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

Sukanta Chaudhuri
Editor

Bichitra: The Making of an Online Tagore Variorum

Springer

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

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 Amritesh 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

The Story of the Bichitra Project

1

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

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_1

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,