Evo Edu Outreach (2011) 4:350–353
DOI 10.1007/s12052-011-0333-6

INTERNET COLUMN

A New Digital Edition of Darwin's 1859 *Origin of Species*

Adam M. Goldstein

Published online: 3 May 2011

Abstract The Darwin Manuscripts Project (http://darwin.amnh.org) is publishing a new digital edition of the London 1859 edition of Darwin's *The Origin of Species*, which, alongside a copy built of scanned pages, extends opportunities for Internet readers. The new digital edition is warranted by the absence of good digital copies of the 1859 London edition online: though many will find this hard to believe, it is nonetheless true. The new edition is described in the context of the reader's experience and in the theoretical context of the nature of texts as a kind.

Keywords Digital publishing · Darwin · Origin of species

Introduction

The Darwin Manuscripts Project (http://darwin.amnh.org) is publishing a new digital edition of the London 1859 edition of Darwin's *The Origin of Species*, which, alongside a copy built of scanned pages, extends opportunities for Internet users. The new digital edition is warranted by the absence of good digital copies of the 1859 London edition online: though many will find this hard to believe, it is nonetheless true. This state of affairs is described in more detail in the next section, subsequent to which the nature of the new edition is described in further detail. An interesting consequence of the means used to produce the new edition, digital typesetting software, is that the idea of a text takes on a new shape.

A. M. Goldstein (✉)
Department of Philosophy, Iona College,
715 North Avenue, New Rochelle,
NY 10801, USA
e-mail: z_californianus@shiftingbalance.org

Motivation for the New Edition

There are many copies of *The Origin of Species* online: nonetheless, as I will explain, all fail to meet standards which both general and scholarly readers ought to expect to be satisfied by a digital work suitable for everyday use. The problem is particularly acute in the case of the 1859 London edition. Consider some of the demands that an online text ought to be able to meet. First, it ought to be free. The copyright on the 1859 *Origin* has expired, so the text is no longer owned by anyone with a right to limit its distribution. This criterion is indeed met by at least a few of the editions available at the bookseller site `Amazon.com` for the Kindle electronic book reading device (Amazon.com 2011)—but this of course does not include the cost of the Kindle itself. Paperback copies can be obtained from `Amazon.com` for under US $5.00, a cost justified by the work on the part of the publisher to manufacture and distribute the book.

A second criterion that a good digital copy of the *Origin* ought to meet is that it be authoritative. The reader must be able to trust that the text is absolutely faithful. As well, the reader must be able to determine which of the six editions is

reproduced in a digital copy. One site at which Internet readers can find such an edition is the well-known http://www.darwin-online.org.uk, "The Complete Work of Darwin Online." Here, users can view an image of each page of the *Origin*, and the edition number is clearly indicated. In a two-column view, the image can be seen next to an HTML representation of the text. Page breaks are marked in the HTML text, which is presented as a single, long web page. In addition to being authoritative, this text is suitable for scholars because the original pagination is preserved. Pages referred to by Darwin's contemporaries or others working with a copy of the first edition can be located.

The page numbers in printed books are designed to give readers a sense of where they are in the book, which is also provided by the thickness of the pages remaining in the book. Page numbers help, but in the absence of pages themselves, they do not give the reader as much information about how far he or she has read as in a printed book. Perhaps the solution is to print out part or all of the web page at the Darwin Online site mentioned above. The page numbers of the *Origin* do not correspond to the printed pages, which the web browser will simply fill with text when it formats the page for printing. The same is true for other major divisions in the book, such as chapter divisions. In printed books, these start on a new page, often on the same side of a page, for instance, on the page that is face-up to the right. Font choices, justification and hyphenation, and other design elements appropriate for a web page are not usually appropriate for a printed work. This results in a less-than-optimal reading experience and a lower level of understanding than would be obtained if a well-designed print work were consulted.

In order to create a more familiar reading experience by obtaining a text of the *Origin* in a more book-like form, the reader might elect to download and print out the PDF copy at the Darwin Online site. In this way, the reader can have a copy of the *Origin* much like someone in 1859 would have had: the PDF copy is composed of high-quality photographic reproductions of a copy of the 1859 *Origin*. The PDF copy is different in an important regard from what a reader in 1859 would see: in the intervening time, the pages represented in the PDF copy have yellowed. For the reader of today, this enriches the reading experience because the yellowed pages have a warm, antique look, which does not obscure the print. This yellowing is not a virtue for a digital book, however. To capture this look, the PDF file must be large—92 megabytes. Even over the fastest connection to the Internet, downloading such a large file will take a long time. Once obtained, printing the images will be slow and the quality poor.

