The Editor and the Aggregate

Nigel Lepianka

ABSTRACT

As big data becomes more prominent in the humanities, scholars engaged with the method have begun to question the history of the data they use. A line of thinking about the transmission of data is inherently bibliographical and editorial, and presents an opportunity for textual scholars whose methods have long relied on understanding the way data circulates. Using Lyle Wright's American Fiction *bibliography, and the collections and digital projects and it has informed, this essay traces the way literary data can be affected by the goals and ideas of those who compile it.*

MY PURPOSE IN THIS ESSAY IS TO ARGUE THAT TEXTUAL SCHOLARS ARE ABLE TO TREAT aggregate collections of data as texts, and thus, these data may be analyzed and critiqued within the framework of textual criticism and editorial theory. A dataset, like any other text, can be subject to the same phenomena as individual texts: a dataset can be transmitted, revised, and read, it can be published, it can be circulated, and it can even be corrupted. The differences between an individual work, for example a novel, and a collection of numerous units of data, such as an enumerated bibliography of novels, are not as stark as they may appear. This approach will be necessary as the humanities continue to push forward with adopting computational and algorithmic modes of reading texts and grappling with our cultural history at a scale beyond the individual text. As textual critics and editors, we can apply our methodologies to aggregate data in order to both ensure accuracy and make explicit the way data is affected by the culture that produces it.

Textual scholars have expressed some reticence towards large-scale analysis of text but have also served to aid the practice by providing standards for encoding (via the Text Encoding Initiative), or in the creation of collections of digital texts and databases of literary materials that afford text mining and other forms of computational analysis. However, I would argue that textual scholars are in a prime position to also treat the emergence of data as a concept in the humanities, as an opportunity to expand the sphere of what counts as textual scholarship. Textual scholars are already equipped with the skills and knowledge to critique the editing and transmission of data, and the familiarity with adjacent communities of book history and bibliography can serve the textual scholar well when approaching data, especially when it is

revealed where that data originally comes from. That is, in many cases, from print bibliographies, catalogs, and other physical or analog sources that precede digital forms of data.

This essay will first state what is under the umbrella of the terms “data” and “dataset”, and how they can fit under the current model of textual scholarship's approach to text. Secondly, it will argue for a means of analyzing a collection of data' creation and transmission using a case study of the Wright *American Fiction* bibliography, a list that has formed the foundation of notable resources for scholars, including library collections and catalogs, microfilm collections, and most recently digital repositories and databases of literature. Lastly, it will connect these ideas of the history of data to the ways in which scholars can approach the editing of collections of data.

To approach the first point, let me address the concept of data and clarify where I see its textual qualities to be most pertinent to the discussion. My preferred definition of data is that offered by the Open Archival Information System (OAIS) for its clarity, breadth, and nuance.[1] The OAIS definition declares data as:

> A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen. (Consultative Committee for Space Data Systems 2002, 1-10)

The term "reinterpretable" is most significant here, partially because of the repetition of its use in the definition, and how it invokes two different meanings in this context. The first of the meanings is communicative; this sense of reinterpretability refers to the act of reading and the derivation of meaning from data on the part of the reader. The second is the more innately textual meaning of reinterpretability; the ability for data to move (or, to be emphatic, to *be moved)* across media, brings to the foreground the bibliographical nature of data.

While the OAIS definition does hint at the unfixed nature of data, it bears only a trace of the fact that data itself has undergone interpretation already when it has been sorted, arranged, pared down, etc. In its use of the term "representation", the prefix re-, or "again", suggests a state prior

---

1 OAIS was adopted as ISO standard 14721 in 2002. Originally a product of the Consultative Committee for Space Data Systems, the OAIS prescribes a system for archival workflows and digital preservation. While I am primarily concerned with the way it has defined the term data and how this may help us approach bibliographical descripton, the OAIS covers a large, interdisciplinary and complex model of archives, preservation, and access. (Consultative Committee for Space Data Systems 2002)**.** For the OAIS in a humanities context and its relevance to the archives, see (Kirschenbaum 2013)

to that of data. Thus, we may say data is not innocent of human involvement, tampering, or subjectivity, and thus, data itself is not a neutral object, but susceptible to ideology via the methods and practices that inform the person producing the data. Johanna Drucker has argued for more humanistic approaches to data in her recent work. Drucker has stated in multiple venues that "data", derived from the Latin *datum* ("that which is given"), is instead taken, and thus is *not data,* but *capta* (2014, 128-9; 2011, 3). Drucker's aim in the use of the word *capta* to describe what is more commonly known as data as "always interpreted" as "no data pre-exists its parameterization" (2014, 129). Parameterization, according to Drucker, is a construction, and the term *capta* opens up the possibility of recognizing and "acknowledging the constructedness of the categories according to the uses and expectations for which they are put" (129). The parameterization that Drucker locates as inherent to data/capta is synonymous with the formalization and representation the OAIS definition prescribes.

Accepting the OAIS definition and Drucker 's addendum, we can then frame a discussion of data with textual critical theories. The act of creating data, of parameterizing it to use Drucker 's term, is an act that codifies the information into a readable form. In accordance with Jerome McGann's textual condition, this act would constitute a textual event, in which the linguistic and bibliographical codes inflect the way data is represented and read. To clarify what I mean, let me list a few notes of integral information that solidify the properties of a dataset when considered as a text. To provide examples, I will discuss a hypothetical bibliography of fiction:

- Aggregate data can be read as a single entity but contains various parts that compose the whole. **Ex.** A listing of fiction will necessarily contain multiple references to individual fictional works such as Susanna Rowson's *Charlotte Temple* (1791).

