where work title is *Achalayatan*' is sent from PHP to the MySQL database, which in turn returns the results to PHP. The result set is then formatted according to the interface layout.

Our basic aim in designing the website was to keep the database in the best possible state for fast querying and data mining—in technical phrase, to maximize the input-output efficiency, yielding most data in least time through the smallest number of mouse clicks. To this end, we had to keep the physical storage of the data tables and associated indexes at optimal level at all times.

One of the most fundamental ways to improve querying speed relates to the structure of the database. Broadly speaking, the database comprises several inter-linked tables, each with a pre-allocated memory limit. Every time a record is added or deleted, the tables need to be reorganized and checked for several technical integrities. This process needs to be reviewed and optimized at regular intervals to reduce storage space, ensure best organization and automatically repair any discrepancies.

Client-Side Scripting and Other Strategies

With images and data running into terabytes, our task was to ensure a smooth user experience with minimal server response time. The process of data transmission between the server and the user had therefore to be minimized. To ensure this, we adopted the following strategies:

Table-less structuring: The Bichitra site uses no tables except the bibliographical Full Table and the Index of Manuscripts. In these two cases, scroll bars with a

fixed first column are used to allow access on screens of less than optimal size. For the rest, as in all modern websites, we avoided tables, as HTML pages in tabular form lose some of their formatting when viewed on screens of different sizes like smartphones and tablets.

AJAX: We made liberal use of Asynchronous JavaScript and XML (AJAX) to fetch the required dataset without reloading the whole page. AJAX is widely used on sites like Google Maps, Gmail, YouTube, and Facebook. Conventional web applications transmit information to and from the server using synchronous requests: that is to say, by directing the user to a new page with new information from the server. Here the whole page must be reloaded. AJAX allows web pages to be updated asynchronously—that is to say, in the background, without moving away from the main page—by exchanging small amounts of data with the server behind the scene. Hence it is possible to update parts of a web page without reloading the whole page (see Fig. 9.8). For instance, if you click on a particular letter in the Alphabetical Index of the Bibliography, you can see the search results without reloading the page.

jQuery®: Along with normal JavaScript, jQuery® javascript library is used throughout the site to improve functionality. Using JavaScript on a website becomes much easier with jQuery®. We can enhance the performance of the user interface without writing hundreds of lines of code. Moreover, jQuery® is fast and easily extensible to meet customized needs. It resizes the display of the website to suit different screen sizes to ensure a constant level of browser display—a specially important consideration for the bibliographical tables. (See ‘Table-less structuring’ above.) All this helps to make the web page as compatible as possible with the greatest range of browsers. Hence jQuery® is used by companies such as Google, Microsoft and Netflix.

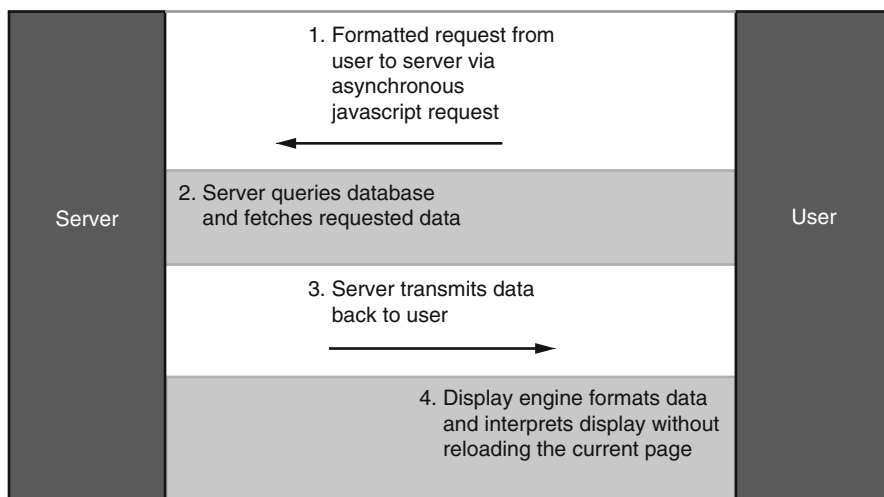


Fig. 9.8 Data request and supply using AJAX

Fig. 9.9 Style declaration for headings

```
.heading{
margin:0px;
font-size:12pt;
text-align:center;
letter-spacing:1px;
color:#3B0002;
font-weight:normal;
}
```

jQuery© is also used to provide the basic options of an image viewer with an optimized payload. It is jQuery© that enables the seamless traversal from one page image to another, as described above, without reloading the rest of the layout like the transcript panel and toolbar controls. It also enables functions like resizing and rotation, as well as the image slider on the home page.

CSS (*Cascading Style Sheets*): The Bichitra site is built with CSS hooks so that it can be readily redesigned by changing features like colours and backgrounds. Common points of style like text colour, font and alignment are defined in a single file, which is referenced through all the pages of the website, ensuring a uniform layout across the site. These style declarations are stored in the browser cache, so they do not need to be fetched from the server each time when loading a new page. Figure 9.9 shows an example of a style declaration for the text of headings. The entire set of features is stored in a single file, which is referenced through all the website pages.

The End in View

We had a twofold purpose in our website design: to attract casual browsers and inspire in them an interest in Tagore's works, but also to provide serious or scholarly users with varied material for their research. In the Bengali literary world, not all such scholars are at ease with computers: they needed an accessible and user-friendly interface. But casual browsers would often be sophisticated website users: the site had to meet their expectations as well.

