

Development of HBW's BBIP Metadata Creation Manual

White Paper for Award HAA-281030-21
History of Black Writing (HBW)
Building Literacy and Curating (Critical Cultural) Knowledge in
Black Humanities (BLACK DH)
National Endowment for the Humanities: Office of Digital Humanities
Digital Humanities Advancement Grants

December 2024

Authors: Erin Wolfe, Maryemma Graham, Ayesha K. Hardison

Project Summary

The Black Book Interactive Project (BBIP) is the digital component of the History of Black Writing (HBW), which aims to increase the availability of Black-authored texts for scholarly engagement and teaching. By developing (1) a comprehensive corpus of these texts and (2) a metadata schema that accounts for the complexities of race and race-related issues in the context of these works, BBIP addresses a critical need to document and analyze Black literature—a body of work that has historically been underrepresented in literary scholarship as well as within contemporary digital humanities.

Primary Collaborators:

Erin Wolfe, University of Kansas

Maryemma Graham, University of Kansas, History of Black Writing

Ayesha Hardison, Indiana University Bloomington, History of Black Writing

Project Origins and Goals

Black authors and their contributions have often been marginalized or excluded in mainstream literary canons, which leads to an incomplete understanding of their impact on literary traditions and cultural history. This exclusion becomes notably more pronounced in digitally accessible sources and, as a result, digital scholarship. This lack of representation and availability not only limits scholarly research but also hinders broader access to these works by educators, students, and the public.

BBIP's significance lies in its capacity to bridge these gaps by identifying underrepresented authors and creating structured metadata that serves as the foundation for providing a nuanced record of these works. Metadata is crucial for ensuring that these works are discoverable and accessible for scholarly activity. It allows researchers to locate and analyze materials, providing the basis to engage with the texts on an individual basis and to look for trends. Furthermore, metadata supports the integration of these works into larger datasets to enable comparative studies and interdisciplinary research.

In alignment with the mission of HBW, BBIP not only contributes to the preservation of these works but also empowers researchers to explore new perspectives on Black literature, fostering a deeper understanding of the cultural and historical contexts in which these works were created. Through this work, BBIP helps to ensure that Black literature receives the recognition and critical attention it deserves. This work will also enrich the broader fields of literary studies and digital humanities by expanding access to this vital aspect of literary heritage.

BBIP has been funded by the University of Kansas, the National Endowment for the Humanities (NEH), the American Council of Learned Societies (ACLS), and the Mellon Foundation. The strength of this support underscores the project's central role in addressing the challenges of working with digital humanities in Black American literature, which has historically been underrepresented in bibliographic and literary studies.

Definitions and Rationale

Metadata refers to structured information that describes and organizes datasets, providing essential context and facilitating the discovery of information. In the context of the Black Book Interactive Project (BBIP), metadata plays a pivotal role in documenting literary works, particularly those authored by Black writers. By systematically capturing key details such as a text's authorship, publication history, thematic elements, and cultural context, BBIP ensures a more accurate representation of Black literature within larger literary and bibliographic frameworks.

The significance of metadata extends far beyond mere organization; it is a foundational element that supports scholarly inquiry and engagement. By making underrepresented works more accessible, metadata fosters cross-disciplinary research and encourages collaboration among scholars from various fields. It also facilitates the integration of Black literary works into broader datasets, allowing for comparative analyses and interdisciplinary studies that can illuminate connections between different literary traditions and cultural narratives. The creation and standardization of metadata for Black literary works is essential for preserving this vital aspect of literary heritage.

Previous Activities

The need for a customized metadata schema has been a recognized key aspect of BBIP since before its inception in 2013, even dating back to a 2011 HBW prototype, "The 100 Novels Project." Designed by former HBW staff member Kenton Ramsby, the prototype examined spatial and temporal data and word frequency. In 2015, BBIP was awarded a National Endowment for the Humanities grant to help fund the development and refinement of a metadata schema that would pay particular attention to the complexities of race and difference as it intersects with cultural, historical, and textual features identified through the scholarship on African American fiction.¹ During this grant, the first draft of an extensive schema was developed by the BBIP staff, with input from Eric Radio, a metadata consultant partner from the University of Kansas (KU) Libraries. In 2017, the HBW team, in consultation with KU Libraries Metadata Librarian Erin Wolfe, standardized the metadata elements and descriptions into a working draft schema. This original schema included 51 elements and was designed to describe a variety of metadata elements in five major areas: author biographical data; genre and textual (language) features; geography (place/setting); subject content and theme; and criticism. [See Appendix A for a list of all elements of the 2017 schema.]

This ambitious schema was tested by HBW staff and students on an initial subset of 75 titles from the larger novel corpus during 2017-2018. Although comprehensive, the requirements to fully meet this schema proved to be too time consuming to apply widely for reasons including:

- Extensive external research, with varying amounts of information available, depending on the popularity of the book and/or author (e.g., author biographical metadata, contemporaneous book reviews, etc.)

¹ Maryemma Graham. 2015-2019. "Black Book Interactive Project." National Endowment for the Humanities. Digital Humanities: Digital Humanities Start-Up Grants. HD-248607-16.

