INTRODUCTION

This book is a practical guide for using digital tools and techniques for historical research, but it should be useful for anyone from across the humanities who is interested in working with collections of texts at scale. We do focus on text and so, although our examples will be historical, their applicability to the study of literature, for example, should be evident.

This book is aimed at non-programmers and it does not teach or require the use of any programming language. The practical examples we work through substantiate our belief that you can do a great deal without programming, by leveraging the work of others. Learning to program is interesting and useful but it is not essential to doing digital history and many digital historians choose not to. There are certainly things that can only be done with coding but we have purposely omitted all such things. We hope you will be surprised by the power and flexibility of the approaches we show you.

We have deliberately chosen to focus almost entirely on tools which have been used for decades and which we expect to continue to be used for many more years. They are mature and there is an abundance of advice available online to supplement what we show in the book.

We do insist that much of the work in digital history projects is likely to consist in preparing the data for the interesting part of the project: the part that produces interesting results (whether that be an idea, a graph, a map or a website). This is often known as 'data cleaning'. Actually, that term is slightly misleading because it implies that the problems with the underlying data are always errors of some kind. Sometimes the data needed for a project arrives in a clean and well-organised condition, but simply happens to be in the wrong format for its new use and so requires preparatory work. Data preparation is not much discussed when digital projects are written up, but this book is a practical guide and so we have tried to give it due prominence.

Our book follows the general structure of other volumes in the IHR Research Guides series. Chapter 1, 'The context of digital history', describes the history of our subject and some of its milestones. What is available as a digital source, and how, is a product of the history of the subject and, more importantly, early drivers of digitisation, such as the commercial value of genealogical sources. In writing this chapter we found it hard to disentangle digital history from the digital humanities more broadly considered and we suggest that trying to impose a clear demarcation is unhelpful.

Chapter 2, ‘Formulating your research questions’, will help you to think through your research ideas in the context of digital history. What techniques will you need? What is already available in terms of data and tools? A critical and judicious approach to early decisions here (both in terms of the research project you pursue and any resources you employ) can save you valuable time and energy and we give lots of advice on how to make those decisions. The last section may seem to skip ahead to some thoughts on where and how you might publish your research. We think that it is best to have a rough idea of this from the very beginning and we also suggest that you should not think of publishing solely in terms of final outputs.

Chapter 3, ‘How a digital project begins’, is a nuts-and-bolts discussion of how a research project might go from books on a shelf to digital output. In our experience not many people are confident with the process of digitising material because they have no experience of doing so. This chapter describes that process. We digitised part of a book specifically for *Doing digital history* and have made our files freely available online (see Appendix 1). The book is *The Post Office London Directory for 1879*, and we have tried to use it for our historical questions and for our practical examples wherever we could throughout the book.

Chapters 4 and 5 go into detail on how to work with digitised text automatically and at scale. We show how you can use the *command line*, which gives you access to hundreds of small programs written by other people, and with which you can accomplish an enormous amount without writing a line of code. We make no apology for talking at length about the command line: it is the Swiss army knife of computing, beloved of most programmers. Learning even a little bit about how to use the command line can transform the way you work. Plain text is covered in Chapter 4 and structured text in Chapter 5. We will show that plain text is harder to deal with, although perhaps easier to get hold of, and that structured text is preferable when available, even if at first glance its appearance may be more forbidding. For structured text we concentrate on XML

(Extensible Markup Language), but the approaches we take should transfer reasonably easily to other formats.

Chapter 6, ‘Caring for your digital history project’, covers the practicalities of managing your data and sharing it effectively. Our section on research data management spends a fair amount of time on using the Git tool to manage your data, because we think this is simply the best option available. Further, a great deal of reusable data can now be found in the form of a Git repository, so a basic understanding of what that means and how it works is becoming essential. We also look at documentation and metadata.

Chapter 7, ‘Visualising your data’, gives an overview of visualising historical data with some advice on practical aspects, such as the use of colour. Here we use the Post Office data to create some visualisations of our own, in the form of charts and a map, with detailed information on how we went from dataset to visualisation and why we made the choices that we did.

Chapter 8, ‘What next for digital history?’, is our attempt to predict what new technologies are around the corner for historians and how they might affect your work. This chapter ignores George Eliot’s advice, in *Middlemarch*, that ‘of all forms of mistake, prophecy is the most gratuitous’. We hope that, even if this chapter eventually proves to be laughably wrong in some of the details, the generalities will affect historical practice one way or another. We put the finishing touches to this book while much of the world was in lockdown because of the COVID-19 pandemic. We have not revised our predictions, even though our expectation of ‘gradual evolution and embedding rather than of revolution and disruption’ already looks dated. We think it is too early to say what long-term changes the pandemic will bring to the practice of history.

We have also included three appendices: Appendix 1 describes the data repository we have created for this book – what is in it, how to get a copy of it and how you might want to use it to practise your skills. Appendix 2 is a table of command line tools and how to use them. We give a human-language description of what each command does and some more extended examples of how you can use them for common tasks. We hope this will be a useful ready reference for day-to-day command line work. Appendix 3 is a summary of the syntax of regular expressions. We think getting to grips with regular expressions, or regex, is essential for working digitally with text. We introduce regular expressions bit by bit in Chapter 4, but Appendix 3 provides a convenient summary in one place. Regular expressions are not easy to learn but we encourage you to keep practising and referring back to this appendix any time you need to.

There are many things we have not included in this book. Equally, our readers do not need to master everything we do cover. Digital history has many facets, some more appropriate to a particular field or congenial to a particular researcher than others. If this book encourages some historians to try something new or go further with a digital approach than they had previously then it will have been worth writing.