Ritwick of Pixel Poetics, who masterminded the website design, described his role as that of the 'listening bird', in contrast to the singing birds creating the material. Yet his role was no less creative, for he had to bring order to this mass of material on his own terms—but also on *its* own terms, as the nature of the material determined the structure of the site. His dialogues with the file manager, image processors, and the search engine and collation programmers could fill another volume. As remarked elsewhere in this book, how vastly easier our task would have been, were we producing a series of offline CDs for separate titles! But as also remarked there, that would have defeated our purpose. Our material called for not only digital but cybernetic embodiment in dynamic and transmissible form: that is to say, a website. In Tagore's own words quoted on the Bichitra home page, we wanted to let the life breath of his music run from sky to sky.

Sukanta Chaudhuri

Improvements

We recognize that Bichitra could do with some facilities that it lacks. The principal ones, as we see them, are laid down below. We would be grateful to be told of more. And needless to say, we would be more than grateful for the funds and opportunity to add these features, and the more ambitious value additions contemplated in the next section.

Direct Access to an Image Within a Folder

Suppose you want to access page 216 of a 300-page book. As of now, you have to click through the images till you reach the one you want. There is a shortcut we have indicated in earlier chapters as a tip to users, but it is a tiresome process all the same. In any case, you can only reach a particular image, not a particular page as numbered in the book or manuscript. The page number may or may not be close to it.

An analogous problem is the need to access directly the page carrying a particular item (short poem, story or essay) in a large collection. With the help of students in our School's Digital Humanities course, we have set about ensuring this facility, though it may take some time and several classes to complete the job. The drama texts, at least in the standard version, should be processed within this academic year.

S. Chaudhuri (✉)
Department of English, Jadavpur University, Kolkata, India
e-mail: schaudhuri@english.jdvu.ac.in

Simultaneous Scrolling of Image and Transcript

Again, users may find it a nuisance to scroll through the continuous text of the transcript till they reach the page they want. Ideally, the transcript should scroll simultaneously with the images, automatically opening at the point transcribing a particular page. Technically, this is a simple application, but a laborious one to put in place for such a large corpus.

Integrating the Stand-Alone Features

Bichitra is an integrated website—that is to say, it was created in one go, and its various parts are linked to and accessible from each other in very large measure. But a few items in the menu are stand-alone, like the Index of Alternative Titles, the Contents of Collected Works, and the Timeline. Either these were late additions, or else they involved more work on the site structure than was feasible within our time-span. But it would obviously be a good idea to link these items with the main bibliographical tables and indexes, not to mention the images and reading texts.

Value Additions

The above points would improve the site within its present structure and purpose. But Bichitra can provide a launching ground for exciting new ventures in textual computing. Its priceless asset is a large corpus of texts to be mined for all kinds of data, which can then be analyzed in various ways.

In what follows, ‘we’ does not refer to the Bichitra team or the School that houses the project. We can hardly explore on our own the endless horizons opening up in various directions. These are adventures for the whole community of digital humanists. We are proposing a common endeavour in which we too hope to play a part.

Empowering Non-Latin Fonts

Users of Bichitra face the same problem as its makers: the texts are in a non-Latin font, whose characters (as explained in Chap. 3) work in a very different way from the Roman or English alphabet. Most major tools for textual computing and data mining are not adapted to non-Latin fonts. But this challenge offers a great opportunity: the Tagore corpus can provide a laboratory for extending a whole raft of textual computing programs to new languages. It can be no more than a model: it will need fresh adaptation even for other abugida alphabets, let alone those organized on different principles. But even that model would be a major step in integrating all the world’s languages into the computerized universe—not just for collecting numerical or other textually restricted data, but for processing in all languages their full range of texts of whatever size and complexity. Such a development would truly unleash the potential of the Unicode standard.

In everything said below, this crucial factor must be kept in mind. Whatever we do by way of mining and analysing the text implies a second endeavour in constant parallel: the major empowerment of a non-Latin font for such purposes. This will remain as a lasting benefit of the exercise, irrespective of our success in terms of data mining and textual analysis. In fact, we have already gone beyond extending programs from the Latin to a non-Latin font: the programs we have created, above all the collation program Prabhed, were originally created for texts in the Bengali font and only later applied to English texts in the Latin or Roman font. These may be the world's first textual computing programs to use a non-Latin font as their primary medium.

Multimedia

Bichitra is already a multimedia database insofar as it incorporates a very large store of still images. It also uses some fairly complex diagrammatic visualization in its collation program. It would obviously gain from two other types of multimedia intervention: video and audio clips, separately or in combined audio-visual form.

A notable attempt at such integration was made by Spandana in her Ph.D. dissertation, approved in January 2014. She carried out a textual study of a group of intermeshing Tagore plays: *Raja* (*The King*, published in translation as *The King of the Dark Chamber*), *Arupratan* (*The Invisible Jewel*) and *Shapmochan* (*The Lifting of the Curse*). Besides a genetic and stemmatic study along traditional lines, she created a digital platform comprising a full multimedia editorial program for dramatic texts. Its opening page is modelled after the traditional stemma charting the relation between various versions of the work (in this case, several related works, each in several versions). By clicking on various icons, this page opens others with images and transcriptions of all the versions; collation results using Bichitra's collation program Prabhed; links to audio, visual and audiovisual clips; and annotations, with the option of editorial intervention through a Notes Editor. All these functions are exclusive to this framework, which Spandana created using open-source softwares and libraries like Dojo toolkit, JQuery© library, and other applications of JQuery©.

Spandana calls this editorial platform 'Pathdarpan' or 'Textual Mirror'. It is yet another program created for Bengali material and so far only available with Bengali apparatus. We hope it finds wider applications before long, and opens the way to still more advanced multimedia editorial software.

Topic Modelling

Topic modelling is the type of computer operation that examines the frequency of certain sets of words in a corpus of texts, with a view to determining the topics common to them. It thereby allows us to detect the subjects or concerns operating in a discourse. Conversely, if we know what those subjects or concerns are, it allows us to map in greater detail the vocabulary associated with them, hence what further concerns or associations the discourse might carry.