- Aggregate data will attempt to make its individual parts conform to an overarching standard of format and parameters **Ex.** recognized standards for citation, e.g. Chicago or MLA. Or, the near-universal standard of listing authors first in a citation.

    - These standards can be local to the aggregate or exist within a larger culture or ecology that defines such standards. **Ex.** A bibliography that uses the legal names of authors over their more well-known pseudonyms, leading to placing *The Adventures of Huckleberry Finn* under Clemens rather than Twain.

- Individual pieces of information may be more important than other pieces, or possess a different level of cultural value. **Ex.** A list of fiction may contain both Hawthorne's *Scarlet Letter* (1850) and James Knorr's *Two Roads* (1854). The former is a canonical standard, the latter is essentially unknown.

- A dataset can be read as a full, totalizing text or accessed in discrete parts. **Ex.** A bibliography can be mined to develop a graph of books published by year, using the entire bibliography for information. On the other hand, one may look up an individual title such as *Moby-Dick* in order to locate its publication year specifically.

This argument is necessarily two-pronged, as it connects two different, and oftentimes at odds, fields within the general area of digital humanities: that is, textual editing and big data.[2] Some textual scholars have indicated a hesitance towards how the methodologies, including the algorithms and practices, of distant reading work with textual materials. By way of example, we may look at the process (or pre-process as some might call it) of "cleaning" data before its ingestion by an algorithm. Cleaning refers to the process by which data is made more uniform in terms of its structure as well as more idiosyncratic changes to the text to make the output of any computational process more suitable to the goals of the process. A corpus of plays may have their stage directions or cues removed. A collection of novels may remove paratextual information such as chapter headings. Punctuation, in general may also be removed or the verbs of texts truncated to their stems. These examples would all be done in order to aid in the process of machine reading, in order to make sure the information is all read the same by the algorithms. Amanda Gailey, for instance, has expressed a very real and important concern with the way texts are handled by algorithms. Responding to the need for data within a set to be "cleaned" and standardized in order to provide meaningful results from an algorithmic process, Gailey argues: "Standards run the risk of homogenizing information; they track similarities and require conformance in order to perform their function of enforcing both rigor and economy" (2012, 354).

Gailey's concerns may also be an opportunity for textual scholars, bibliographers, and book historians whose jobs it is to critique the circumstances that surround the presentation and distribution of data. Certainly, the removal of punctuation, paratextual matter, and other such

---

2 Other familiar terms may be algorithmic criticism, cultural analytics, or distant reading. I will default to big data as my term of choice in this essay as I believe it encapsulates and alludes to some of the qualities of traditional bibliography that I will discuss.

components of textual information should present challenges to those invested in the material conditions of texts, as scholars within the fields of book history, bibliography, and textual scholarship insist that these conditions reflect upon the semantic content and the interpretation of that content. Additionally, the threat of conformity and standardization can be seen as equally concerning because it presumes that the uniqueness of a text may be removed in the process of its cleaning. A text that has a publication or printing history that differs from the presumed standard models will find its voice drowned out in favor of an overarching format. This is a condition of the structure of data and datasets, but can be viewed and should be acknowledged for what certain processes may do in terms of what they erase and omit. This sort critique and the work associated with it have been argued by scholars working with big data. Sarah Allison has recently published a short but provocative essay in the *Journal of Cultural Analytics* where she argues for the methodologies of textual scholarship (though she does not call out the field by name) to be practiced by those who use datasets. Allison advocates for what she calls a "turn to the *byproducts* of cultural analytics—to more project-specific tools, documentation, and discoveries." Allison here likens the idea of finding someone else's data as "like a new manuscript: an unexplored object that deserves attention in its own right." This metaphor, as Allison later explains in a conversation with Andrew Goldstone, also demands the asking of certain inherently bibliographical questions: "is this an authoritative source? what process created it? what is the chain of transmission by which it reaches us?" and so on (2016).

This is a feasible task for the textual scholar because we can reconcile the processes by which data and critically edited texts come before their readers. That is, when Johanna Drucker says that "no data pre-exist their parameterization" she is calling attention to the fact that the discrete interpretable units we refer to as data have *already been interpreted* and made to fit into a framework that allows for their processing by others. This is the process of scholarly editing distilled into other terms. Compare Drucker's claim about the already-interpreted nature of data to the definition of scholarly editing provided by Peter Shillingsburg: "editorial efforts designed to make available for scholarly use works not ordinarily available or available only in corrupt or inadequate forms." (1996, 2) Shillingsburg's definition denotes that editing focuses on providing a text through the editorial work, an act that itself requires an interpretation of the text, in order to disseminate the text in both a more accessible and comprehensible form for the purpose of proliferating interpretations and readings of a text. Both scholars have argued the same points, with Drucker casting a wider net than Shillingsburg: there is a process by which some data are

worked by one person, incorporating all of the human fallibility and interpretative capacities, and then that data is disseminated as it has been read by another and taken up by others for use.

