Evo Edu Outreach (2009) 2:122–135
DOI 10.1007/s12052-008-0113-0

DARWIN RESOURCES ON THE WEB

Charles Darwin's Manuscripts and Publications on the World Wide Web

Adam M. Goldstein

Published online: 20 January 2009

Abstract This paper describes three Internet resources that publish manuscripts and published works by Charles Darwin. The authority, content, and design of each is outlined, and each site is assessed according to how effectively its design and organization of the material promotes exploration of Darwin's works and ideas. The presentations of the materials are compared with the traditional presentation of text materials in books. It is concluded that, while these three sites offer a large amount of important work by Darwin, access to it is sharply limited because, particularly in the case of two of the sites, the material is not presented in a structured manner that might direct users to particular ideas or texts, especially if those ideas or texts are not well-known to the user.

Keywords Charles Darwin · World Wide Web · Darwin manuscripts · Digital publishing · Origin of species (work by Darwin) · Book (form of information delivery)

Introduction

The aim of this essay is to orient Internet users to three sites providing access to the work of Charles Darwin. "The Darwin Correspondence Project" (http://www.darwinproject.ac.uk/) offers letters to and from Darwin; "The Darwin Digital Library of Evolution" (http://darwinlibrary.amnh.org) offers a small but growing and well-organized selection of Darwin manuscripts, edited to a high standard of scholarly excellence, and an extensive bibliography of evolution particularly notable for its listing of works used by Darwin himself; and "The Complete Work of Charles Darwin Online" (http://darwin-online.org.uk/) is an excellent source for Darwin's published works. These sites are accessible free of charge, and much of their content is available for download. Before discussing the sites, I offer a few words on the strengths and weaknesses of Internet publishing.

Internet Publishing: An Unfulfilled Promise

The Internet's use as a means for the distribution of written texts and other products of human intellectual and imaginative work is an advent equal to the development of the printing press with movable type. It is not far-fetched to imagine that at some point in the not-so-distant future, much of the content of the world's libraries will be readily accessible online by anyone in the world. Nonetheless, there are important respects in which it is not clear whether distributing a text online really does provide more access to its content, in contrast with its print form. I sketch two of these respects in which Internet publishers are still struggling to equal the achievements of Gutenberg, his predecessors, and his inheritors responsible for what has been an enormously durable, remarkably stable publication medium, the humble and familiar book.

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

Moving Beyond the Book?

Internet publishing possesses several advantages over publishing in print. Works published online reach a broader audience, print works being limited by the range of their physical distribution. This is particularly important for archival materials such as letters and manuscripts, which exist in single copies. Rather than visit an archive in, say, Cambridge, England, readers all over the world can view Darwin manuscripts online. Some archival materials, such as Darwin's correspondence and notebooks, have been transcribed and published in print form. Nonetheless, most readers will have a difficult time obtaining them. These expensive works of many large volumes are not likely to be held in neighborhood public libraries or college libraries; the best source for them is the library of a large university with an extensive research collection, or the library of a large natural history museum. A second advantage the Internet provides over print is that online documents can be hyperlinked. A scholarly editor can direct readers to related works, notes, or other editorial devices intended to clarify the texts by linking to related resources directly. Third, digital documents can present different views of a document, for instance, displaying a manuscript with and without annotations, as though they were overlays; presenting it at different sizes; or rotating it.

These advantages are partially offset—in some cases, completely offset—by the absence of an online format that matches the intuitive ease of use and familiarity of the printed book. Jumping 200 pages ahead in a book is a simple matter; online, infelicities of web site design may obscure mechanisms of "page" turning, leaving the user puzzled about how to move forward or backward more than one page at a time; or, if the mechanism's usage is clear enough, it may be so clumsy that users have a difficult time viewing the page in which they are interested. Likewise, electronic "bookmarks" lack the simplicity of the real thing: How to mark one's place in a digital text? Another serious problem is that the digital medium degrades a text's readability. Most computer screens cannot display a full page at full size, destroying information about one's location on a page or within a paragraph. Readability is further compromised by the abandonment of ages-old principles of book design such as those concerning the size and placement of the type block, page size, placement of page numbers (folios), and type design. Many of these elements of book design cannot be reproduced online, even if digital publishers wanted to do so; in their absence, alternative technologies for readability have yet to be fully developed.

In my account of each site below, I will touch on the user experience, alerting site visitors to how well each site's designers have met the challenges of distributing digital copies of books, letters, and manuscripts. I will focus on areas of greatest concern to the user: navigation to the various parts of the site, readability of the texts and navigation through them, and the ease of printing and downloading texts. The idea is not to assess the sites from an aesthetic point of view, but in terms of those design elements that might help or hinder a reader's understanding of a digital text. I aim to sketch each site's design principles from a broad point of view, so that someone might be able to use my account of each site as a preliminary "road map" on his or her initial visits.

Wide Distribution—but No Access?

Internet publishing neatly solves the problem of distributing works to many people, simultaneously, over great distances. But does this mean that there is more access to the information in each work? At the very least, it is clear that after a work is distributed online there is more access to it than before it was distributed online, in the following sense: Before it appeared online, many people had no access to it at all, whereas after it appears online, many people have some access. Can a more precise statement be made concerning *how much* access to a work is provided by distributing it online? There is a strong case to be made that, particularly when a large body of work is distributed online, the increase in access is minimal. The problem is that it can be difficult for a user to find suitable entry points to the online work. Identifying information of interest poses a significant problem, which gets worse as the volume of texts in a given collection grows.

First, note that unless a researcher has narrowed his or her search down to one or a few works, browsing is futile. Scanning a single work such as the *Origin* for passages relevant to one's questions or research problems takes a significant amount of time, though it is profitable do so, as many researchers have done many times. The problem is that searching, say, *all* of Darwin's published works by browsing would be an impossible task, given the amount of time available to most researchers before their projects must be complete.