In a comprehensive survey, the results of which are reported in the editor's introduction to the digital edition at the Darwin Manuscripts Project, no copy of *The Origin of Species* meets the criteria sketched above. A good digital copy of the *Origin* should be authoritative, clear as to the edition and to the publisher that produced the book represented online, easy to read, and easy to obtain in both digital and print forms. Some of the digital editions of the *Origin* are hybrids of the first and third editions, and many are mislabeled, apparently purporting to be a copy of an 1859 edition, but being in fact either the third or the sixth. The problem of authority and accuracy is most acute. Copying and recopying a digital file can result in the loss of character formatting information, and in many online editions, accented characters are deleted, or else have lost their accents. In producing the Darwin Manuscript Project's digital edition, thousands of errors were discovered in the initial working digital copy, obtained from the Oxford Text Archive. These include the missing characters just mentioned and also include missing and transposed sentences and phrases.

The New Digital Edition

The Darwin Manuscripts Project, as its name suggests, aims to create and publish manuscripts, that is, texts by Darwin not published, such as his reading notes, marginalia, and records of his observations in his scientific notebooks. Correspondence, although not published by Darwin or intended for publication, is the province of the Darwin Correspondence Project; The Complete Work site mentioned above has taken responsibility for publishing digital copies of works of Darwin's available in print, including their translations and multiple editions.

Nonetheless, the Darwin Manuscripts Project does have an interest in providing easy access to good quality digital editions of a small number of Darwin's previously published works. The Project has collected the existing leaves (sheets) of Darwin's near-final copy of the first edition *Origin*; and, those of Darwin's manuscripts especially important to understanding his line of thought leading up to the 1859 *Origin* are collected on the site in the form of a scholarly edition. An ongoing effort at the Darwin Manuscripts Project is to link references to books made by Darwin in his manuscripts to bibliographic records for those books. These records can be linked to the digital full text of the works to which they refer. Site users will expect to find a copy of Darwin's finished work ready to hand.

The Project has obtained a high-quality scan of a first edition of the *Origin*, which is displayed page by page. Like the scanned images at the Complete Work site, these pages have an antique glow, and remain as clear and readable as they did when they were printed in 1859. By the same token, as in the case of the Complete Work site, the scanned edition rates poorly on ease of downloading and printing. The decision was made, in light of this, to publish a second text, a text which conforms more closely to the style of a modern book, and which is made for portability in digital and print forms, and which adds internal navigation not possible with a scanned edition.

Authority ranks high among the criteria for a good digital edition, as already noted. The text started as a plain (ASCII encoded) copy of what purported to be the 1859 text, obtained from the Oxford Text Archive. Annotations to this text file reported that it had been proofread against the facsimile of the first edition, edited by Ernst Mayr (Darwin 1859). In fact, thousands of errors were discovered; in preparing our text, it was proofread three times. Before proofreading could begin, the text was prepared for processing by the LaTeX digital typesetting system. It takes files in ASCII format as input. It outputs PDF files, typeset in accord with markup and other typesetting commands embedded in the text. As will be seen, using this system enables the text to be used for a wide range of applications, in addition to generating the printed text published at the Darwin Manuscripts Project.

The output generated by processing LaTeX renders the text in the format of a printed book. The first proofreader read this text; when errors were discovered, the proofreader made the changes in the underlying text file input to the processor. After re-processing the text to generate a new PDF file reflecting the first proofreader's changes, it was passed along to the second proofreader; after repeating this process, the second proofreader passed the text to the third proofreader. The text was checked against the edition by Mayr mentioned above; Mayr states in his editor's introduction that the facsimile was created by the first edition of the *Origin* held by Princeton University in the Firestone Library.

The general editorial principle that guided the proofreaders is that the text of the digital edition deviate from the facsimile if, and only if, that deviation would improve access to Darwin's ideas for the reader of today, but would not influence the meaning of the text. This means that every word, sentence, paragraph, and section heading or other similar devices be copied exactly as it appears in the facsimile. In contrast, no effort was made, for instance, to duplicate the page size or hyphenation of the 1859 *Origin*; and some 19th century conventions of word spacing and punctuation were changed because they are no longer used today and would be a distraction. All content is preserved but the typesetting is not. Every deviation from the 1859 text, no matter how small, is described in the editor's introduction.