Proceeding beyond this familiar ground, we may think of determining the common concerns and contrasts between discourses on different matters or of different periods or provenance. We can try to model ‘topics’ on a syntactic or grammatical rather than semantic basis: that is to say, the way sentences are structured or, more broadly, words are related to other words. Grammar and syntax relate not only words but the ideas embodied in those words: the way those relationships are defined within the sentence, or in a text composed of sentences, indicates the way the writer views the elements making up his or her world. In broad terms, a language embodies a world view. This kind of enhanced topic modelling might help to define that view. The next section explores these possibilities.

New Directions

[This section may have succeeded in being outdated and futuristically speculative at the same time. It consists entirely of Sukanta’s thoughts, and is therefore phrased in the first person singular.]

Brief though they are, the above remarks on topic modelling ended on a speculative note. There is also a speculative dimension to our further thoughts and ventures in the field of collation, as outlined at the end of Chap. 8. Sooner rather than later, these new ideas about collation take on applications beyond the comparison of texts. At their most ambitious, they envisage a comprehensive means of comparing not only words or text blocks in natural languages, but any kind of strings denoting entities of any nature in all their combinations and variations. Once we devise a basic logic for such an operation, it can be adapted to all kinds of material besides natural languages. Among the latter, it would need to be modified for different languages, or different states of the same language. The last case (different states of the same language) would relate to the task of paraphrase modelling, and perhaps extend the scope of the latter by indicating equivalences (and differences) between different historical states or registers of utterance.

It may be my own focus on language that makes me feel that finally, natural language will provide us with the most complex and challenging applications of any such exercise. I have had stimulating if casual discussions with biologists on the structural similarities and differences between verbal texts and DNA molecules or other biochemical structures and compounds. A theoretical life scientist pointed out to me that all proteins are composed from 20 amino acids, each in three classes. Why, she asked, should the permutations of these 60 base units be less complex than those of the 26 or (counting capitals) 26×2 characters comprising the Roman alphabet, even if you throw in the punctuation marks?

My response was that characters are not organic to the word as amino acids are to the protein. A word can differ in spelling or phonology, beyond what would be acceptable as mutant forms of the ‘same’ biological unit. Its semantic identity—in plain language, its meaning—is unstable, and its many possible syntactical functions can support an infinity of potential sentences. Words built of characters, and sentences built of words, represent endless syntheses of non-homogeneous elements.

I have stated my argument in grossly simplified and naïve form. But it may help to indicate the literally endless refinements possible in a collation program for a natural language. More fundamentally, it indicates the theoretical possibilities of such a program. There is a recognized (though relatively recent) branch of conventional textual studies called editorial theory. How far will this extend to take in a brave new world of theoretical digital humanities? Such a pursuit might lose itself in vapid speculation. On the other hand, it could give the necessarily data-specific substance of digital humanities a firm direction and perspective, a supporting structure of intellectual principles to validate the algorithms on a broader conceptual plane. A certain philosophic grounding never does a discipline harm.

I may carry the argument a stage further. *Mutatis mutandis*, the open-ended divergence and contingent nature of textual data is found in all humane or human data, in vastly bigger sets: as relating, for instance, to epidemiology, disaster prediction and management, surveillance, finance, administration and welfare. The bed-rock unit of measurement of all such data is the human being, each one unique in constitution and circumstance, and further changing over time—not only from age to age but ‘from day to day, from minute to minute’, as Montaigne remarked (‘Of Repentance’, Montaigne 1958, 611). No wonder we find the same endless variation in the language in which human beings express their lives. Hence advances in textual data mining may have crucial implications for the interpretation of all big data.

‘Big data’ refers to the unprecedentedly vast, often uncoordinated data generated in far-ranging, often global databases: the immense daily grist of the Google mills, or the ‘raw’ databases of international media, finance and surveillance. Needless to say, even the biggest textual databases are minute in comparison: we have to scale down the application of the term ‘big data’ till it seems scarcely applicable. For big textual corpora, we may prefer Hope and Witmore’s more cautious term, ‘Very Large Textual Objects’ (Hope and Witmore 2004). The paradox of textual non-big big-data banks can offer a special advantage in investigating big-data management generally.

Because of its immense extent, big-data mining, almost by definition, is not fine-tuned nor analyzed at the point of initial processing: there is just too much of it. It is profiled by quantity and location—that is to say, by patterns of occurrence, by ‘where’, ‘what kinds’ and ‘how many’ but not ‘precisely what’ or ‘why’. This is clearly an inadequate way of processing any data, but its sheer volume makes closer analysis impossible at the point of first access, both for the computer and the human brain.¹ With textual data, however, such heuristic processing may prove specially futile if not extended to the ‘precisely what’ at a very early stage. The reason, as stated above, is the infinite syntheses of non-homogeneous elements in textual data.

¹Programs like ‘Knowledge Vault’ (Dong et al. 2014) parse big data semantically and/or syntactically to extract knowledge, but they chiefly extract the data automatically by measuring multiple occurrence. Google’s Knowledge Graph goes far in this direction. But all such programs extract knowledge *out* of the text, rather than probe more deeply *into* the verbal dimension per se. Their thrust is informational, not textual. Yet such in-depth textual analysis would be invaluable in creating a new generation of knowledge-extraction programs.

Ultimately, this applies to the substance of all or most other big data. But textual data may offer a feasible key to resolving this dilemma of unmanageably large accretions of uniquely small units.