Abandoning the strategy of searching a large collection online by browsing, it is natural to consider free-text searching in the full text of the works in the collection, which seems to hold great promise. Search software scans the entire contents of the texts in the collection; a "hit" occurs if a pattern matching the

morphology[1] of the user's search key is identified. Results of searching are presented as a list of passages, page numbers, or other information about the location of hits in the collection, with a link to each one. There are two difficulties with this strategy. First, there is a wide scope for mistaken hits; second, there are many relevant passages in the collection that cannot be identified by morphological pattern-matching. A single morphological pattern of letters can have multiple meanings; free-text searching will identify all such patterns, regardless of whether they are of interest to the user. These "false positives" are mistaken hits. In contrast, many concepts or objects are often known under a range of descriptions that have no morphological similarities. For instance, someone interested in natural selection would want to be alerted to passages that use the phrase "survival of the fittest;" but no pattern-matching software looking for "natural selection" will identify such phrases. This is a case in which relevant passages are not identified as hits. Searching a very large collection such as Darwin's correspondence or all of his published works, these problems grow significantly, the number of hits returned by a search increasing enormously, and the potential for missed opportunities becoming acute.

Expert indexing of works offers an alternative. Works in a digital collection can be identified by subject by selecting terms from a list that gathers together synonymous terms and phrases, for instance, cross-referencing "survival of the fittest" to "natural selection." To apply index terms, a scholar or trained indexer determines the subjects addressed by the work and identifies other variant types of search keys that users might search for the text, indexing it according to these as well. In a digital text, such indexing can occur at a fine-grained level. A journal article or book chapter can be marked up so that it is visible to the search software. Expert indexing is time-consuming and costly, relying on individuals with rare skills. The possibility of training machines to mimic the indexing behavior of human indexers is exciting. Work toward improving the capability of machine indexers is currently under way by artificial intelligence researchers studying natural language processing.

In my account of the three Darwin sites below, I will consider the extent to which the organization of the site contributes to solving the problem of access. Are links to works organized to provide natural entry points to the collection, given its format, the subjects it covers, its authors, or some other intellectually useful means? Once a work of interest is identified by a user, is any means provided for identifying similar or related works? As will be apparent from my comments in conclusion to my discussion of each web site below and my conclusion to the paper, I conclude that, even though significant progress has been made toward solving problems of distribution, there is a great distance yet to be traveled toward improving *access* to Darwin's works.

[1]By "morphology," I just mean the *shape* of letters and punctuation, arranged to form words and phrases. Search software can match such patterns in a digital text to a user's search key without having any information about their context or meaning.

The Darwin Correspondence Project

> I have read heaps of agricultural & horticultural books, & have never ceased collecting facts—At last gleams of light have come, & I am almost convinced (quite contrary to opinion I started with) that species are not (it is like confessing a murder) immutable.
>
> Letter 729—Darwin, C. R. to Hooker, J. D., [11 Jan 1844]

In correspondence, a scientist struggles to explain his or her ideas as they develop; assesses them; poses questions to other scientists, and responds to those asked by others; and reveals his or her personality. The public is fortunate that Darwin wrote many letters and that many have been preserved. The Darwin Correspondence Project aims to provide online access to transcriptions of all letters by Darwin and all those he received. Darwin's correspondence is particularly important for scholars. Working from his home in Down, Darwin acquired information from all over the world by exchanging letters with naturalists, collectors, physicians, and others in remote locations that provided him with first-hand accounts of plants and animals to which he had no direct access, vastly extending the number and diversity of the observations around which he built his theory of evolution.

Personnel

The Project began in 1974 under the direction of Frederick Burkhardt and Sydney Smith, and it is now directed by James Secord, a professor in Cambridge University's Department of History and Philosophy of Science; Prof. Secord is assisted in his work by a full-time staff of eight, which includes three Ph.D.- and four Masters-level scholars. As well, there is a full time staff

member in the USA, at Cornell University, and three volunteer US staff.[2]

The Collection

The collection of letters online at the Darwin Correspondence Project's web site is a companion to the ongoing publication of the letters in print, estimated by the Correspondence Project to take until 2025 and to require 30 volumes. Letters appear on the web site four years after they appear in a print volume. At present, there are approximately 14,700 documents in the database, including 7,600 letters from Darwin and 6,500 to Darwin, in addition to "relevant third party letters" and "memoranda." As in the print volumes, the web site presents transcriptions of the letters, the original copies of which are held in archives of Darwin correspondence all over the world, primarily at Cambridge University Library in England and the American Philosophical Society library in Philadelphia.[3]

The Correspondence Project adheres to high standards: A letter's transcription is proofread by four different individuals before it is published.[4] The verbatim text of each letter is published, including misspellings, punctuation such as dashes, and abbreviations such as "sh^{d}" for "should." The date, signature line, "My dear Sir," and the like appear as Darwin wrote them as well. The text of each transcription appears in a large, readable sans-serif font, with no annotations or interpolations in the letter text; these appear in hyperlinked footnotes. Other footnotes provide biographical information about a person mentioned in the letter, or describe an event mentioned in it; still others direct readers to recent scholarship about persons, events, places, or biological groups. Darwin's ideas or words appearing in his published writings or other letters are also pointed out in the scholarly footnotes. At the top of each letter's page, the editors summarize it briefly. A sidebar to the left on each letter's page provides catalog data, such as who the letter is to and who it is from, where it was sent from, a physical description of the letter, information about where the original is held, and cross-references to the *Calendar of the Correspondence of Charles Darwin* (Darwin 1994) and to the print *Darwin Correspondence Project*. Subject headings direct readers to related letters. This is an important tool for users wishing to explore the collection.

Users can search the letters using an "advanced search" tool. Help pages describe the search query language, including how to conduct boolean searches, searching within a date range, searching for a letter to or from a particular person, letters about particular places, locations from which letters were sent, or subjects of the letter. This is only a partial list of the fields describing each letter that a user can search. Because the collection is so large, narrowing a search using some one or other means of advanced search is almost certain to be necessary in almost every instance.

The site does not offer any special format for printing or downloading letters. Users should find it easy to acquire letters for themselves, however, because most are short, and can be printed directly from a web browser or saved as a web page or text file directly to a hard drive.

