Curatorial Statement

The Index of Major Literary Prizes in the US includes two related datasets.

The first is a dataset of the winners and judges of prizes for prose, poetry, or unspecified genre between 1918 and 2020 with a purse of $10,000 and over. The data was collected by hand mainly from institutional websites. Gender and higher education data for individuals was collected from author biographies, interviews, and other materials. Some information about judges not listed on websites was obtained through correspondence with institutions.

The dataset includes details about the winners of fifty-two unique prizes awarded by twenty-two institutions. For a subset of thirty-nine prizes, it includes details about judges; not every prize has complete judge data. It does not include prizes awarded specifically for children's literature, nonfiction, drama, or translation.

Claire Grossman, Juliana Spahr, and Stephanie Young are the principal investigators, did the majority of the data gathering, and are responsible for any errors. They were assisted by Jennifer Chukwu, Clare Lilliston, Jordan Pruett, Esther Vinarov, and Betty He. Richard Jean So provided significant support for this project.

Please report any errors and/or corrections here.

Filename: winnersandjudges.csv (7134 rows; 18 columns)

- person_id: unique numeric identifier for each name; assigned alphabetically by first name
- full_name: pen names were used; in case of name change, most recent name was used
- given_name: first name; includes middle name, if used
- last_name: last name
- gender: provisionally labeled by research team based on pronouns used by author in biographical notes at the time research was completed; it is possible a judge/winner's gender identity and/or pronoun may have changed subsequently; intended for study of broad patterns over time and not as definitive statements on any individual identity; values are "male," "female," "nonbinary/he," "nonbinary/they," "unknown," and "No Winner"; nonbinary was used only when the term appeared in the individuals' biography.
- elite_institution: individual mentioned they attended (even if they did not graduate from) one of the listed institutions; intended for study of broad patterns over time and not as definitive; values are "Barnard College," "Brown University," "Columbia University," "Cornell University," "Dartmouth College," "Harvard University," "Princeton

University," "Radcliffe College," "Stanford University," "University of Pennsylvania," "University of Chicago," "Yale University," "No Winner," or blank (means unlikely as individual listed higher education affiliations in biographical notes but did not include an elite institution or unable to locate any educational information about the individual); intended for study of broad patterns over time but not as definitive.

- graduate_degree: individual mentioned they attended (even if they did not graduate from) a graduate program (includes masters, PhD, JD, and medical degrees); values are "graduate," "No Winner," or blank (means unlikely as individual listed higher education affiliations in biographical notes but did not include a graduate degree or unable to locate any educational information about the individual); intended for study of broad patterns over time but not as definitive.
- mfa_degree: individual mentioned they attended (even if they did not graduate from) an MFA program; values are name of institution, "No Winner," or blank (means unlikely as individual listed higher education affiliations in biographical notes but did not include an MFA or unable to locate any educational information about the individual); intended for study of broad patterns over time and not as definitive.
- iowa_mfa_person_id: values are either a number that corresponds to the Post45 Iowa Writers' Workshop "People" table, "missing" (means that the individual's biographical materials suggest they attended Iowa for an MFA but a corresponding entry could not be found in the Iowa dataset which ends in 2014 and does not include graduates of the MFA in playwriting), "unknown" (unable to locate any educational information about the individual), "No Winner," or blank (means that the individual did not list University of Iowa in their biographical notes or unable to locate any educational information about the individual)
- stegner: individual mentioned they were awarded a Wallace Stegner Fellowship at Stanford; the Stegner program does not award degrees but it resembles an MFA program in pedagogy except it is not unusual for those admitted to already have an MFA; we thus treat it as the equivalent of an MFA (and not a prize); values are either "Stegner," "No Winner," or blank (means that the individual did not mention the Stegner Fellowship in their biographical notes or unable to locate any educational information about the individual)
- role: values are "winner" or "judge"
- prize_institution: nonprofit organization that oversees the prize
- prize_name: name of prize; for the Gold Medal Awards from the American Academy of Arts and Letters, we only included awards categorized as fiction and poetry; for the Morton Dauwen Zabel Award from American Academy of Arts and Letters, we excluded periodic awards given specifically for "Criticism"; for the National Book Award, we only