A consequence of this editorial principle is that the page breaks in the new digital edition do not match those in the 1859 publication. To remedy this, page breaks are noted in the text. There are two indices, one that refers to page numbers as they appear in the digital republication, and one that refers to the page numbers in the 1859 edition. The new edition is appropriate for use by students, professors, historians, or any other researcher who needs to be able to locate text in the original publication. Entries in both indices are hyperlinked to the pages in the text they refer to. The table of contents is also hyperlinked.

The resulting digital edition of the *Origin* complements the scanned copy. Users can download the entire PDF at once, or download a single chapter. Browsing the scanned copy of the *Origin*, users of the site are given the option of downloading the PDF copy. The scanned pages, each approximately 400 kilobytes in size, add up to approximately 185 megabytes, for the entire *Origin*. The new PDF copy is a featherweight at 6.4 megabytes.

Further Applications

The digital copy of the *Origin* in the PDF format is only part of the new edition or perhaps, one of the many products or results of a larger digital entity that better deserves the title "digital text." To understand this paradox, consider the manner in which it is created. As I mentioned above, the PDF document is generated by processing text files with the LaTeX typesetting system. This is analogous to the manner in which a web browser such as Firefox or Apple's Safari displays the text and images on a web page. The browser queries the address requested by the user. A text file at that address is read by the browser. This file contains computer code instructing the browser what to display. The browser can be viewed as a computer application whose task is to interpret a text file with instructions about what to display on the user's screen. The LaTeX system works in precisely the same way, except that code interpreted by the processing application consists of the text of the document to be created, with commands about how to display text, such as chapter headings, paragraph breaks, page dimensions, and type size and style. A

word processor such as Openoffice.org's Writer application does much the same thing but produces the output as the user enters the text input.

The text files containing the *Origin* text and encoded for processing by are available by way of the Darwin Manuscripts Project web site. They are licensed for redistribution and use by anyone, so long as neither they nor their products are sold for profit, and so long as any changes that are made to the code are also made available to others on the same terms. This means that anyone who can code can create his or her own digital edition of the *Origin*, to whatever specifications desired. The PDF output published on the Darwin Manuscripts site is clearly a digital edition of the *Origin*, which is formatted for print output as a book. Nevertheless, intuitions about what makes a text must adapt to keep pace with the times. The source files, together with the interpreter and a display such as a computer screen or hard copy, are a digital, virtual entity that can in principle, an in all likelihood actually will, produce different kinds of entities that are each, in and of themselves, texts of the *Origin*. The source code of the digital *Origin* is best considered the digital text, its machine-interpreted products as extensions of the more basic digital entity.

Many alternative forms of output come immediately to mind. A copy reformatted for display and ease of use on the many portable digital reading devices such as Apple's iPad or Amazon.com's Kindle would be of interest to many people. A large print edition for the seeing impaired also suggests itself. Other applications include custom editions for use in the classroom by inserting questions, comments, or underlining or numbering passages. Excerpted texts for classroom use might also be created. These might include hyperlinks to external information sources. The source text can be "translated" into another form of digital markup, such as HTML for display on the web, or custom XML markup designed for a particular application. Indeed, the Darwin Manuscripts Project has used the source code for this very purpose, creating a database of the *Origin*'s contents, to be read by the site's search engine. Finally, perhaps most intriguing, the source code might be read directly by a machine indexer to identify patterns, search for particular phrases, or generate input for a natural language processor.

Concluding Remarks

Two new digital copies of the 1859 London publication of the *The Origin of Species* are available to the public at the Darwin Manuscripts site. Together, these texts present Internet users with a unique opportunity for a particularly rewarding experience. The copy built from scanned pages contextualizes Darwin's work, evoking the late nineteenth century. The text created by a modern digital typesetting application makes up for what the scanned copy lacks, utility for the working scholar, student, or interested member of the general public.

References

Amazon com. Amazon.com. 2011. http://www.amazon.com. Accessed 18 March 2011. Search for "Origin of Species" on the site.

Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. In: Mayr E, editor. Facsimile reprint. London: John Murray; 1859. (Harvard University Press, Cambridge; 1964.)