Eclectic Pathways into the Collection

At the start of this paper, I raised the issue of access. There is no question that the Correspondence Project makes enormous strides toward distributing important texts to a wide audience. But how does a user find an entry point into the collection? The Correspondence Project web site offers some eclectic topic-oriented pathways into the collection. The topics are religion and science and ecology. The religion section is broken out into further subtopics, for instance, design in nature, belief, and ethics and society. Pages for each subtopic offer interpretative overviews of the nineteenth century context, highlighting key individuals, places, and events relevant to the subtopic at hand. Most importantly, each subtopic section of the site links to letters of particular relevance. At present, the section on science only contains one major subsection on ecology, which connects Darwin's ideas about the subject with those of today by way of short articles accompanied by photographs.

These are eclectic pathways because they are not of central interest to someone researching Darwin. Religion and science is an important topic, but many wanting to learn about Darwin would no doubt want to know about the development of his ideas about evolution itself or natural selection. Other topics of central interest to researchers include Darwin's correspondence concerning the acquisition of specimens from locations around the world or his discussion of his experimental designs and reading in natural history.

[2]This information about project staff can be found at http://www.darwinproject.ac.uk/content/view/112/110/.

[3]The information reported in this paragraph can be found at http://www.darwinproject.ac.uk/content/category/3/30/36/, a page of "FAQs;" consult those concerning the correspondence.

[4]See http://www.darwinproject.ac.uk/content/view/17/89/ and http://www.darwinproject.ac.uk/content/view/74/42/ for more on the editorial policies of the Correspondence Project.

The home page also links to a section of the site named "Teaching Materials," which links to course materials used at Cambridge University in an upper-level undergraduate course in history and philosophy of science. Links to letters are organized into topical groups. The main subdivisions are "Scientific Networks," "Scientific Practice," "Controversy," "Religion," and "Beauty." Within each main subdivision, there are further subtopics. For instance, "Scientific Networks" contains "Friendship and Trust" and "Friendship and Information Exchange." There are 21 subtopics in all, each citing a dozen or so letters. A bibliography on topics including the nineteenth century practice of letter writing is also found on this part of the site. Users wanting an initial orientation to the collection are urged to visit this part of the site, which offers a topic-oriented gateway created by Darwin scholars and intended for those that are not already experts.

Concluding Remarks

The Darwin Correspondence Project (http://www.darwinproject.ac.uk/) offers transcriptions meeting the highest standards for accuracy and it presents a large number of transcriptions. Darwin scholars can look forward to the day when all of Darwin's correspondence is available online. This increased distribution increases access only incrementally, however. Subject classifications supply users with a valuable tool for exploring the collection, but researchers must first identify a letter on a topic of interest in order to enter into the network of subject keywords. The correspondence is not organized by any theme or general scheme of classification; this is reflected in the web site's design, which has few navigational elements. Most users will probably enter the collection by way of the search tool.

The Darwin Digital Library of Evolution

The Darwin Digital Library of Evolution (DDLE), a project of the Research Library at the American Museum of Natural History (AMNH), was launched in conjunction with the AMNH's 2005–2006 Darwin Exhibition. The DDLE has the ambitious aim of providing a unified access point from which Internet users can explore the literature of evolutionary science, including its initial elaboration by Darwin in his manuscripts and published works, its subsequent development in books and articles by evolutionists, and responses to it in culture and society. The DDLE site will store and distribute Darwin manuscripts but not other Darwin texts; its role as a gateway or portal to information on evolution served by linking to other resources all around the Internet.

In order to attain the ambitious project described above, the DDLE incorporates two component projects:

1. The *The Darwin Manuscripts Project* aims to provide the public with access to Darwin's manuscripts, organized by theme, and around Darwin's authorship of particular works, such as *The Origin of Species* and *The Descent of Man*. The Project aims to edit manuscripts to the highest professional standards of excellence. This component of the Darwin Digital Library can be reached by following the "Library" link on the site's home page.
2. *The Literature of Evolution* aims to provide the public with access to the literature of evolutionary biology from the present day to Darwin's precursors dating from the seventeenth century. Links to full-text digital copies of works listed in the bibliography are provided, if such copies are available; more are becoming so as the Biodiversity Heritage Library, a large-scale digitization project, proceeds. This component of the Darwin Digital Library can be reached by following the "Bibliography" link on the site's home page.

Although the site has advanced only slightly beyond its "opening day" content and design, editorial staff are, at present, preparing major updates to both content and design in anticipation of presenting them online in 2009 to mark the anniversary of *The Origin of Species* and Darwin's 200th birthday. Here, I describe background and content of each of its two component projects as they appear on the site at present (November 2008).

Personnel

The DDLE's Editor-in-Chief is Dr. David Kohn, Emeritus Professor of History at Drew University; Dr. Kohn's experience includes contributions to print publications of Darwin's notebooks (Darwin 1987) and to the Darwin Correspondence Project's print publications (Darwin 1985–1987, 1994). The project's scientific advisor is Dr. Niles Eldredge, Curator of Invertebrate Paleontology at the AMNH, whose recent publications include *Darwin: Discovering the Tree of Life* (Eldredge 2005) and *Reinventing Darwin* (Eldredge 1995). Dr. Adam M. Goldstein, a philosopher of science specializing in the study of scientific method, explanation, and chance in the context of evolutionary biology and

a librarian focused on cataloging and indexing, is the site's Associate Editor.[5]

The Darwin Manuscripts Project

Here, I describe Darwin Manuscripts Project editorial practices and the materials that are available online at present. An underlying theme of my account of the Project is that its editors aim to provide access to Darwin's works by selecting materials that reflect a trend in his thought or his progress toward the creation of a given publication, such as the *Origin*.

Project Editorial Practices

The aim of the Darwin Manuscripts Project is to create and distribute digital copies of Darwin's manuscripts suitable for use by historians, digital copies that preserve all information appearing in the originals. This represents a significant challenge, because Darwin's manuscripts possess a three-dimensional *layering* structure difficult to reproduce in two dimensions. To understand what this three-dimensional layering structure is, why it is important, and why it poses a challenge for someone wishing to publish Darwin's manuscripts, consider how Darwin produced them and what value they have for the historian.