Traditionally in pre-computer times, we analyzed texts by paying heed to their unusual elements, valuing any component insofar as it stood out by its rarity—either within the work, or across a body of works (say, the special components in the writings of an outstanding author). This corresponds very roughly to the statistician's concept of the 'long tail'; but here the tail was considered the most important appendage of the textual beast. I am not talking of authorial originality but of the words constituting the text. Hope and Witmore talk of 'the relative infrequency of the salient items we [usually or traditionally] select in order to understand or interpret a text', making us 'cognitively biased towards the unique' (Hope and Witmore 2004, 2). We focus on three oak trees in a wood rather than the 1000 birches surrounding them. Instead, the big-data approach demands (and technologically enables) equal attention to all components of the corpus.²

Yet this is and is not the case. To mine or analyze data at all, we have to search and select, play off matches against differences, weigh x against whatever is not- x . All data mining assumes this implicit deconstruction. Even when tallying a 's and the 's, we want to know their relative frequency and distribution: i.e., how they stand in relation to everything else. The more we refine a search, it approaches an exercise in collation, just as (more obviously) collation is a species (or many species) of advanced search. It has been argued that the search function is fundamental to all computer operations. In textual computing, we may lay special stress on the sub-function of collation in addition to 'direct' search.

With Very Large Textual Objects, the field of search-cum-collation expands incrementally. We now have a wood with 1000 oaks and 100,000 birches, or 100,000 oaks and 10,000,000 birches. We end up by counting oaks in the same way that we count birches. It may even be hard to tell one from the other in a distant perspective. Alternatively, we may have a wood with 1000 birches, 97 oaks, 15 elms, six beeches and 20 pines. The 'tail' has got longer, and needs to be interpreted differently from a simple oak-and-birch situation.

The great benefit accruing to textual studies from mining Very Large Textual Objects lies in an exercise hitherto impossible in practice even if we conceived of it: what Franco Moretti calls 'distant reading' (Moretti 2005, 1), as opposed to the familiar exercise of 'close reading'. (See also Mimno 2012.) Distant reading looks at the production and dissemination of a very large number of works over a long expanse of time and/or space; and by organizing this great mass of data, draws wider conclusions about historical and cultural processes than unaided human or 'manual' study could possibly attain. '[B]ig data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value' (Mayer-Schönberger and Cukler 2013, 6).

Distant reading is one of the clearest examples of how, simply by solving the external or mechanical challenge of quantity, computers can bring about qualitative

²In fact, infrequent items may be excluded on the grounds of small sample size (Mimno 2012, 5).

changes in the agenda of a discipline. Literary historians have traditionally focussed on a relatively small range of material: a single book or author, at most a single genre, theme or period, and within that a (usually canonical) fraction of the total output. As the range or scope of the study expands—beyond author to genre, beyond genre to period—the number of works studied closely decreases in inverse proportion. But by storing the full data and metadata of the total corpus in a computer database, we can mine it to grasp the big picture of textual production and textual transactions.

When marshalling an indefinitely large body of complex material, we can logically adopt one of two courses. Either we collect a large and varied body of discrete data, and select the closest match to a given case; or we abstract general patterns or principles from the data, and apply them to the case. This tension between the atomistic or accretive approach and the generalizing or abstractive approach underlies many paradigms of knowledge: the medieval debate between nominalism and realism, the relation of observation to law in the natural sciences, ultimately the basic opposition of induction and deduction. We noted the contrast between heuristic and holistic approaches in the context of collation programs in Chap. 8. The tension takes its distinct form in the humanities, whose objects of study are exceptionally various, undefined and open-ended, though often reducible to a finite number of elements: the letters of an alphabet or the notes of the musical scale. The annotated corpora used for natural language understanding systems are usually quite small in size. But their possible alternative components are effectively infinite. Further, each of these alternatives is unique, and significant by virtue of its uniqueness. Hence a complementary type of corpora is needed to populate the annotations.

We are back at Hope and Witmore's point about three oaks versus a thousand birches. But instead of contrasting the rare or unique with the ubiquitous, I want to draw them together. At one level (most practised to date, sometimes almost trivially), distant reading concerns itself with metadata alone: when, where and by whom the texts were created, and how they circulated. The model for such large-scale inquiries might be the varied output of the Stanford project 'Mapping the Republic of Letters'. But the most valuable results will emerge when we proceed from the metadata to the content of the texts. This is already happening in a rudimentary way in databases like the Iraq War Logs, where each document is categorized by 'the three most "characteristic" words in that report' (Stray 2010). The 'Read the Web' program under development at Carnegie Mellon University aims to raise the scale to the epic dimensions of the entire Web. (See Read the Web.)

With a sufficiently versatile data mining program along the lines indicated below, we could sift the content in great detail: distant reading and close reading would come to merge. Even after raising the stakes to 1000 oaks and 100,000 birches, can we possibly—just possibly—look at every one of those oaks, indeed those birches, as a distinct organism, a unique manifestation of life—in our metaphoric context, the life of utterance and communication?

This does not assume a romantic notion of originality or individuality, let alone individual talent. It only implies an intrinsic factor of contingency, the unique position in time and space occupied by each text, artefact or other human construct; and,

within it, the unique location and context of each component: words, musical notes, but also actions, behaviour and observances, ethical and intellectual constructs. In other words, this premiss applies not only to works of art but to all expressions of the human consciousness, the material of psychology and the social sciences as well as of literature, music or the visual arts.

Having reached the brink of a very deep precipice, let me pull back to the safer ground of hands-on textual computing. Work is under way on much of what I suggest below: sometimes in mature programs, more often in ones under development or at an experimental stage. I have not made any survey of such activities. In what follows, I am only drawing up a wish list for textual data mining.³

The basic object of virtually all these inquiries is **fine-tuning of contextual data**: assessing the significance of *each separate use* of a term by noting precisely where it occurs, and what other terms occur around it. This is where we have hitherto most presumed on the rarity of the search term. In traditional textual analysis, one could manually locate and contextualize every occurrence of what we considered the salient term or terms. It was sometimes long and hard work, but it could be done. With Very Large Textual Objects, a term or feature that is rare in proportional occurrence can be numerous in overall count (100,000 oaks amid 10,000,000 birches). There may be *so many* instances of the search term that contextual distinctions get blurred. To restore these distinctions must be the guiding aim of advanced data mining. Such mining can be at three levels.