Following the provocation of Allison, the comparison of Shillingsburg and Drucker's approaches to information, we may then accept the claim that a dataset may be considered a text. However, it is also necessary to recognize that in the conceptualizing of datasets as texts is also to accept two implicit premises of textuality:

1. The process of creating a dataset involves subjectivity; the creation of data is interpretive and subject to human decisions that may subsequently be depicted as fact.
2. A dataset makes an argument, or several, in its composition and organization.

As a primary case study, I examine the example of Lyle Wright's three-volume *American Fiction,* an enumerative bibliography of over 10,000 titles from 1774 to 1900. The first volume of *American Fiction* was published in 1939, with volume 2 following in 1949, and volume three in 1966. Wright's bibliography stands as one of the foremost attempts at creating a comprehensive list of American fiction and was used quickly after its publication not only as a reference work, but also as a guide for libraries seeking to expand their collections. Wright's bibliography was used by Research Publications LLC in the 1960s and 70s to construct a microfilm collection of early American imprints. Today, the Wright bibliography can be found on the web at the University of Indiana's *Wright American Fiction Project* or in Gale Cengage's *American Fiction 1776-1920* dataset. Both of these digital iterations of Wright's work afford the ability for computationally minded scholars to explore the texts Wright describes. Additionally, however, what these versions, including both the digital and microfilm, of Wright's work do is replicate some of the Wright bibliography's underlying issues. A listing of fiction is just as susceptible to interpretive judgments, including bias and errors, as any other piece of communication. It would be naive to expect a list such as Wright's to be perfect, as he himself implies with the bibliography's coy subtitle-"a contribution toward a bibliography." Projects that use Wright, however, fail to acknowledge Wright's limitations and in doing so, implicitly accept the argument that Wright's work was complete.

American Fiction has its deficiencies, however; some of them are by design and others are erroneous. As Wright notes in his preface, "In general, it has been intended to omit annuals and gift books, publications of the American Tract Society and the Sunday School Union, juveniles, Indian captivities, jestbooks, folklore, anthologies, collections of anecdotes, periodicals, and

extra numbers of periodicals" (1939, vii-viii). Wright's parameters for the bibliography purposefully excluded materials that were published in serial extras, leading to the exclusion of Walt Whitman's *Franklin Evans* (1842) or Edgar Allan Poe's "The Balloon Hoax" (1844) to name canonical exclusions, alongside an untold number of non-canonical works that have gone undescribed. Wright was also interested in listing primarily fiction meant for adults, and not juveniles. This means that some authors have absences in their lists that may seem odd to a human reader, such as Louisa May Alcott, whose *Hospital Sketches* (1863) is listed, but *Little Women* (1868), *Little Men* (1871), and *Jo's Boys* (1886) are absent (Lyle Henry Wright 1957, 7).[3] Wright's inclusion of autobiographical slave narratives such as Harriet Jacobs' *Incidents in the Life of a Slave Girl* (1861) and Solomon Northup's *Twelve Years a Slave* (1853) presents one of the more standout errors that denotes these works as fictional and actively works against the aims of those titles and may cause a modern reader to view *American Fiction* with some skepticism as to its accuracy (Lyle Henry Wright 1957, 179, 242). The previous errors are compounded by the fact that other works that are fictional by black authors, such as Martin Delany's *Blake, or the Huts of American* (1859-1861) are absent in a seeming oversight. All of these individual cases reflect on the dataset as a whole, asserting the conditions of its creation and the validity as a comprehensive source of American fiction titles. Each of these individual cases reflects an interpretive stance towards these items and posits an argument about them as to their apparent necessity in being recorded in a list.

Wright's project did not start with the scope it ultimately encapsulated, and the first two volumes had a few subsequently revised editions. This sort of narrative is not unique, but is often not discussed and theorized when it comes to a reference work such as a bibliography. While it may be obvious, let me state explicitly, that a work such as Wright's underwent a process of composition that involved multiple decisions about what to list, what counts as American fiction, and what was permissible to exclude. This process was iterative and required Wright to reach out for advice, to correct his information in the face of new evidence, and to change between volumes his assumptions about what was feasible to accomplish. Between the first and second volume (1774-1850 and 1851-1876), Wright changed his mind about listing multiple editions of the same title. In the first volume, multiple editions of works are listed in an attempt to provide an overview of the publication history of the work in America within the time frame covered by

---

3 This particular criterion is not always enforced, however, as Mark Twain's *Tom Sawyer* and *Huckleberry Finn* are both listed.(Lyle Henry Wright 1966. 109)

the bibliography, but with the second volume, Wright decided to only list multiple editions of a title if there were "significant differences" (a statement that itself is subjective).

Such decisions are the result of necessary quandaries that Wright and other bibliographers encounter in their work. The process of distilling down the complex worlds of publishing, cataloging, and genre that converge in the case of *American Fiction* was a necessary one in making the task feasible for Wright and navigable for readers. This process, however, is also informed bythe ideologies of the one making the list and the places from which they receive their information. Those who take and adapt the data constructed by others may miss or pay no mind to these decisions, and in doing so introduce the possibility that their reproductions of the data alter the originals, changing the context, purpose, and affordance of that data in order to suit their own agenda for how it should be used.