The distinguishing feature of a manuscript, in general, is that it is not intended by its author for publication. Darwin's manuscripts touch on topics that one would expect a working scientist of the late nineteenth century to encounter: field observations; experimental setups and results; reading notes, including those in the margins of books ("marginalia"); draft writings; and notes about connections between current and past work. Starting with a fresh manuscript page, Darwin put down, most often in blue or black ink, what may be termed the "base layer" of the manuscript. As would be expected of anyone developing his or her ideas, Darwin often returned to his manuscripts, putting down additional layers of writing on top of the base layer. These additional layers take many forms: crossing out, adding new text between lines of existing text ("interlining") or in the margins, inserting cross-references to the work of others or other manuscripts of his own, or commenting on the ideas he expressed in the base layer. Darwin also used the "upper layers" to *index* manuscripts. By marking them with numbers or letters, Darwin would identify a manuscript or set of manuscripts that belong together because they address a given topic. He also grouped manuscripts that he intended to use as the basis for a given chapter of a book in progress. The layering of the manuscripts preserves the history of Darwin's thinking: if one could determine when each layer was added, one could reconstruct his thought. Manuscripts indexed by the book chapter for which they were intended in a published work offer a rich source of information. It is possible to trace a phrase or idea from its first appearance in the base layer to the form in which it appears in *The Origin of Species* or *The Descent of Man*.

Why is it difficult to produce digital copies of Darwin's manuscripts? A Darwin scholar in the Cambridge University Library able to inspect a manuscript directly can easily identify its layers with the unaided eye. As mentioned above, Darwin's base layer is typically put down in ink. Other layers can be distinguished from the base layer because they are created with a variety of other writing instruments, such as red or brown crayon or pencil. With the original manuscript immediately before one's eyes, no special viewing apparatus is required to see that Darwin had written a remark in brown crayon on top of the base layer in black ink. The crayon mark passes over the ink mark; one can confidently deduce that the brown crayon appeared later. Loss of the layering information—"flattening" the manuscript—would destroy the historical record of Darwin's thought, causing an unsuspecting reader to make the mistake of believing that all writing on the page appeared at the same time. Someone aware of the layering would be helpless to identify it, and would most probably want to reserve judgment about the historical significance of one or another manuscript passage. The problem is that all means of duplicating Darwin's manuscripts "flatten" them: photography, scanning, and microfilming obscure the differences between the marks left by the various writing instruments, and also make it impossible to tell which marks pass over which others. Another drawback of photography, scanning, and microfilming is that they destroy information about paper characteristics, which can aid in dating a manuscript.

The DDLE adopts the solution to this problem used by print publications of Darwin's manuscripts, for instance, the 1836–1844 notebooks (Darwin 1987), a work coedited by DDLE Editor-in-Chief David Kohn. Manuscripts are transcribed and presented on screen in a text format of the kind usually found in printed books, and the transcription is annotated to indicate

[5]This information can be verified at http://darwinlibrary.amnh.org/index.php?globalnav=people. Christine Stephenson is also listed as a project staff member; this is incorrect. She has left the project.

interlining, insertions, deletions, and other layering effects created by Darwin's revisions. The annotations also indicate which writing instrument was used and describe the paper on which the manuscript is written. Annotations are written directly into the transcription; readers can refer to a readily available key to learn typographical conventions for representing the various types of annotations. For instance, Darwin's deletions are represented in angle braces. Manuscripts are presented in several views: Scanned images of the original, a transcription without annotations, and a fully annotated transcription. This range of formats offers the reader the opportunity to confirm that the transcription is correct, at least as far as the "flattened" manuscript image allows comparison with the three-dimensional layering evident in the transcription. The transcription also provides a further benefit. It provides for the identification of words, phrases, and punctuation. For someone not trained in reading Darwin's handwriting, these minimal units of meaning can be difficult to discern.

The collection at present

The manuscripts now online have been selected and organized to reflect the development of Darwin's thought in the years leading up to the publication of *The Origin of Species*. This is in keeping with the DDLE's larger aim of offering Internet users natural pathways into the literature of evolution. The collection takes Darwin's 1836–1844 notebooks as its starting point, distributing a volume of manuscript transcriptions previously only available in print (Darwin 1987). This volume, coedited by Dr. Kohn, is annotated to show the manuscripts' layers, and scholarly footnotes provide context and further connections. Next in sequence, still in preparation for online distribution and indicated by a link not yet activated, are the "sketches" of 1842 and 1844. The volume intended for online distribution is *The Foundations of the Origin of Species* (Darwin 1909), edited by Darwin's son Francis and published in 1909. This work, one of the earliest in which Darwin records his developing thought on evolution and natural selection in essay form, exhibits both intriguing differences and similarities, when compared with the *Origin* in organization and strategy.

After the *Foundations*, the site's central contribution to the online publication of Darwin's manuscripts is presented, The "Natural Selection Portfolios." Darwin organized some of his notes in portfolios—a kind of document box—grouping notes on a given topic together, or, in some cases, grouping notes together because they form the basis for a given chapter of the *Origin*. A selection of notes from portfolios organized by Darwin starting in 1854 is presented on the DDLE site. The manuscripts have been selected to highlight Darwin's thinking on divergence, described in Dr. Kohn's notes to the manuscripts as "Darwin's attempt to explain how, through natural selection, one species living in a single range would split into separate descendant species."[6] This principle plays a central role in his views about the evolution of adaptation and the origin of species.

The scholarly footnotes and annotations of portfolio material appearing on the site have been prepared specifically for online distribution at the DDLE site. The manuscript transcriptions have been edited to a high standard, proofread by several people trained to read Darwin's handwriting and trained in marking up digital copies of transcriptions for presentation online. The transcriptions take advantage of the digital medium by presenting documents in three views. First, transcriptions may be viewed fully annotated, showing all layering information; second, transcriptions may be viewed with all annotations hidden, which makes them easier to read, though less informative; third, a scanned image of the original manuscript may be viewed alongside the transcription; the scan appearing in a separate window.

The final work on the site in the sequence of works leading up to the *Origin* is *Charles Darwin's Natural Selection* (Darwin 1975). This work does not appear online anywhere else, the print copy being reproduced at the DDLE site. This volume, edited by R. C. Stauffer, is well known to Darwin scholars as the work he abandoned in order to write the *Origin*, which he viewed as a compact account of his work in *Charles Darwin's Natural Selection*. The *Origin* itself appears, completing the sequence; the DDLE offers a digital copy "borrowed" from the Oxford Text Archive.