(A) By the number and location of words, taking them at face value. We may look for

- Specific words or word groups in exact forms.
- Variant forms of words or word groups, through wildcard and fuzzy searches.
- The contexts and associations of words, through proximity searches.
- Compounds, phrases and word clusters, which are really special types of proximity.

All these searches can be either specific, or non-specific and open-ended. We can specify search terms in exact form or through wildcard and fuzzy searches. Or we can set broad parameters and see what the program trawls from the database.

We are progressing from word searches per se to topic modelling, identifying recurrent clusters of words across a corpus from which we can deduce the subject or subjects of discourse. Once we have identified the topics, we can carry out proximity searches to map the wider vocabulary associated with them, hence the further concerns and associations of the discourse. Going farther still, we may think of combining or comparing the topics found in discourses on different matters or of different periods or provenance. We can build up to a study of changing discourses and epistemologies across time, geographical space, disciplines and epistemic categories. Hence, conceivably, we can redraw their boundaries.

³See Mimno 2012 for an illustration of some basic possibilities in practice, especially the 'broad areas' indicated on page 17.

(B) By syntactic function and grammatical structure.

As suggested earlier in this chapter, we might also investigate such ‘topics’ on a syntactic or grammatical basis. The meaning of a word is not limited to its isolated definition in a dictionary. It acquires further meaning in interaction with other words in a sentence, or in the larger context of a text. As each sentence has a unique combination of words, the function of each word within it is necessarily unique, if only minutely so in most cases. No less importantly, the way the words are ordered within a sentence reflects the way ideas are related to each other. Different languages relate words—that is to say, ideas and experiences reflected in words—in different ways: different languages *think* differently. A language embodies a world view. So do different kinds of discourse in the same language: a poem, a scientific treatise and a sports report structure the language differently, because they look differently at the world and define their creator’s position in different contexts of utterance. The context of utterance also distinguishes spoken from written language. Such stylistic variance is a concern of the traditional subdiscipline of stylistics. Computers have enabled us to make stylistic analysis in minutely specific detail. The more we map the intricacies of textual composition, the better we will understand how language works—not only in general or abstract terms, but in terms of specific topics, texts and discourses. This last difference indicates the basic contrast between linguistic and textual computing.

(C) By more complex semantic criteria.

1. We can extend the operation of (A) above through semantic and not merely literal searches—i.e., by progressing to the mapping of concepts. ‘Concepts’ does not only mean abstract ideas, but the ideas of anything, even material objects. It is these ideas that provide the currency of verbal communication. Words do not refer to things but to the concepts of things: that is a fundamental premise of modern linguistics. The word *book* (whether the sounds making up the word, or the letters representing the sounds) does not relate to a particular object but to the idea or concept of a class of objects.⁴ The human brain translates material objects into concepts or mental entities in the process of apprehending them. It is these concepts that are fed by the human brain to the computer: either through verbal input, or through recognition programs whereby the computer views an image, or even an actual object through a camera, scanner or other device, and links the object to the idea. The computer apprehends that idea in terms of its own binary code, as the human being does through the spoken or written word. The computer may, of course, link its own formulation to the verbal medium for the benefit of human users.

To advance the computer’s understanding, it needs to grasp these ideas in as wide and complex a form as possible. I have just said that humans think in words, but we also think *beyond* words, in terms of categories demarcated

⁴ ‘The linguistic sign unites, not a thing and a name, but a concept and a sound-image.’ (Saussure, 1975, 66). This notion underlies Ogden and Richards’s ‘semantic triangle’ linking an object, the thought of that object, and the word or sign representing it (Ogden and Richards 1923, 5).

by a range of words. (On a different level, that category might have a word to indicate it.) The word *dog* relates to the same category as the word *canine*. But *canine* also extends outward to take in creatures like the wolf, fox and jackal. And *dog* extends inward, if I may so phrase it, to take in creatures like the spaniel, labrador, terrier and mastiff. When we hear of any of these breeds, we automatically relate it to the category ‘dog’.

Computers can make all these associations if provided with suitable lexicons, which must reach far beyond humanly pre-set wordlists and dictionaries to results drawn from the proximity, alternation and variation of words in ‘big-data’ textual corpora. Programs for advanced semantic parsing attempt go further than identifying topics in terms of complex ontologies.⁵ They place the semantic value of textual components against their syntactic values—very simply put, the meanings of words against their position and function in the sentence. But finally, these results can only be analyzed through human intervention. In any topic modelling exercise, the computer can identify related groups of words; the topic they represent must be validated by the human brain. (Sometimes there proves to be no viable topic at all.) There are immense possibilities here of continuing human-computer interaction. With more than a quibble on words, we may see this as a lesson all computer users can draw most particularly from digital *humanities*.

2. There is a ‘text analysis environment’ named ‘Docuscope’, created by David Kaufer and Suguru Ishizaki, that groups words not by semantic or grammatical function but by rhetorical categories—which can only be determined by assessing human response. It allows us to ‘annotate a corpus of text against any dictionary of regular strings that are classified into a hierarchy of rhetorical effects’ (see [Docuscope](#)). Its pioneering critical application, to an analysis of Shakespeare’s plays (Hope and Witmore 2004, 9–29), uses a ‘generic dictionary’ classifying the English vocabulary in rhetorical or affective categories through human judgment as validated by multiple testing and consultation (whose efficacy can, of course, be contested). It then maps the presence of these various categories in a text to help analyze its nature and purpose.