As a first step in understanding how the editing and adapting of Wright's work to other forms present interpretive judgments about texts, it is necessary to understand Wright's own approach to his work. Identifying the framework he employed in compiling his multiple volumes of American fiction helps to understand some of the implicit arguments a work such as *American Fiction* presents to its readers. Wright was a full-time employee at the Huntington Library for most of his career; he did publish a few pieces of scholarship that were informed by his research and production of *American Fiction.* His first essay, "A Statistical Survey of American Fiction, 1774-1850" displays a few key ideas the undergird the first volume of *American Fiction's* argument. The first of these ideas is the concern for contemporary taste that informs the population of his bibliography, rather than the designation of status and cultural value. As he begins in one section of his analysis: "The writings of many of the forgotten authors, true enough, may not be literary masterpieces, but the point so often overlooked is the contemporary taste for such literature" (Lyle H. Wright 1939. 312). Key early American writers, in Wright's terms, are Timothy Shay Arthur and Joseph Holt Ingraham, both of whom published prolifically between 1830 and 1850, with Arthur publishing 50 works in that time and Ingraham producing 79 works. These are not canonical authors, but authors who were successful in the marketplace, in producing, publishing, and seemingly in selling. What Wright recognizes is that their ability to fill out his list so substantially is a byproduct of that success and not their literary value.

The question of literary value is antithetical to *American Fiction,* and this stance dovetails into the second of Wright's ideals: the ability for a list of works to potentially (though not realistically) include everything. In his essays, Wright demonstrates concern for how those he

calls "literary historians" treat the wider world of American publishing beyond the canon. At the conclusion of his discussion of Arthur and Ingraham, Wright asks, "How many literary histories even mention their names?" (Lyle H. Wright 1939, 312). This sort of subtle jab at literary scholars is characteristic of Wright in his other writings. In an essay derived from his work on the second volume of *American Fiction,* Wright states, "Literary historians will say, I am sure, that some of these titles were better forgotten, but that is a bibliographical impossibility" (Lyle H. Wright 1955, 75). This statement erects a dividing line between Wright's conception of literary scholarship and bibliography. Bibliography is meant to compile and bear witness to all the print publishing that it can; to Wright, bibliography is perfectly egalitarian with its ideal of proclaiming everything worthy of being listed. Literary scholarship, on the other hand, is exclusionary.

One of the arguments we might then assume about the three volumes of *American Fiction* is in the ability for the list to testify to the abundance of American fiction in any capacity, rather than just in terms of the aesthetic and cultural value of the items listed. This appeals to the natural affordance of bibliographies and why they exist in the first place, because they provide easy reference to a large amount of information. But we can also view this as an ideological principal as to how a reader may approach the subject of American fiction. Wright's view of the topic of American fiction is largely more aware of market success, contemporary taste, and those who numerically contributed greater to the total sum of the American literary tradition, rather than to the canonical figures whose presence fills out Wright's bibliography, but is dwarfed by that of other authors and texts. The information that *American Fiction* provides, as Wright implies, is useful not for the possible literary merit that could be discovered, but because of what it will provide a means to access holistic knowledge of American culture:

> From these tales of varying degree of literary merit, and I do not consider all of them literary outcasts, a great deal can be learned about the way of life of the people, the clothes they wore, the food they ate, and their daily gossip. (77)

The culture of early America as displayed by its fiction is not necessarily congruent with the way literary history has been constructed by scholars, but a reference work that makes more accessible the ability to see beyond the mainstay titles of American literature can help to alleviate that problem. Wright's position as the compiler of American fiction titles places him outside of the position of the literary historian who closely reads a small subset of those titles.

As a reference work, Wright's bibliography takes it upon itself to not only list works and testify, via their descriptions, with the data present as evidence that these works exist, but a significant part of the descriptions Wright includes is their locations in the libraries Wright surveyed and visited while he was researching and compiling the volumes. In all three volumes, each description is appended with codes of libraries where Wright was able to physically locate the texts. The census of where the works are available lends both a provenance and argument as to the work that went into the compilation of the bibliography, but also works to help the reader. For an early to mid-twentieth century work, *American Fiction* was particularly effective in pointing researchers in the direction of the wealth of American fiction that existed, but still endeavored to aid in the access to rare or lesser-known texts. The census reinforces the overall trajectory of Wright's work in terms of its commitment to showing the broadest cultural context of American literary writing possible. Reading the entries in the census itself, however, provides an addendum to the data beyond the innately bibliographical information (i.e. title-page, year of publication, etc.) that speaks to the influence of certain titles in terms of their placement around the country in various academic libraries. Unsurprisingly, a first edition of a title such as *Moby-Dick* (1851) was found in thirteen of the nineteen collections Wright perused. Titles with less or even no cultural and institutional backing can at times appear in only one library.

The issue of access becomes central to the way Wright's work is used and re-imagined in subsequent generations of scholarship and projects that use Wright's work as a foundation. Wright's own work was a significant undertaking that has yet to be replicated or redone, even with the expansion of multiple libraries and the arrival of digital repositories and databases that might turn up new discoveries that fit Wright's definition of "American Fiction." Rather than a vertical expansion of Wright's list to expand the number of entries of titles between 1776 and 1900, the transmission of Wright's work instead relied on a more horizontal expansion, taking the titles and furthering the data and information associated with those already listed by Wright. The most obvious and primary way of doing this was to make more accessible the texts of the titles Wright enumerated, particularly by finding, scanning, and marketing a collection of images of the texts via microfilm in the 60s and 70s, and then in digital form by the twenty-first century. When the data that Wright had presented to the world moved from his hands into the hands of others interested in the possibilities *American Fiction* presented, Wright's data became the subject of inevitable changes inherent in the process of transmission.