Navigation through the portion of the site containing the manuscripts is crude but effective. The site is simple due to the small number of works presented on it. Using the sidebar navigation links, users can follow "Library" back to the "Publications and Manuscripts" page, which has the *Origin* and associated manuscripts. The sidebar appears on all pages presenting manuscripts and those presenting the *Origin.* In fact, the sidebar appears on almost every page on the site, making it difficult for a user to lose his or her way.

Printing either the *Notebooks* or *Natural Selection* results in output that is worthy of publication in a

[6] See http://darwinlibrary.amnh.org/index.php?globalnav=manuscripts&sectionnav=viewer&unit_id=745#text_h10253.

distinguished print volume. Multiple pages can be printed at once; the site uses the user's operating system print interface by way of the user's web browser, which, in most cases, will allow the user the option of printing a range of pages. The Natural Selection Portfolios and the *Origin* are more difficult to print, apparently designed primarily for online viewing. Each of these works is presented online one page at a time, each page being presented on a single web page. The user's only option is to print the page while viewing it. The output is nicely presented; most of the manuscript pages are short, fitting in the top portion of a printed page. Scholarly footnotes may continue on for several pages, however.

The site search facility is useful. Terms entered into a search box in the upper-right-hand corner of almost every page of the site will generate a list of page numbers in each work in which the search key, or some derivative of the search key, appears. The search is comprehensive, searching the full text of works in the database. Users must locate hits on the pages presented; they are not highlighted or otherwise indicated. This can be confusing in some cases. For instance, searching for "variation under domestication" finds some pages with this complete phrase, but it also finds pages in which these words to not appear together. A user must scan the text of the page to determine whether it represents a "direct hit" in which the search key is matched exactly or nearly exactly.[7]

The Literature of Evolution

The historical scope of the Literature of Evolution project extends beyond Darwin, aiming to create a bibliographic database for references to works in the tradition of evolutionary thinking started by Darwin, including those of the present day. The project also has the important aim of organizing references by subject using an entirely new subject classification designed specifically for organizing works about evolution. I provide some general background to the project's aims, including discussion of the novel classification scheme it will employ. Then I describe the materials online at present by the project at the DDLE site.

Background

The aim of the Literature of Evolution project is to create a bibliographic database containing records for all works on evolution, beginning with its antecedents in the seventeenth century and extending to the present day, and to provide links to works that are available online. If a work is free to the public, the bibliographic record will link to it directly; if the work requires registration or payment, the bibliographic record will link to a page from which the user can decide whether to register or pay for the work. Users at an institution such as a university or library that subscribes to the journal or collection in which the desired work is held may be able to "click through" directly. Organizing the database of references by subject is also a central goal of the project.

The aim of assembling a comprehensive database of bibliographic information for works produced during a 300-year-long period is indeed ambitious. There is already a substantial number of references in the database—approximately 3,500—which is nonetheless a small fraction of the total. In order to build the database, a variety of methods for importing records *en masse* will be used: extraction of records from library catalogs, such as the AMNH's; extraction from literature indexes such as PubMed; and optical character recognition software together with custom text-processing scripts, for importing the bibliographies of print works.

Locating and linking to digital copies of works listed in the database also presents challenges. Copyright ownership is one such challenge. Works printed before the early 1920s are free of copyright restrictions. Those that have been digitized and presented to the public can be linked to directly. Some copyright owners may offer later works online for free, which can be linked to directly as well. Journal literature provides a particular challenge. A subscription is typically required to access commercially published articles.

Of course, even if copyright barriers to distributing a work online can be surmounted, that work must be digitized and stored in a repository accessible by way of the Internet. Older works pose a problem because they may be rare or delicate. A recent work can be scanned by machine, but a rare or delicate work cannot be, because such works may be damaged by the scanner. They require scanning by hand, carefully opening the book, turning its pages, and imaging each page, one by one.

Fortunately, the DDLE has partnered with an important digitization project, the Biodiversity Heritage Library (BHL), whose home page may be found at

[7] At the time of this writing, the search tool generates an error when the user key contains more than one term in quotation marks ("natural selection," for example). AMNH IT support is looking into this problem.

http://www.biodiversitylibrary.org/. The BHL is a joint project of ten of the world's most important libraries of natural history.[8] These libraries have agreed to digitize their *entire collections*, to which they have pledged to provide online access, free of charge. The BHL has already made significant progress. As of late November 2008, 24,131 volumes, making up 8,664 titles, for a total of 10,062,186 pages, have been scanned. This is a fraction of the estimated two million volumes to be scanned by the end of the project's digitization phase.[9] Some of these include entire runs of several journals. Special provision will be made for scanning older, delicate works, such as first editions of the *Origin*.

As the discussion of the partnership with the BHL suggests, the Literature of Evolution is not a digitization project. The Literature of Evolution's aim is to inform Internet users about which texts are relevant to the study of evolution and, wherever possible, to direct users to digital copies of those works. It is a portal or gateway. A fraction of the two million texts to be digitized by the BHL concern evolution. The aim of the Literature of Evolution project is to identify and link to just those works, separating them from other works on natural history that do not concern evolution. In this sense—because it is a portal, not an archive—the Literature of evolution differs from the Darwin Manuscripts Project, the Correspondence Project, and the Complete Work of Darwin Online (see below).

Especially as the size of the Literature of Evolution database grows, the central problem to be overcome will be organizing it in some way. Subject organization is particularly important. Users must be able to navigate the database—a gateway to the vast collection of works on evolution available on the Internet—by more than just title or author. This is adequate for known item searches, that is, searches conducted by a user who needs to locate a work whose contents he or she believes will be useful, the only issue being the location of the work online. A database of bibliographic records must also be able to direct users to works that they do not already know about, but that may be relevant to their interests.