The point is not the efficacy of a particular application but the immense potential of the tool. It allows us to carry out semantic mapping at a connotative level, going beyond even the most advanced ontology based on factual or denotative criteria.⁶ Through continual mining of textual corpora, we can hope to locate more and more accurate and impersonally demarcated categories of utterance, defined in connotative or associative rather than denotational terms. We can thereby obtain a new set of maps of the same

⁵ See, e.g., Kwiatkowski et al. 2013, Das et al. 2014.

⁶ Even the elaborate ontology of the ambitious ‘Read the Web’, designed to comb the resources of the entire Web, currently comprises factual or firmly denotative categories alone (Read the Web, <http://rtw.ml.cmu.edu/rtw/kbbrowser/>)—almost necessarily in view of the vast corpus involved.

textual territory, just as one might map a geographical territory in terms of, say, its topographical contours rather than political boundaries.

3. Another major line of potential data mining would be to assess the metaphoricity of language. For this we would need to model a different set of topics, defined by the contextual proximity or affinity of words from different categories in terms of denotation or 'literal' meaning. Metaphoricity is not only a local phenomenon. Discourses and entire disciplines, the operation of language itself, are governed by metaphors so embedded as to be invisible. Of every hundred uses of *see* or *stand*, ten or less—maybe none—might refer to those physical actions; the rest will be metaphoric in one way or other. A scholar of the Renaissance pointed out to me how the medical term *remedy* was commonly used in that age (and indeed others) in the contexts of theology and law. *Pastoral*, literally relating to sheepkeeping, is more often used (through the influence of the Christian gospels) in contexts of church governance.

If large textual corpora could be comprehensively mined for metaphoricity, it would revolutionize our understanding of language. It is the subtlest and most ambitious of the inquiries I have suggested here.

All these exercises involve nothing more nor less than asking the computer to do more of the data mining that we now conduct manually. By the creative tautology of their being, computers can compute. They can count and calculate—at a speed, across a range, and with a degree of accuracy unthinkable by any other means. But all their miraculous output is finally nothing but counting and calculating. The clarity of that output precludes any bias in its presentation. It therefore allows humans to make clearer and more unbiased interpretations of the data than would otherwise be possible. To help us do so (or conversely to bear out our own biases), we can program the computer so as to present the output in the form best suited to our needs. But it is for us to interpret that output, taking care that the thrust of the program (which too is a human construct) does not unduly condition our interpretation.

As recounted at the start of Chap. 8, the man who gave me my first computer lesson began by warning us that the computer is an idiot. The idiot status of the computer is a boon, for it demands a painfully precise articulation of every minute step in its operation, fully and exactly defined, excluding all other possibilities. Much more evidently than natural language, the language of mathematics provides the most important field of integrated formal articulation. A mathematical formulation is precisely itself and nothing else. Digital humanities mathematically computes types of data that we have hitherto absorbed (and still commonly do) in a more impressionistic way. The mathematics is itself the tool of a still more fundamental algorithmic logic.

Thus digital humanities provides the notoriously intractable material of the humane disciplines, not to mention the disordered stuff of human lives, with a logical and mathematical underpinning. It forces us humans to think more fully and precisely about the terms of our inquiry. We can then ask the computer to provide

the relevant data, sorted and processed, whereby *we* can test those terms. As with all empirical testing by whatever means, the data may or may not bear out the hypothesis underlying our terms of inquiry. We thereupon examine the computer's findings, assess them by our judgement, and accordingly frame the next question. We thus engage in a continuous dialogue with the computer.

Eventually, this passes into a dialogue with ourselves. We assess the answers we obtain to the questions we have ourselves put to the computer. This is no different from traditional humanistic inquiry: our analysis of texts has always stemmed from our own epistemic position, and helped to define it further. The large textual database affords unprecedented resources for carrying out our traditional tasks, if now defined in new ways. It also puts those exercises in tandem with others *conceptually* possible only in the new perspectives opened up by the computer. To apply Katherine Hayles's term in a new context,⁷ the computer is not only the instrument but a 'material metaphor'—or should we say a virtual metaphor?—for the very practice of the digital humanities.

References

- Das, Dipanjan et al., 2014. Frame-Semantic Parsing. *Computational Linguistics* 40:1, 10–56. http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00163. Accessed 12 January 2015.
- Docuscope. <http://www.cmu.edu/hss/english/research/docuscope.html>. Accessed 24 December 2014.
- Dong, Xin Luna et al. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. *Proceedings on the Conference for Knowledge Discovery and Data-Mining*. <https://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf>. Accessed 12 January 2015.
- Hayles, N. Katherine. 2002. *Writing Machines*. Cambridge, Mass: MIT.
- Hope, Jonathan and Michael Witmore. 2004. The Very Large Textual Object: A Prosthetic Reading of Shakespeare. *Early Modern Literary Studies* 9.3/Special Issue 12, January. <http://extra.shu.ac.uk/emls/09-3/hopewhit.htm>. Accessed 24 December 2014.
- Kwiatkowski Tom et al. 2013. Scaling Semantic Parsers with On-the-Fly Ontology Matching. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, October: 1545–56. <http://homes.cs.washington.edu/~lsz/papers/kcaz-emnlp13.pdf>. Accessed 12 January 2015.
- Mayer-Schönberger, Viktor and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.
- Mimno, David. 2012. Computational Historiography: Data Mining in a Century of Classics Journals. *Journal on Computing and Cultural Heritage*, 5:1, 1–19.
- Montaigne, Michel de. 1958. *Complete Essays*. Trans. Donald M. Frame, 1957. Stanford: Stanford University Press.
- Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso.
- Ogden, C.K. and I.A. Richards. 1923. *The Meaning of Meaning*. London: Kegan Paul.
- Read the Web. <http://rtw.ml.cmu.edu/rtw/>. Accessed 12 January 2015.
- Saussure, Ferdinand de. 1975. *Course in General Linguistics*. Trans. Wade Baskin, 1959. London: Fontana/Collins.
- Stray, Jonathan. 2010. A Full-Text Visualization of the Iraq War Logs. <http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs>. Accessed 12 January 2015.