Soon after its publication, *American Fiction* was being used to facilitate the process of creating research materials for scholars. As a bibliography, *American Fiction* was a reference tool that attested to the existence of titles, but did not help beyond that. Libraries and archives, such as the *American Antiquarian Society,* would use Wright as an aid in assembling materials and expanding their collections, but a formidable wall still existed that hindered access to some of the rarest materials Wright listed. A company known as Research Publications undertook the task of creating a microfilm collection of every title Wright listed. [4] By 1974, the task had been completed, and the set of microfilm that provided facsimiles of 10,827 titles was available for purchase for institutions interested in having the texts listed by Wright for considerably less effort than acquiring physical copies of every text. In addition to the microfilms, Research Publications published a cumulative index of the texts, "in keeping with [their] policy of providing the best possible bibliographic tools for use" (Lyle Henry Wright 1974, n.p.).

The Research Publications' project marks a significant textual moment for *American Fiction*: a moment of transmission and interpretation into a different format that added to *American Fiction's* own presence and influence, as well as to the history of American fiction in library settings. The packaging of the entirety of the unique titles listed by Wright into a single product was meant to drastically improve a library's American fiction holdings upon acquisition of these microfilms. Especially for libraries that could not obtain first editions or copies of rare titles. The process was not without its own decisions that affected the way one would approach Wright's titles. The project was not a complete replication of the bibliography in itself. The focus was on the facsimiles rather than the descriptions. Any publication information available upon viewing the scanned page images that related to the bibliographic data was still available, but any information that required a critical research beyond what was overtly visible (i.e. the census information, related publications, reprintings, etc.) was not included. The index, understandably, presented only the slimmest information in reference to the titles. The below text is a representation of the index entry for Susanna Rowson's *Charlotte Temple* (1794):

> ROWSON, Susanna (Haswell). Charlotte. 2d ed. Philadelphia, Printed for M. Carey, 1794. I.2159, Reel R-5 (318)

The codes at the end of the index reference represent, first, the Wright volume number and entry number: thus, *Charlotte Temple* is to be found in the first volume, as entry number 2159.

---

4 Research Publications is now known as Primary Source Media and operates as a branch of Gale Cengage.

This assumes the 2nd edition of Wright I, since the first was not enumerated. The second code refers to the microfilm collection's own reel. Any additional bibliographical information helpfully points to the exact place in Wright where a scholar could find it, and helps to locate the place in the microfilm for the scholar to find the text, but it is not a full description. This is not the chief point of interest in the microfilm's collection of texts versus Wright's bibliography, however. The instance of *Charlotte Temple* demonstrates an important condition of the microfilm collection's creation. The index only lists one instance of *Charlotte Temple,* and the microfilm contains only one facsimile for the text. That is, the second edition as it clearly denotes in the index. Wright I, however, enumerates 82 different editions of the *Charlotte Temple* published in America alone, all with unique bibliographical descriptions (1948, 232-8). The Research Publications' microfilm, however, omits all other entries in favor of simply supplying one text for reference. This choice suggests a reasoning that helps understand what the goals and aims of the microfilm collection are in comparison with the Wright bibliography. The numerous editions of *Charlotte Temple* are of interest to collectors, librarians, editors, and bibliographers interested in knowing the publication history of the text, but the literary critic may not care or know to care about the importance. Instead, one copy of the text should suffice, lest it suddenly become an overbearing labor to wade through over a hundred different scans of the same novel. The cost of such work would also present an appealing incentive to cut the idea of scanning entries with multiple editions listed.[5]

This sort of change demonstrates where the most important additions to Wright's work could be made, and whom the work was intended to help. Those interested in comparing editions of *Charlotte Temple* gain little, unless they had something other than the second edition available already and sought out the microfilm collection. The microfilm itself, while although alluding to Wright, erases the work he (and others) have done in compiling the list, as well as hides the expansive publication history of the novel, barring access to one pivotal piece of information and the culture in which the text of *Charlotte Temple* is circulating-a culture which could not get enough of the novel for decades-while simultaneously attempting to enable access to the text in order to provide researchers and readers with a firsthand artifact that observed that same culture.

Moving from microfilm to the emergence of the web and digital technologies, Wright's work saw continued use when Indiana University introduced the *Wright American Fiction Project.*

---

5 *Charlotte Temple* is by far not the only one here. Other notable entries would include James Fenimore Cooper's various novels, which also have quite a few entries per novel. Other authors are not as widely reprinted as these two and may only have an additional one or two editions that Wright also lists.

The project gives scholars a repository of both .pdf facsimiles of the Wright titles found in volume two (1851-1875), as well as some XML-encoded texts, alongside some of the original metadata Wright initially described. The project continues with some of the same ideals that Research Publication's microfilm collection did, though updating the concept to fit with both modern scholars' expectations and the affordances of the digital environment. More so than the microfilm copies of the texts presented, the *Wright American Fiction Project* is a task of editing, and demonstrates an interpretation of the items that compose the project as a whole.