The solution to the problem of organizing such a large literature adopted by the Literature of Evolution is to build a subject index for works on evolutionary biology. No existing subject index is adequate for describing a body of literature about evolution; existing indexes such as the Library of Congress Subject Headings or the National Library of Medicine's Medical Subject Headings do not provide a description of evolutionary biology fine-grained enough to organize its literature. The subject index being constructed for the Literature of Evolution is termed "The Evolution Ontology" (EO) and, as its name suggests, will take the form of what is known as an "ontology." An ontology is a description of a discipline or field of knowledge or practice. EO contains terms describing the theory and practice of evolutionary biology. An ontology is particularly well suited for promoting exploration of a database of bibliographic records because ontologies represent relationships between concepts particularly clearly. Records can be indexed by the topics addressed in the works they represent; the ontology allows researchers to obtain a clear view of related subjects and to navigate to works on those subjects. EO is presently in its initial stages of development.[10]

The collection at present

Though still in its infancy, the Literature of Evolution offers a unique reference list which both the Darwin scholar and someone interested in evolutionary biology more broadly will find particularly useful. The list of bibliographic references is accessible by way of the "Bibliography" link on the DDLE site's left-side navigation panel; this brings the user to the Literature of Evolution's main page. The reference list, approximately 3,500 references long, can be viewed online from this page in alphabetical or chronological order; EO keywords have not yet been applied to the works. The references are presented on two long web pages, one for chronological and one for alphabetical order. Users can navigate these pages by decade or by letter of the alphabet. These pages do not fit into the overall DDLE design, lacking the sidebar navigation of the rest of the site, but a liberal supply of links back to the DDLE home and main bibliography page make it easy for users to return to central locations. Lacking subject

[8]The ten member institutions are as follows: the AMNH (New York, NY); the Field Museum (Chicago, IL); Harvard University Botany Libraries (Cambridge, MA); Harvard University, Ernst Mayr Library of the Museum of Comparative Zoology (Cambridge, MA); the Marine Biological Laboratory—Woods Hole Oceanographic Institution (Woods Hole, MA); the Missouri Botanical Garden (St. Louis, MO); the Natural History Museum (London, UK); the New York Botanical Garden (New York, NY); the Royal Botanic Gardens, Kew (Richmond, UK); and the Smithsonian Institution Libraries (Washington, DC). See http://www.biodiversitylibrary.org/About.aspx.

[9]The BHL's home page always displays a count of the volume of material scanned in its upper-left-hand corner; for information on the total number of volumes to be scanned, see http://www.biodiversitylibrary.org/About.aspx.

[10]For more information on EO, contact the author of this paper, who is the lead researcher on the project.

organization, the bibliography pages are nonetheless simple and effective. PDF copies of alphabetically and chronologically organized bibliographies are available by way of links on each bibliography's web page; these PDF copies are designed to be downloaded and printed.

The bibliography is constructed from a variety of sources. DDLE editors consulted the bibliographies of works current in evolutionary biology to identify a core set of works frequently referred to. These works focus on paleontology, population genetics, and works of the architects of the "Modern Synthesis" such as Sewall Wright and R. A. Fisher. Histories of evolutionary biology were also consulted. Second, a significant proportion of the Literature of Evolution's references describe works referred to by Darwin in his correspondence. The list of these works was provided by the Darwin Correspondence Project and was digitized by DDLE editors. A third significant portion of the reference list describes works owned by Darwin; references in this group are derived from a list of works provided by *Darwin's Marginalia* (Darwin 1990). Annotations to references from this group indicate key facts about the work of interest to Darwin scholars. For instance, some works are annotated to show whether Darwin obtained them before or after his *Beagle* voyage, providing a picture of the background of ideas against which Darwin's trip took place. Other works are annotated to show whether they were inscribed by Darwin or another person. A third type of annotation shows the current location of Darwin's copy of the work. Of course, works in many languages appear on the reference list.

Because many records were obtained *en masse*, and because these references originate from a range of sources, quality control presents a challenge for the DDLE editors. Checking references to make sure all meet a high standard of completeness, accuracy, and readability is ongoing.

Concluding Remarks

The DDLE (http://darwinlibrary.amnh.org) was founded with the aim of creating a portal to the growing body of literature online concerning evolutionary biology. The Darwin Manuscripts Project aims to create content developed for storage and distribution at the DDLE site. Manuscripts are selected according to their role in Darwin's thought and the history of his ideas. This approach is exemplified by the initial content deposited at the site, which is organized around Darwin's preparations for writing the *Origin.* The methods for editing and annotating Darwin's manuscripts are intended to produce transcriptions whose content is equal in value to the manuscripts from which they are derived.

The Literature of Evolution complements the Darwin Manuscripts Project, intended to organize the literature of evolution by subject. At present, however, the approximately 3,500 references are organized only by alphabetical and chronological order.

The central organizing principle of the DDLE reflects the technique of expert indexing referred to earlier in this paper. The accumulation of texts and references is balanced by their organization into themes or natural units for the history of Darwin's thought. This is exemplified by the site's current manuscript content, aimed at offering researchers a view of the development of the line of thought that Darwin took in his final approach to the *Origin*.

The Complete Work of Charles Darwin Online

The Complete Work of Charles Darwin Online (hereafter, "CW") has the aim of obtaining digital copies of as many of Darwin's works as possible and distributing them online (except in the case of correspondence, handled by the Correspondence Project). One of the site's central achievements is to have digitized the entire body of Darwin's published works, including the various editions of each and their translation into a range of languages. The site also provides scanned microfilms of a large collection of Darwin's manuscripts; few transcriptions are provided, however. The site continues to grow as more material is rapidly added. Original copies of most of the materials on the site were obtained from the Cambridge University Library.

Personnel

Dr. John van Wyhe is the CW project lead. Dr. van Wyhe has a range of affiliations with Cambridge University, including a position as an Affiliated Research Scholar in the University's Department of History and Philosophy of Science. Before he began concentrating on offering Darwin's work online, his research interests were in the history of phrenology.[11] He is assisted by Research Associate Dr. Kees Rookmaaker, a zoologist whose research focuses on the rhinoceros, and who has contributed to the bibliography of the subject. The site's technical director is Dr. Antranig Basman, a Cambridge University Information Engineering Ph.D.

[11] Curiously, Dr. van Wyhe's "about" page does not provide any information about his academic training.