⁷Hayles 2002: see especially Chap. 2.

Appendix: The Bichitra Team

Adviser: Sankha Ghosh

Project Co-ordinator: Sukanta Chaudhuri

Executive Heads: Subha Chakraborty Dasgupta, Samantak Das

Technical Credits

- Electronic master text of Tagore's collected Bengali works prepared by Anirban Raychaudhuri using an OCR program developed by him under a Jadavpur University project.
- Base collation program prepared by Siddhartha Chaudhuri. Full collation engine developed and interface prepared by Sunanda Bose with guidance and assistance from Arunasish Acharya (School of Education Technology, Jadavpur University), Sukanta Chaudhuri and Spandana Bhowmik.
- Search engine prepared by Arabinda Moni and Prakash Koli Moi of the Department of Computer Science & Engineering, Jadavpur University, under the guidance of Chandan Mazumdar and monitoring by Dibyajyoti Ghosh. Search engine index prepared by Purbasha Auddy (Bengali) and Debapriya Basu (English).
- Transcription filter for extracting the final reading text of a manuscript prepared by Bhupati Roy of the Department of Computer Science & Engineering, Jadavpur University, under the guidance of Arunasish Acharya and Dibyajyoti Ghosh.
- Avro keyboard software prepared by Omicron Lab, Dhaka, Bangladesh, and gratefully used in terms of their Creative Commons licence.
- Bengali bibliography and index of manuscripts prepared by Purbasha Auddy, assisted by Subrata Sinha, Smita Khator, Sahajiya Nath and Rajib Kundu.
- English bibliography and master list prepared by Debapriya Basu, assisted by Debapratim Chakraborty and Lav Kanoi.
- Hindi website text prepared by Lav Kanoi.
- Advice on servers and hosting from Chandan Mazumdar, Department of Computer Science & Engineering, Jadavpur University. Server operations by Samit Pahari and Shyamal Bose.

- Website development and database design by [Pixel Poetics](#) (Chandannagar, West Bengal).
- Website design and logo design by Pinaki De.
- Website management by Dibyajyoti Ghosh, Purbasha Auddy, Ritwick Pal, Kawshik Ananda Kirtania and Lav Kanoi.

Supervisors

Purbasha Auddy, Spandana Bhowmik, Dibyajyoti Ghosh, Smita Khator, Debapriya Basu, Rianka Roy, Hrileena Ghosh

Project Operators

Amritesh Biswas, Anupam Roy, Anwesha Sarkar, Anwita Nandi Chowdhury, Aparupa Ghosh, Aritra Chakraborti, Arnab Ghosh, Arpan Das, Arunava Banerjee, Ashoktaru Panda, Baisakhi Topdar, Bithika Sahana, Daya Chatterjee, Debapratim Chakraborty, Debdip Dhibar, Debotri Ghosh, Gourab Chatterjee, Ishita Basu Mallik, Kawshik Ananda Kirtania, Lav Kanoi, Manasi Roy Chowdhury, Meghdut Rudra, Moumita Banerjee, Partha Sarathi Nandi, Piyali Chakraborty, Rajib Kundu, Raktim Sarkar, Riya Biswas, Rohan Islam, Sahajiya Nath, Saikat Banerjee, Samarpita Ghatak, Sambuddha Ghosh, Semonti Neogi, Shubhayan De, Somnath Chakraborty, Sriyayee Bhattacharjee, Sujata Datta, Swagata Roy, Swarnali Barik, Tayana Chatterjee, Tuhin Bhattacharya, Unmesh Mandal, Vinayak Das Gupta

Administrative Staff

Anuradha Mondal, Indrani Burman

List of Tagore's Works Cited¹

Printed Works: Bengali (Poetry: Volumes)

Arogya (Recovery). 1347/1941. Kolkata: Visva-Bharati.
Gitanjali (Song Offerings). 1317/1910. Kolkata: Indian Publishing House.
Janmadine (On My Birthday). 1348/1941. Kolkata: Visva-Bharati.
Kabi-kahini (The Story of a Poet). 1284/1877-8, *Bharati* (journal). 1285/1878, Kolkata: Prabodhchandra Ghosh.
Katha o kahini (Legends and Tales):
Katha, 1306/1900, Kolkata: Adi Brahmosamaj Press; 1310/1903, 'Katha' & 'Kahini' sections in the collection *Kabya-grantha*, vol.5, Kolkata: Majumdar Library; combined *Katha o Kahini*, 1315/1908, Allahabad: Indian Press, & Kolkata: Indian Publishing House.
Kshanika (Momentary Pieces). 1307/1900. Kolkata.
Lekhan (Writing). 1334/1927. Berlin/Balatonfüred (Hungary).
Mahua. 1336/1929. Kolkata: Visva-Bharati.
Manasi (The Woman of the Mind). 1297/1890. Kolkata: Adi Brahmosamaj Press.
Prabhat-sangit (Morning Songs). 1290/1883. Kolkata: Adi Brahmosamaj Press.
Prantik (On the Margins). 1344/1937. Kolkata: Visva-Bharati.
Punascha (Postscript). 1339/1932. Kolkata: Visva-Bharati.
Rogshajyay (On My Sickbed). 1347/1940. Kolkata: Visva-Bharati.
Sonar tari (The Golden Boat). 1300/1894. Kolkata: Kalidas Chakrabarti.
Sphulinga (Sparks). 1352/1945. Kolkata: Visva-Bharati.
Utsarga (Dedications). 1321/1914. Kolkata: Indian Publishing House.