included prizes for poetry and fiction; for the Academy of American Poets, we only included the Academy of American Poets Fellowship, the Lenore Marshall Poetry Prize, and the Wallace Stevens Award; for the Poet Laureate Consultant in Poetry to the Library of Congress, we included the US Consultants in Poetry but did not include the three Special Bicentennial Consultants that served in an advisory role from 1999-2000 and excluded William Carlos Williams (who was named as Laureate, but did not serve); for the Pulitzer Prize, we only included prizes for fiction and poetry; for the MacArthur Fellowships, we included those who were categorized by the MacArthur website as "poetry" and most of those categorized as "fiction and nonfiction" (if a writer exclusively published journalistic nonfiction or essay, they were not included).

- prize_year: year awarded; in the case of the Poet Laureate Consultant in Poetry to the Library of Congress, which begins in September and continues until May, we included entries for the Laurate under both years
- prize_genre: values are "poetry," "prose" ("prose" includes prizes for "short stories," "essays," "fiction," and "novel"), and "no genre" (prize has no genre requirement, as in the MacArthur Fellowship or the Whiting Award)
- prize_type: values are "career" (prize is awarded to author on basis of overall career) or "book" (prize is awarded to author for a specific book)
- prize_amount: value here is the amount of money awarded in 2022; amounts change over time, which we do not track
- title_of_winning_book: if "prize_type" is "book," then the awarded book title is listed (if the jury awarded more than one book in same year, titles for both are listed); other values are "No Winner," and blank (prize was not awarded for a specific book)

The other dataset was assembled by Jordan Pruett.

Filename: hathitrust_prizewinners.csv (5569 rows; 11 columns)

The *hathitrust_prizewinners.csv* dataset contains records for volumes in the HathiTrust Digital Library written by authors who won a prize in the prize winners dataset. It includes many duplicate titles, as in the case of later editions of the same work.

Matches were produced by performing an exact string comparison between the *last_name* and *given_name* columns of the prize dataset to the *author* column of the HathiTrust dataset held by the Post45 Data Collective. Fields were forced to lowercase and stripped of punctuation and spacing before comparison. This conservative matching process is likely to produce two kinds of

errors: missed matches, in the case of authors who appear under different names in the two spreadsheets; and false positive matches, in the rare case of two authors with identical first and last names. The first type of error was considered acceptable: the spreadsheet aims to maximize true positive matches rather than minimize false negatives. Since matches were assessed restrictively, researchers can be confident that the vast majority of the entries in this dataset were in fact authored by somebody who also appears in the prize dataset. In order to estimate the rate of the second, more problematic type of error, a random sample of 100 entries was taken from the final spreadsheet and checked manually for accuracy. This sample contained no errors, though it did contain one match that could not be verified, since no secondary literature could be located for the author in question.

Finally, it is worth noting that *hathitrust_prizewinners.csv* does not distinguish between the types of prizes won by authors nor the point in their careers that those authors won those prizes. For each author in the prize dataset, it simply lists every HathiTrust volume authored by that author that could be located.

- hathi_id: HathiTrust item identifier number
- shorttitle: the short title of the work as listed in HathiTrust
- prize: name of prize; either NBA (National Book Award) or pulitzer
- author: name of the author of the award-winning work
- person_id: unique numeric identifier for each name; assigned alphabetically by first name
- inferreddate: earliest publication date for this particular volume
- imprintdate: the date of this edition of the text
- oclc: a unique identifier for this volume as registered in WorldCat
- full_name: pen names were used; in case of name change, most recent name was used
- given_name: first name; includes middle name, if used
- last_name: last name
- gender: provisionally labeled by research team based on pronouns used by author in biographical notes at the time research was completed; it is possible a judge/winner's gender identity and/or pronoun may have changed subsequently; intended for study of broad patterns over time and not as definitive statements on any individual identity; values are "male," "female," "nonbinary/he," "nonbinary/they," "unknown," and "No Winner"; nonbinary was used only when the term appeared in the individuals' biography.