Dr. Basman's experience includes work in the technology industry as an executive, and he is the project lead on a web application software package known as RSF, which is an important part of the CW site's digital infrastructure. The Associate Editors are James Secord, director of the Darwin Correspondence Project and a faculty member of the Department of the History and Philosophy of Science at Cambridge University, and Janet Browne, a biographer of Darwin and a faculty member at Harvard University's Department of the History of Science.[12]

The Home Page

The bulk of the content available on the CW site is distributed across two component collections, one for published works, the other for manuscripts. Understanding this is essential to successfully navigating the site. These two content-rich sections of the site are accessible by way of the home page, which also provides a central section of links to texts likely to be of interest. Before elaborating on the two content-rich sections of the site, the content and design of the home page is worth considering because it provides a useful point from which users can orient themselves as they explore the site.

Users visiting the site home will find themselves taken to a page with sidebar navigation and a set of buttons across the top. The sidebar links to what are best thought of as site utilities: an introduction for new users, "What's New," "Feedback," "Press" (for news reports about the site), and a link to information about a census of Darwin texts being conducted. The buttons along the top take the user to site content. One button takes users to Darwin's publications, the other to manuscripts; another directs the user to biographical information about Darwin. Users must look to the sidebar for a link to a third type of content, illustrations. The main body of the page contains a brief statement about the site's contents, with links to site highlights: All six editions of the *Origin*; selected volumes of the *Journal of Researches*, also known as the *Voyage of the Beagle*, in which Darwin and others report the scientific findings during their time on the *Beagle expedition*; the *Descent of Man*; and a selection of Darwin's notebooks, labeled "Evolution notebooks."[13] Following each of these links brings the user to a work's listing on one of the main sections of the site, the main page for published works, or the main page for manuscripts.

The "Publications" and "Manuscripts" buttons appearing horizontally across the top of the site bring the user to one or the other of the central content locations on the site, viz., the site's central access point to publications and manuscripts, respectively. Recognizing this is the key to finding one's way about the site, which can be daunting. The site's minimalist architecture provides few navigational pathways, works generally being presented in chronological order in long lists, no other categorization or organization being provided. A user unsure of his or her location in one of these lists can reorient him- or herself by returning to the home page and re-entering the section of interest by following either the "Publications" or "Manuscripts" links.

Publications

The "Publications" button on the home page brings the user to a single long web page, which is divided into three main sections. The user can jump to any one of these sections by following links at the top of the page: "Books," "Articles," and "Published Manuscripts." A fourth link for "Supplementary Works" takes the user away from the long web page for publications to another long web page.

Books and Articles

The section of the site for books lists those published by Darwin, in the order in which their first editions appeared. Derivative works—editions after the first and translations—are listed with the first edition. For instance, the six editions of the *Origin* appear grouped with the 1859 first edition, as do translations of the *Origin*. In general, British editions of English-language publications are presented, but in some cases, American publications are also offered. A small finch icon appearing next to some works' listings indicates that the illustrations from the work can be viewed together as a group, independently of the text, in a separate window.[14]

The interface to the published works offers users a choice: a scan of each page may be viewed alone, a transcription of each page may be viewed alone, or the two may be viewed side by side in a single split web page. It is possible to synchronize each side of the split so that advancing the transcription one page also advances the scanned image one page. A page on editorial

[12]This information about the site's staff was obtained from the site's "Acknowledgments" page.

[13]An important *caveat lector* concerning the "evolution notebooks" is offered below.

[14]Illustrations can also be accessed by following the "Illustrations" link on the site's home page.

practices states that each transcription is intended to be an exact copy of all information in the original, with the exception of line breaks. Pages of the transcription and those of the original differ in width, so lines will break at different points on each. End-of-line hyphenation will also differ because of this. Transcriptions indicate page breaks, however, so that someone reading the transcription can identify which page of the original he or she is looking at. Of course, because a scanned original can be readily seen, it is easy to confirm that the transcription is correct. Illustrations appear in the transcription and, of course, in their context in the scanned images of the original.[15]

A PDF file of an entire scanned work can be downloaded. These PDF files are quite large, the *Origin* first edition weighing in at approximately 97 MB. These files are appropriate for printing; users are advised to use the interface to their operating system's print tools to scale the page images so that each fills an entire physical page of printed output. The actual size of each image is quite small, requiring scaling-up before printing. After downloading the PDF, a user intending to print it will select "Print" from his or her PDF viewer's menus, bringing up the viewer's print dialog. Attentive users will note that most PDF viewers provide the user with a means of choosing the scale at which the document is to be printed, usually in the form of a small text-entry box in which a percentage by which the document is to be increased or decreased in size can be entered. The "natural size" of the documents downloaded from CW is quite small; if the printing scale is set at 100%, a miniature of the work is printed, each page being only a few inches on each side. Instead of setting a scaling factor, users should select the option provided in most PDF viewers that will scale each PDF page to fit the physical page, increasing its size so that it fills an entire printed page. This option will usually appear as a check box or "radio button" entitled something like "scale each page to fit paper." Printing at the proper scale, an Internet user can obtain a clear, readable copy of one of Darwin's published works formatted just as Darwin and his nineteenth century publishers intended.

The section for articles offers links to works by Darwin that appeared in periodicals. Like the books, the articles are listed in date order. The viewing interface to the articles and their transcriptions is the same as for books. Texts can be read in a text-only view, an image-only view, or a side-by-side text and image view.

Published Manuscripts and "Supplemental Works"

The third main section of the publications section of the site provides published manuscripts. As discussed above in connection with the Darwin Manuscripts Project, manuscripts are works by Darwin not intended for publication, written in his own hand, including notebooks, reading notes, sketches of ideas, and the like. The paradox of a "Published Manuscripts" section disappears when one recognizes that this section of the site presents scanned images of print works consisting of manuscript transcriptions.

A *caveat lector* ("let the reader beware") is in order here. The problem is that an outdated edition of the transcription of some of Darwin's notebooks is presented on the site. The site provides scanned images of the pages from de Beer's 1960 print publication of transcriptions of some of Darwin's early notebooks. Darwin scholars have ceased to refer to the de Beer volume; the new standard work is the 1987 volume of transcriptions of these same notebooks edited by Barrett and colleagues (Darwin 1987). The CW site warns readers of this, noting that the works at issue have been superseded by the 1987 work. However, CW does not link to this later work. Users are advised to visit the DDLE, which presents the up-to-date Barrett work in its "Library" section.

