I. Introduction

The Darwin Correspondence Project is a digitized collection of all known letters sent to and from the famed nineteenth century naturalist, Charles Darwin. The dataset includes letters sent by and to Charles Darwin (transcriptions as well as scans of the original), detailed explanations and contextual information for each letter, including scientific concepts, biographical details about the correspondents, and historical context, and metadata for the letters such as dates, correspondents, and synopsis of topics discussed. The collection has over 15,000 letters exchanged with nearly 2,000 correspondents, from 1821 to Darwin's death in 1882.

II. History

The webpage describing the project's history explains that the project began in 1974. It was initiated by American scholar Frederick Burkhardt, his wife Anne Schlabach Burkhardt, and zoologist Sydney Smith from the University of Cambridge in the U.K. Their goal was to compile all letters written by and to Charles Darwin, aiming, initially, to publish summaries of these letters – but they eventually decided to publish full transcripts of these in chronological order.

By 1975, the project established a dedicated team of researchers and editors in both the U.K. and the U.S., with the U.K. team based at the Cambridge University Library. This library holds the most extensive collection of Darwin's manuscripts and his personal collection of books and journals. In the U.S., the project relied heavily on volunteers working from Frederick and Anne's home in Vermont, until 2013, when a professional team operated out of Harvard.

III. Purpose

The collection aims to shed light on Darwin's thought processes, the development of his theories, his interactions with the scientific community, and the cultural context of the nineteenth

century. As a History concentrator, I could potentially use the dataset to explore the relationship between natural history and colonialism, or to probe the personal dimensions of Darwin's work.

IV. Processing

The Darwin Archive in Cambridge contains approximately 9,000 letters. A decade was spent searching for additional letters, acquiring copies, and transcribing them into a digital format, creating an electronic archive that supports the project's publication efforts. This includes printed volumes titled "The Correspondence of Charles Darwin" and online resources. A significant task was to date the letters, as less than half were fully dated by the sender and identifying unknown correspondents. Processing the data thus involved locating, obtaining, scanning, and transcribing letters into an electronic format, followed by rigorous research and annotation for contextual clarity.

V. Quality

The dataset is very easy to access and use, offering not only transcripts alongside scans of the original letters but also contextual annotations that explain historically situated terms and give insight into the cultural, scientific, or personal backdrop of each letter. The letters are also cross-referenced and hyperlinked, making it easier to find related letters and discover connections between different correspondences.

The dataset's advanced search function is another asset that allows users to filter letters by year, term, location, correspondent, etc. This is especially useful in the context of historical research. For instance, I searched the term "Tierra del Fuego" where I visited this winter, and where Darwin had stopped in 1832-33 during the *Beagle* voyage. I read Darwin's (deeply racist) observations about the indigenous people of Tierra del Fuego, as well as his reflections on the landscape. I could draw insights from these about his scientific and sociological convictions.

The chronological representation of the letters in bar graph format also provides a visual overview of Darwin's correspondence over time, highlighting periods of intense communication. These shed light on his working habits and the development of his thought.

VI. Limitations

As with any historical archive, the collection might not be (almost certainly isn't) exhaustive. The letters reflect the biases of Darwin and his correspondents, who were primarily part of the Oxbridge-educated scientific elite. While this is instructive from a scholarly standpoint, it is concerning for a public access archive (especially one that is encouraging use in schools) if it isn't engaged with a critical eye.

Moreover, although the collection is widely accessible and has a straightforward interface, the sheer volume of pages and articles on the home page is slightly overwhelming at first. I also noticed, while scrolling through the letter timeline, that the page tends to jump around at times. As a result, I kept losing my place on the timeline which was slightly frustrating – and because I was collecting letters on Tierra del Fuego, I'm concerned that I may have skipped over a some relevant letters because I landed on a different part of the timeline each time.

VII. Conclusion

Despite these minor navigational challenges, the overall quality and accessibility of the dataset makes it an invaluable resource for scholars, especially historians of science. This is because of the insights it potentially offers into the social networks of science in the Victorian era, and possibly also the ways in which these were steeped in colonialism. Interactive features like maps and thematic collections allow for a richer nuanced understanding of the geographic and topical spread of Darwin's correspondences. And as a pedagogical tool, the dataset allows students to immerse in a historical archive and learn how to critically engage primary sources.

The project is continually evolving, with efforts to locate missing letters and update the database. The team collaborates with other archives and private collectors worldwide to ensure the collection's growth, and this is complemented by educational initiatives like workshops and public lectures. In these ways, the Darwin Correspondence Project not only preserves a rich archive but also helps democratize access to information without the barriers of location or cost.

VIII. Spreadsheet

Dataset name	Darwin Correspondence Project
Link to data source	https://www.darwinproject.ac.uk/letters/darwins-letters-timeline
Link to storage source	The physical archive is in the University of Cambridge Library. The digital archive is on the project's website.
Who collected the data?	A team led by the founders of the project, namely Frederick Burkhardt, Anne Schlabach Burkhardt, and Sydney Smith. Contributions were also made by researchers and editors from the U.K. and the U.S.
Who owns the data?	Cambridge University
How was the data collected?	Archival research including collaborations with libraries and private collections. Data processing involved digitization of letters for transcription and annotation.
Sample size	15,000+ letters with nearly 2,000 correspondents
Who was included or excluded from the sample?	Fellow scientists, family members, government officials, administrators, friends, acquaintainces, etc.
When was the data collected?	Began in 1974 and the bulk of collection took from 197 to 1984, but documentation efforts are ongoing
When was the data last updated?	Continually updated as new letters are found or existing entries are revised and/or annotated. But the last major publication, marking the completion of the print edition, was in 2022.
Why was the data collected?	To provide insight into Darwin's life and work, the development of his theories, and perhaps also to document the scientific and cultural context of the nineteenth century

Notes on data quality	Transcripts are easy to read, contextually annotated, cross-referenced, and advanced search functionality allows filtering by year, term, location, correspondent, etc.
Notes on data usage conditions	Widely accessible and straightforward interface, although the number of pages and articles is slightly overwhelming and the page tends to jump while scrolling through.