The digital project was pushed with only the second volume of Wright viewable. While the other two volumes are labeled as eventual goals, the project is seemingly dormant, with no major updates since it released the digital texts of every Wright II title. The choice of Wright II is significant as a first choice for providing digital records because it coincides with literary movements and historical moments that emerged after Wright's time. Wright II encapsulates F.O. Matthiessen's "American Renaissance," the publication of *Uncle Tom's Cabin* and its imitators and contractors, westward expansion and imperialism, and of course the Civil War. On the subject of F. O. Matthiessen, Wright published Wright II in 1949, 8 years after *American Renaissance* was published (1841), but Wright does not make any explicit mention of Matthiessen in *American Fiction,* so it is indeterminate whether or not he was aware of or familiar with the Matthiessen's arguments. Given that Wright II follows naturally from Wright I, beginning in 1851, and ending in 1875, a date not particularly relevant to Matthiessen's thesis, the overlap is certainly coincidental from a composition perspective, but nonetheless significant from the perspective of readers and researchers. To prioritize Wright II is to prioritize the likes of Melville, Stowe, and Twain; the project is not shy about this as the home page shows portraits of the aforementioned authors alongside Bret Harte, William Dean Howells, and Nathaniel Hawthorne that are hyperlinked to search results of these authors' works. The project itself understands and anticipates who the most prominent members of its corpus are and what texts will be the biggest hits, so to speak. This is almost in direct conflict with Wright's ideal of the egalitarianism of bibliography. The issue is only complicated when one looks at state of more popular and canonical texts when compared to those that are virtually unknown.

The *Wright American Fiction Project* provides XML-encoded files of the texts enumerated by Wright II in addition to .pdf and plain-text formats. Details about the status of the collection's encoding can be found in the site's "Encoding Overview," which reveals a division in the way certain texts were treated:

> The online collection of nearly 3,000 volumes consists of two different groups of texts. The larger group of approximately 1,800 electronic texts was created by Prime Recognition Optical Character Recognition (OCR) software. These texts are minimally encoded and largely unedited, and rely on the facsimile page images as the main access point. The other group of approximately 1,200 texts has been fully edited and encoded, and also includes facsimile page images. In addition to being corrected, these files allow for better document-centric navigation by identifying chapter or story divisions within each work and having a hypertext linked "Table of Contents." Both groups of texts are available for bibliographic and full-text searching as well as browsing. (2018)[6]

The choices of which texts to include in the "largely unedited" group and in the group of edited and encoded is not entirely arbitrary. The selection process for the more robustly edited and encoded texts is largely opaque to the viewer, but some of the choices are obvious as to why certain texts were chosen for advanced encoding. Again, familiar names come up as to which texts are beneficiaries of the resources and labor required for electronic editing.

Let us compare Moby-Dick to the text that is found directly after it in American Fiction, Matthew Merchant's *How Bennie Did It* (1869). For this comparison, I would like to begin with the text of Merchant's work and the associated XML the *Wright American Fiction Project* presents to readers who may encounter this text. As the project's "Encoding Overview" states, the digital text was created via OCR, rather than transcription. This process, depending on various factors including how the software was trained, the quality of the print scanned, among other contingencies, can naturally produce errors in the text files that are created. *How Bennie Did It* is no exception. Below is the opening paragraph of the body of the text and its XML encoding:

```
<pb n="0" xml:id="VAC8367-00000005"/>   HOW BENNIE DID IT.
CHAPTER L IT'S a singular story,     think you will say,
before having rea it through;. Singular, however, though, it
may be, we hope there are none who may read it but will do so
with both pleasure and profit. The BENNIE STOUT Of the story
was a youth, the record of whose life, peculiarly interesting
and eventful as it is, might well be accepted. as more of a
study than a story: for, connectedy with, or in fact giving
rise to, the very features of his history which will
doubtless interest us most, there is something deeper than
story, and more earnest than entertaing. ment. It will not
only be the youthful reader,  ho will stop, and wonder, and
```

6 In addition to these two collections, the encoding has changed, from the original SGML to the various iterations of the TEI guidelines P3 through P5 in order to keep in it in line with standards and best practices at the time.

```
probably aask met   twl  o  nyb teyuhu    edr   &#xD;<pb
n="8-9" xml:id="VAC8367-00000006"/>
```
(2018)

What becomes apparent from this section is how minimal the "minimal encoding" is. This portion of text is only marked up to refer to a page break, i.e. the <pb> tag which simultaneously refers to the page image that is associated with the text. The rest of the text is largely unmarked, leading to an untidy portion of text that does not delineate between the body of the text and its paratext. The all-caps "HOW BENNIE DID IT" is seen after every <pb> tag in the XML, indicating the running head of the page being scanned and picked up by the OCR, but its physical location on the page and the meaning inherent to that being lost. A similar phenomena occurs with the "CHAPTER L," which is both an OCR error produced by reading an I with a full stop as an L, and a phrase that has been dissociated from its bibliographic function as what demarcates separate textual sections.[7] The end of the quoted passage additionally shows some of the difficulties that come with OCR. The final words of the page itself are "probably ask" before it continues the sentence on the next page. However, the OCR, and thus the XML file and text visible on the project's site include the unintelligible series of characters "met twl o nyb teyuhu edr." No words follow the "probably ask" of the page image, and so the machine introduced text that is not present in its source, and the project has pushed this text to the public with these additions.