Finally, the supplemental works available on the site are not by Darwin. Rather, they are works selected for their interest to the Darwin historian. These include secondary works on Darwin, listed in a bibliography; works about specimens collected by Darwin; and reviews of Darwin's work.

Manuscripts

The site's manuscripts section—the second main subdivision of site content, in addition to published works, as discussed in the "The Home Page" section above—is accessed from the home page. The difference between this section and the section for published manuscripts discussed above is that the works presented in the latter are transcriptions of manuscripts that have appeared in print. In the "Manuscripts" section, works are listed in the order in which the Cambridge University Library initially processed them, by their "DAR" number. As Darwin's manuscripts were unpacked in Cambridge, they were placed in large archival binders; each binder

[15] Information in this paragraph about editorial practices and formatting may verified by consulting http://darwin-online.org.uk/Print_transcription_policy.html.

has a DAR number. Access to manuscripts is provided by way of a long table, presented on a single long web page. The left column of the table lists DAR numbers, a first central column provides a link to the manuscript, a second central column shows a thumbnail image of a page of the manuscript, and the right column contains a brief description (at most, a short paragraph) of the contents of the manuscript. Users can view one page at a time. Generally, the images are scanned microfilms from the Cambridge University Library.

In addition to that mentioned in connection with the de Beer transcriptions discussed above, a second *caveat lector* is in order for Internet users viewing manuscripts provided in this section of the site. As discussed at length above in connection with the Darwin Manuscripts Project, manuscript images contain a fraction of the information contained in the original. Deletions, marginal notes, and interlining can be identified, but the type of writing instrument, paper characteristics, and the "layering" that reflect the history of Darwin's creation and revision of his notes cannot be discerned. Researchers viewing these manuscripts cannot use them to reconstruct Darwin's thought: Comments written between two lines many years after the initial lines will appear contemporaneous in a scanned image with no annotations. Some manuscripts are transcribed, which will help readers decipher Darwin's handwriting, but no annotations are provided that allow users to identify the type of writing instrument used or paper characteristics.

Users are also advised that the DAR numbers do not correspond to any natural archival unit. Manuscript material is grouped by the DAR numbers solely for the purpose of managing the binders in which the manuscripts are stored. Darwin's notes on a given topic or in preparation for a given published work may span several DAR numbers, and in general, the only way to discover the contents of a collection of manuscript material with a given DAR number is to browse the collection.

Concluding Remarks

The CW is an excellent source for viewing Darwin's published works. Users must come to the site with a minimum of information about the works they are interested in because the site does not provide a high level of organization for its content. Manuscripts offered on the site are less useful than they otherwise might be because microfilming and scanning destroys much of the important information they contain. As well, the transcription of an important set of Darwin's notebooks by Gavin de Beer offered on the CW site is known to be outdated. Users are urged to consult the standard, up-to-date transcription by Barrett and colleagues at the DDLE.

The CW site makes little progress toward increasing access to Darwin's works. It is not constructed around any central organizing principle or site design that would help direct users toward works based on their content. The site's architecture is built around the types of material presented. No topical or thematic categories cut across categories for material type, date-ordered lists providing the only further classification of works. There is no keyword indexing. Users will most likely find it most useful for known-item searches.

Conclusion: So Close, but Still so Far Away

The three sites reviewed here offer the Darwin researcher collections that are both deep and wide. The collections do not overlap, except in the case in which the DDLE and CW sites offer a very small number of manuscripts in common.[16] The CW site is an excellent resource for published works by Darwin. The Correspondence Project clearly ranks as one of the most important digital collections anywhere due to the quality of its transcriptions and its breadth of material. The DDLE has only produced a limited number of texts but offers high-quality work, and its expansion is underway. The Complete Work's "Supplemental Materials" section and the Literature of Evolution component of the DDLE serve as important resources for those looking for both primary and secondary works on Darwin and evolutionary biology.

The architecture of the sites and the organization of the materials collected on each reveal a deep lacuna in the level of Internet access to research materials by and about Darwin. The Correspondence Project and the Complete Work distribute a large number of important materials but offer few or no substantial natural, content-driven pathways into their collections. The DDLE has set for itself the goal of organizing its manuscript collections and the literature of evolutionary biology by subject. Nonetheless, it has not attained this goal at present. On the one hand, digitization efforts by Darwin scholars have come so very close to an important goal, universal or near-universal distribution of Darwin texts and supporting materials. On the other

[16] Those of the DDLE are to be preferred because they are more up-to-date and preserve more historical information.

hand, it is all too plain that they are still so very far away from obtaining an even more central and important aim, improved *access* to the riches of the content of those materials.

References

Darwin C. The foundations of the origin of species: two essays written in 1842 and 1844. In: Darwin F, editor. Cambridge: Cambridge University Press; 1909.

Darwin C. Charles Darwin's Natural Selection; being the second part of his big species book written from 1856 to 1858. In: Stauffer RC, editor. Cambridge: Cambridge University Press; 1975.

Darwin C. The correspondence of Charles Darwin, vol. 1–3. In: Burkhardt F, Smith S, Kohn D, Browne J, editors. Cambridge University Press: Cambridge; 1985–1987.

Darwin C. Charles Darwin's notebooks, 1836–1844: geology, transmutation of species, metaphysical enquiries. In: Barrett PH, Gautrey PJ, Herbert S, Smith S, editors. London: British Museum (Natural History); 1987.

Darwin C. Charles Darwin's marginalia. In: Di Gregorio MA, Gill NW, editors. New York: Garland; 1990.

Darwin C. A calendar of the correspondence of Charles Darwin, 1821–1882, with supplement. In: Burkhardt F, Smith S, Kohn D, Montgomery W, editors. Cambridge: Cambridge University Press; 1994.

Eldredge N. Reinventing Darwin: the great debate at the high table of evolutionary theory. New York: Wiley; 1995.

Eldredge N. Darwin: discovering the tree of life. 1st ed. New York: W.W. Norton; 2005.