Printed Works: Bengali (Poetry: Collections)

Chayanika (Selections). 1316/1909, Allahabad: Indian Press, & Kolkata: Indian Publishing House.
 3rd edition, 1332/1926, Kolkata: Visva-Bharati.
Sanchayita (The Store). 1338/1931. Kolkata: Visva-Bharati.

Printed Works: Bengali (Poetry: Individual Poems)

'Anadrita' ('The Unloved'). 1300/1894. *Sonar tari*.
 'Barshar dine' ('On a Rainy Day'). 1297/1890. *Manasi*.
 'Nirjharer svapnabhanga' ('The Spring Wakes from Its Dream'). 1289/1882, *Bharati* (journal). 1290/1883, *Prabhat-sangit*.
 'Parishodh' ('Reparation'). 1306/1900. *Katha*.
 'Prana bhariye trisha hariye' (translated as 'More life, my lord, yet more'). 1319/1912, *Tattvabodhini patrika* (journal). 1320/1914, *Prabasi* (journal). 1321/1914, *Gitimalya (Garland of Songs)*, Kolkata: Indian Publishing House.
 'Samukhe shanti parabar' ('Ahead, the ocean of peace'). 1348/1941, *Prabasi* (journal). 1348/1941, *Visva-Bharati News* (journal). 1348/1941, *Shesh lekha (Last Writings)*, Kolkata: Visva-Bharati.
 'Tumi sandhyar meghamala' ('You are like a cloudbank in the evening'). 1305/1898, *Bina-badini* (journal). 1307/1900, *Kalpana (Imaginations)*, Kolkata: Adi Brahmosamaj Press.

¹Dates beginning '12' or '13' refer to the Bengali era. CE years have been given alongside. As years in the two eras do not match exactly, the gap between the two can vary.

Printed Works: Bengali (Drama)

Achalayatan (The Inert Bastion). 1318/1911, *Prabasi* (journal). 1319/1912, Kolkata: Adi Brahmosamaj Press.

Arupratan (The Invisible Jewel). 1326/1920. Kolkata: Chintamani Ghosh.

Bisarjan (Sacrifice). 1297/1890. Kolkata: Adi Brahmosamaj Press.

Chirakumar sabha (The Society of Celibates). First published in dramatic form, 1332/1925, Kolkata: Visva-Bharati. Published earlier as prose fiction.

Guru. 1324/1918. Kolkata: Indian Publishing House.

Paritratan (Salvation). 1336/1929. Kolkata: Visva-Bharati.

Prayashchitta (Penance). 1316/1909. Kolkata: Hitabadi Press.

Raja (The King). 1317/1911. Kolkata: Indian Publishing House.

Raktakarabi (Red Oleanders). 1331/1924, *Prabasi* (journal). 1333/1926, Kolkata: Visva-Bharati.

Shapmochan (The Lifting of the Curse). 1338/1931. Santiniketan: Students' Committee for the Tagore Festival.

Shyama. 1346/1939. Kolkata: Visva-Bharati.

Tasher desh (The Land of Cards). 1340/1933. Kolkata: Visva-Bharati.

Printed Works: Bengali (Fiction)

Rajarshi (The Royal Sage). 1292/1885-6, *Balak* (journal: in part). 1293/1887. Kolkata: Adi Brahmosamaj Press.

Prajapatir nirbandha (The Marriage-God's Decree). 1314/1908, Kolkata: Majumdar Library. Earlier, titled *Chirakumar Sabha*, 1311/1904, Kolkata: *Rabindra Granthabali*, Hitabadi Press.

Shesher kabita (The Last Poem). 1335/1928-9, *Prabasi* (journal). 1336/1929, Kolkata: Visva-Bharati.

Gora. 1314–1316/1907–1910 *Prabasi* (journal). 1315/1909, Kolkata: Kuntalin Press (in part). 1316/1910, Kolkata: Indian Publishing House.

Printed Works: English (Poetry: Volumes)

The Child. 1931. London: Allen & Unwin.

Fireflies. 1928. New York: Macmillan.

Gitanjali (Song-Offerings). 1912. London: The India Society.

Lekhan (Writing). 1334/1927. Berlin/Balatonfüred (Hungary).

Stray Birds. 1916. New York: Macmillan.

Printed Works: English (Poetry: Individual Poems)

'The Sunset of the Century'. 1917. *Nationalism*. New York: Macmillan.

'All fruitless is the cry'. 1942. *Poems*, ed. Krishna Kripalani et al. Kolkata: Visva-Bharati.

'I know that at the dim end of some day'. 1914, *Modern Review* (journal). 1916, *Fruit Gathering*, London: Macmillan.

'I hid myself to evade you', 1918. *Lover's Gift and Crossing*. London: Macmillan.

Printed Works: English (Drama)

The King of the Dark Chamber: translation of *Raja* by Kshitishchandra Sen, wrongly ascribed to Tagore. 1914. London and New York: Macmillan.
Red Oleanders: translation of *Raktakarabi*. 1924, *Visva-Bharati Quarterly* (journal). 1925, London: Macmillan.

Printed Works: English (Fiction)

‘Giribala’ and ‘Emancipation’. *Broken Ties and Other Stories*. 1925. London: Macmillan.

Printed Works: English (Non-fiction)

‘An Indian Folk Religion’. *Creative Unity*. 1922. New York: Macmillan.
The Religion of Man. 1930. London: Allen and Unwin.
Talks in China. 1924(?). Kolkata: Visva-Bharati Series.