How Bennie Did It has not been untouched since it was first encoded, but it has not seen the level of revision or editing that a work such as *Moby-Dick* has. Viewers of the *Wright American Fiction Project* would find *Moby-Dick* to be presented with a document that has its table of contents fully linked to the body of the text, the initial "Etymologies" section is formatted as a table and presented in an organized fashion that resembles the same way it is presented in print, and the body of the work is arranged as much as one would expect a novel to be in XML. But of course what makes *Moby-Dick* a prime choice for comparison here is not just its canonicity, but the way it shifts into other modes of presentation periodically, such as the "Midnight, Forecastle" chapter, wherein the text is set as if it were a stage play, or the "Cetology" chapter that replicates a genealogical catalog. These moments of the text reveal the effort required and given to this particular text in order to present a satisfactory digital edition of the work. In the "Midnight, Forecastle" chapter, the initial lines of the dramatic performance belong to the 1st Nantucket Sailor, whose words are not only presented as dialogue but also contain quoted verse.

---

7 XML and the TEI guidelines specifically make space for these textual characteristics to be easily marked. Chapter headings and running heads for pages are both standard tags included in TEI's P5 guidelines.

```
      <sp>
        <speaker>1ST NANTUCKET SAILOR.</speaker>
        <p>Oh, boys, don't be sentimental; it's bad for the
digestion! Take a tonic,follow me! <stage>(Sings,
            and all follow.)</stage>
          <q>
            <lg type="quotation">
              <l>Our captain stood upon the deck,</l>
              <l rend="ti-1">A spy-glass in his hand,</l>
              <l>A viewing of those gallant whales</l>
              <l rend="ti-1">That blew at every strand.</l>
              <l>Oh, your tubs in your boats, my boys,</l>
              <l rend="ti-1">And by your braces stand,</l>
              <l>And we'll have one of those fine whales,</l>
              <l rend="ti-1">Hand, boys, over hand!</l>
              <l>So, be cheery, my lads! may your hearts never fail!</l>
              <l>While the bold harpooneer is striking the whale!</l>
            </lg>
          </q>
        </p>
      </sp>
```
(Melville 2018)

Compared to the encoding of *How Bennie Did It,* the words of the 1st Nantucket Sailor show more than just a minimal level of encoding, and in fact close attention to reading the text. This level of encoding goes beyond presenting just a page break, but the representation of lines, line groups, and the speaker, with both the text that signifies the subject and the words spoken marked with their own distinct tags. The text of *Moby-Dick* therefore is much cleaner compared to *How Bennie Did It.* The readability of the text is improved both by the advanced markup applied to *Moby-Dick* as well as the proofreading the XML file reveals the text to have undergone.[8] This more significant level of attention given to the text of *Moby-Dick* marks the difference between the two groups of texts that the *Wright American Fiction Project* describes, and thus informs us of how the approaches to Wright appear to differ from Wright's vision.

While the *Wright American Fiction Project* uses Wright's initial bibliographic work to populate its database, the extended services and means of accessing those documents troubles some of the initial ideals Wright had about the literary documents that composed his bibliography. The perspective that all the materials enumerated by a bibliography are equal by the virtue of their having been listed in the first place, becomes more troubling when we begin to see how the steps

---

8 Specifically, the XML claims as a change that occurred July 31, 2003: "Finished final proofreading." This was done by Maggie Hermes. The XML file of *How Bennie Did It* does not include a similar note.

beyond listing come into play, either through the limitation of resources or the decisions of the editor/s. For both Research Publications and the *Wright American Fiction Project,* the major move these two initiatives made in expanding Wright was in connecting the bibliographic data to the material texts that data pointed to and in providing the body of the text itself, that which made the texts American fiction to be listed in the first place. In neither case was this done without affecting the titles and materials of Wright's work and thus reconstituting the Wright bibliography as a holistic text and concept. In seeking to provide access to the texts of Wright's titles, both projects made editorial decisions to either limit or restrict some texts, whether it be the multiple editions of *Charlotte Temple* or other reprinted works, or the amount of effort put into editing and presented particular texts in digital reproductions.

The examples I have discussed with regards to both Research Publications and the *Wright American Fiction Project* seem to depart from Wright's initial egalitarian concept that guided his composition of *American Fiction.* Both projects privileged access to titles, though they did not necessarily guarantee equal access, as some entries by Wright are excluded because of perceived redundancy, or are worth less editorial effort because of their status with regards to literary history. While these projects are not afraid to approach the *American Fiction* bibliographies differently from Wright, they do carry forward some of Wright's errors. As I discussed before, some of Wright's imperfections, such as the inclusion of the nonfictional *Incidents in the Life of a Slave Girl* and *Twelve Years a Slave* are noteworthy because of how they stand out against the simple premise of *American Fiction:* that the works listed are fictional. Neither the microfilm nor the digital collections challenge these notions; the digital copy of *Incidents* enables access to the first edition of the text, and attempts to represent the metadata associated with it by Wright.[9]

It is in a situation such as this, however, that an editor may more concretely see how they could intervene in the construction of a collection of American fiction titles. If one were combing through Wright's *American Fiction* (or projects based on Wright) in order to update or revise its data, they would be met with two options for a text such as *Incidents:* to either remove the text, so that *American Fiction* remains accurate as a whole, or append extra information to *Incidents* so that the information is still present but qualified. I believe the latter is the more useful suggestion, as the first would result in consequences for the text of *Incidents* and limit knowledge and

