Stanford University

October 16, 2020

Dear Librarian of Congress,

We are writing on behalf of the Data-Sitters Club, a research group under the Stanford Literary Lab, in support of an exemption to the anti-circumvention provisions of the Copyright Act to allow researchers like us, and the students we teach, to access ebooks for fair use research purposes relating to computational text analysis.

Computational text analysis methods allow literary scholars to ask and answer questions that previously would have taken decades of painstaking research, if they were possible at all. This analysis has value whether it is about the works of William Shakespeare or more recent authors outside the traditional literary canon. The critical analysis of modern, popular texts is a vital part of humanities research; it helps us to understand how books both mirror and shape people's understanding of the world and the major issues of our time.

In the 1980's and 1990's, Ann M. Martin and a team of ghostwriters wrote a total of over 200 children's books, known collectively as the *Baby-Sitters Club* series. It is an iconic depiction of girlhood in the upper-middle-class American suburbs of the time, and was tremendously popular with elementary- and middle-school-age girls at the time. Its distinctive characters personally resonated with many girls; the 2020 documentary *The Claudia Kishi Club* focuses on the impact of a character who was one of the few broadly popular Asian-American role models during those decades. There's been relatively little scholarship written on the series, and what has been published focuses on the close reading of specific, individual texts. Applying the tools and methods of text and data mining to a corpus like the *Baby-Sitters Club* can make it possible to address a different set of questions. It allows researchers to draw upon all the books at once in order to gain an understanding of the totality of this series and how it builds its fictional world.

The Data-Sitters Club has begun to explore a broad agenda of research questions in relation to the *Baby-Sitters Club* series. Each novel is written in the voice of one (or multiple) characters, by Ann M. Martin herself or one of numerous acknowledged ghostwriters. Using computational methods, we are interested in whether each character has a distinct voice, and whether that voice is different across writers. We are interested in whether non-narrating characters themselves have distinct voices expressed through their dialogue, or if they just form classes of character types like "generic mother" or "generic classmate". We would like to find out how the characters' "written" language (shown through the portions of the text in the characters' "handwriting") differs from their implicitly spoken text through the first-person narration. The *Baby-Sitters Club* is (in)famous for its use of tropes, such as Claudia Kishi's "almond-shaped eyes", or

“Mal is white and Jessi is black”. We are interested in what else can we find out about how and where explicit text reuse happened in the most formulaic parts of the book, where the premise and characters are described in order to orient new readers. We are interested in how these books treat religion, race, adoption, divorce, and disability. The instructive role of children’s literature and the popularity of this series make it a particularly valuable one to study as a step towards understanding the worldview of American women currently in their 30’s and 40’s.
Finally, we are interested in adaptations into new media formats: what material was included (and what was removed or significantly transformed) in the creation of a recent graphic novel series, and a Netflix series, based on the original books.

The Data-Sitters Club also has pedagogical aims: we write up our process -- the decision-making and interpersonal aspects of our work, along with the technical steps -- and publish them as “books” on our website. Our goal is for anyone to be able to apply the same methods to texts and questions that interest them, and these “books” have already been incorporated into course syllabi by professors at Emory and Northeastern Universities. There remains one significant barrier for other people to do this same kind of work: access to texts.

Computational text analysis is not possible without text files, whether they come from ebooks, or are digitized from scans of printed books. While a vast amount of literature (including the entire Baby-Sitters Club corpus) is available for purchase as ebooks, which could be trivially easily converted to the plain text format used in computational research, most ebooks are protected by a technological protection measure (TPM). Although TPMs were intended to prevent piracy, for us they are often a roadblock to lawful and socially valuable research. To obtain the text in the necessary format without risking liability under the anti-circumvention provisions, scholars must go to great lengths. Typically, this involves scanning a book, and processing those scanned images using Optical Character Recognition (OCR) software, which generates usable text corresponding to the words that appear in the image. OCR is imperfect, and frequently makes mistakes, particularly if words are distorted near the edge of the page. Scanning a 130-page book (like one of the books in the *Baby-Sitters Club* series) can take 15-20 minutes, OCR can take another 10, and double-checking and correcting the OCR can take anywhere from 10-40 minutes, depending on the number of errors. The OCR error rate is particularly problematic in the sections of the *Baby-Sitters Club* books written in handwriting-style fonts, which OCR very poorly and need to be transcribed manually. These numbers increase when working with longer books, or books with complex formatting like tables. While scholars affiliated with a well-resourced institution such as Stanford may be able to bear the costs associated with paying someone to do this work, the costs are prohibitive for scholars at the vast majority of institutions in the US, including smaller public institutions and community colleges.

While computational methods can allow scholars to ask questions about thousands or even millions of books, the feasibility of doing that work plummets when that requires thousands or millions of hours of scanning and OCRing, even for a version of the text that contains errors. Converting an ebook, in contrast,

takes less than five minutes, and does not introduce any errors in the resulting text file. We purchased books we scanned for the project for a couple dollars, as used copies or library cast-offs. Even books that are generally in poor physical shape are fine for scanning and OCR. But if we were able to circumvent TPM without risking legal liability in order to build a corpus using ebook files, we would be happy to purchase ebook versions from the publisher. Circumventing TPM rather than scanning and OCRing books would enable scholars to spend more time pursuing research questions, allowing them to pursue projects with a more ambitious scope. Were it not for the legal uncertainty created by Section 1201, we could imagine in the next three years expanding the scope of our project to contextualize the Baby-Sitters Club within series books for girls, or even children's literature more broadly. Furthermore, it would become feasible for all of us — regardless of institution — to incorporate computational analysis of modern texts into the curriculum, enhancing students' awareness of the possibility and limitations of digital methods, using material that is more familiar and resonant than the public domain.

We urge you to consider adopting the proposed exception to the anti-circumvention law both to make computationally-supported research feasible without the extreme costs of needless digitization when digitized copies already exist as ebooks, and to support copyright holders in securing the ebook purchases of scholars with an interest in legally building research corpora.

Sincerely,

Lee Skallerup Bessette, Georgetown University
Katherine Bowers, University of British Columbia
Maria Sachiko Cecire, Bard College
Quinn Dombrowski, Stanford University
Anouk Lang, The University of Edinburgh
Roopika Risam, Salem State University