---

9 I use the word "attempts" here because of the error the *Wright American Fiction* project makes in describing *Incidents.* While Wright correctly indicates the author of *Incidents in the Life of a Slave Girl* as Harriet Jacobs, the *Wright American Fiction* project indicates the author as Linda Brent, the pseudonym Jacobs uses in the narrative. The presence of *Incidents* in *American Fiction* notwithstanding, this error demonstrates some departure from Wright for the information about the texts he lists, even though he was correct in the case of the author's name.

therefore access to Jacobs' narrative.[10] Keeping *Incidents,* however, does perpetuate the problem of accuracy, yet it ensures another means of retrieving and recognizing the text. The metadata associated with *Incidents* in both microfilm and digital adaptations of Wright are the bibliographical information Wright records, but more could be said to show how scholarly knowledge (bibliographical, historical, and literary) has expanded since Wright. More information could be applied to note its status within the corpus; information that could notify readers as to its autobiographical nature, allowing those consuming the data to both obtain the information related to the text, but also to know that its inclusion is problematic in nature.

Scholars will continue to make use of data for research as big data methods and questions become more commonplace. But just as the formalization of literary scholarship necessitated critical editions of works, there is an occasion for inquiry and interest for those who are already familiar with the ways in which data are constructed and presented. Because the case of Wright shows how aggregate collections of textual materials have a history that extends back beyond the emergence of digital collections, and touches into areas textual criticism knows well, scholars can take a critical look at the ways in which a collection may be changed, expanded upon, altered, or corrected. These tasks would serve to make the data accessed by other scholars more replete in information while also taking into account the idea of collections as substantive textual objects themselves. At the same time, knowing that the construction of data is itself similar to if not heavily informed by the methods of bibliography, what I have argued is not entirely new but a means of engaging with our intellectual ancestors. Those such as Greg, Pollard, etc. (and I would include Wright himself) whose systematic modes of enumeration and description have impacted the work of big data in the humanities because they formalized the disseminated standards of bibliographical description and critical editing. Because data in the humanities has been gathered through a process of retrieval from library catalogs, collections, and bibliographies, textual scholars and editors are equipped to deal with the byproducts of data, because they are also familiar with the ideas that originally, to use Drucker's wording, "parameterized" the data in the first place.

Texas A&M University

Works Cited

10 Admittedly, a text such as *Incidents* would likely survive due to the recovery efforts of the past few decades in literary scholarship and education, but this rule would hold for other texts in similar situations.

GAILEY, Amanda. 2012. "Editing in the Age of Automation". *Texas Studies in Literature and Language* 3: 340-56.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. 2002. *Reference Model for an Open Archival Information System (OAIS)*. Washington DC: CCSDS Secretariat. Accessed May 12, 2017. https://public.ccsds.org/pubs/650x0m2.pdf.

DRUCKER, Johanna. 2011. "Humanities Approaches to Graphical Display". *Digital Humanities Quarterly* 5.1. Accessed February 17, 2017. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

---.2014. *Graphesis: Visual Forms of Knowledge Production*. MetaLABprojects. Cambridge, Massachusetts: Harvard University Press.

2018. "Encoding Overview". *Wright American Fiction*. Accessed August 13. http://webapp1.dlib.indiana.edu/TEigeneral/projectinfo/encoding.do.

MELVILLE, Herman. 2018. *Moby Dick, or, The Whale*. *Wright American Fiction*. Accessed August 20, 2018. http://webapp1.dlib.indiana.edu/Teigeneral/viewdocid=wright/VAC7237.xml&doc.view=print.

KIRSCHENBAUM, Matthew. 2013. "The.txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary". *Digital Humanities Quarterly* 7.1 (July 1). Accessed February 17, 2017.

MERCHANT, Matthew. 2018. *How Bennie Did It. Wright American Fiction*. Accessed August 20, 2018. http://webapp1.dlib.indiana.edu/TEigeneral/view?docid=wright/VAC8367. xml.

ALLISON, Sarah. 2016. "Other People's Data: Humanities Edition". *Journal of Cultural Analytics*. December 8. Accessed January 1, 2017. http://culturalanalytics.org/2016/12/other-peoples-data-humanities-edition/.

SHILLINGSBURG, Peter L. 1996. *Scholarly Editing in the Computer Age: Theory and Practice.* 3rd ed. Editorial Theory and Literary Criticism. Ann Arbor, MI: University of Michigan Press.

WRIGHT, Lyle H. 1939. "A Statistical Survey of American Fiction, 1774-1850". *Huntington Library Quarterly* 2.3: 309-18.

---.1939. *American Fiction, 1774-1850; a Contribution Toward a Bibliography*. San Marino, CA: Huntington Library Publications.

---.1948. *American Fiction, 1774-1850: A Contribution Toward A Bibliography*. First Revised Edition. Huntington Library Publications. San Marino, CA: Huntington Library.

---.1955. "A Few Observations On American Fiction, 1851-1875". *Proceedings of the American Antiquarian Society* 65.1: 75-104.

---.1957. *American Fiction, 1851-1875: A Contribution Toward A Bibliography*. San Marino, CA: Huntington Library Publications.

---.1966. *American Fiction, 1876-1900; A Contribution Toward A Bibliography*. San Marino, CA: Huntington Library Publications.

---.1974. *American Fiction, 1774-1900: Cumulative Author Index to the Microfilm Collection*. New Haven, CT: Research Publications.