- Knowledge of larger historical literary world (e.g., literary movement, pioneering theme/contribution in Black writing, etc.,)
- Close reading and subjective nature of elements (e.g., presence and nature of supernatural elements, prevalence of vernacular use, etc.)

The complexity of the schema, coupled with external factors such as a change in BBIP's Project Manager and team, as well as challenges stemming from COVID-related shutdowns, led to the effective abandonment of this overly detailed approach. However, it continued to exist within the BBIP administrative goals as an ideal for future development.

In 2021 BBIP Co-Coordinators Arnab Chakraborty and Hamza Rehman, along with other HBW graduate student staff, made a push to reinvigorate the metadata collection process by creating a revised metadata schema. This new schema included 21 elements covering biographical, bibliographic, historical, and contextual information. There was some overlap with the 2017 schema, but many elements were dropped in favor of attempting to place the novel within a larger context of online information sources (Wikipedia, Library of Congress catalog, Google Scholar). [See Appendix B for a list of all elements of the 2021 schema.] The collected data was also cleaned in 2021 to standardize data entry formats; this effort was led by graduate student staff, later BBIP Co-Coordinator, Jade Harrison.

Throughout this time, other BBIP and related HBW projects continued. The BBIP team steadily built the HBW novel corpus through a two-pronged approach entailing title identification and book digitization. Additionally, they attempted to computationally link targeted metadata elements to larger linked datasets, such as Wikidata, VIAF, Library of Congress catalog records, and other sources; regularly produced discrete research projects adjacent to the primary metadata collection; and more. Ultimately, this variety of projects—at times parallel, at others tangential—resulted in the creation of metadata following various (often non-standardized) schema stored in different locations with no effort to bring them together.

The primary goal for this current phase of the project is to create a single location for all HBW novel corpus-related metadata that follows a well-defined and clearly articulated schema. In order to accomplish this larger goal, it requires us to do the following:

- Develop a schema that is comprehensive, practical, and meets the needs of HBW and BBIP users
- Locate all existing datasets where metadata has been created and stored
- Clean and merge existing metadata into a single location
- Create a manual to ensure consistent future metadata creation and entry

Project Activities, Team, and Participants

Current Focus: Standardizing and Refining Metadata Schema

In Summer 2023, HBW Directors Maryemma Graham (MEG) and Ayesha Hardison (AKH) met with KU Libraries Digital Initiatives Librarian Erin Wolfe (EDW) to address the growing decentralization of BBIP metadata. Fourteen discrete metadata sources were identified, including spreadsheets with record counts ranging from 75-6,500+ titles each. Most of these sources were

not actively updated, but a few were ongoing, including the overall Corpus Inventory and the Digitization Tracking Log.²

Element names from each source were collected into a separate spreadsheet, to allow for a high-level cross-document comparison. In all, 280 fields were counted across all sources. After removing duplicate names (e.g., the element “Title” was used in nearly all sources), 166 elements remained. A manual review was conducted to map different element names to describe the same metadata item³ and remove Project Management fields (e.g., digitization tracking). As a result, 90 unique metadata elements were identified that had been used to track metadata across various subsets of the corpus.

To systematize ongoing metadata collection while maximizing previous effort, MEG, AKH, and EDW worked together to evaluate each of these individual elements and identify a list of elements that should be retained and collected in the future. After a lengthy review process, 33 elements were identified to create a new schema (“New Schema”) to replace the others, which formed the basis of the BBIP Metadata Creation Manual. While the New Schema did carry over some of the complexities from the 2017 schema (e.g., external research to locate bibliographic information about the authors) and some of the external linking from the 2021 schema (e.g., Wikipedia entries), efforts were made to (1) narrow the elements to items of highest importance to the project and (2) prioritize elements into distinct groupings that could be completed in separate phases. Specifically, bibliographic and surface-level metadata could be gathered in a first phase by HBW student workers, external research (e.g., identifying awards won by the novel) could be done in a secondary phase, and specialized elements (e.g., topic analysis, historical context) could be done in a separate phase and/or by students with more specialized expertise. [See Appendix C for a list of all elements of the New Schema] In fall 2023, HBW launched the Data Tank with BBIP Co-Coordinator Jade Harrison. Its focus on data collection for the first phase of the New Schema, coordinated by graduate student Abisola Akinsiku, began in 2024. [See Appendix D for a list of all the students who worked on Data Tank in 2023-2024.]

Merging Metadata from Multiple Local Sources

In 2023-2024, EDW undertook a manual review, in combination with custom Python scripts, to isolate and merge metadata that fit into the New Schema. However, the problems of non-standardized metadata quickly revealed themselves. In addition to the variations in element names described previously, the metadata itself was frequently inconsistent. These inconsistencies could come from data entry errors (e.g., typos; varying spacing, capitalization, or punctuation), differences in specific formatting (e.g., formatting the author’s name as “First Last” vs. “Last, First”), or source-specific metadata (e.g., publication information from different editions). Regardless of the source of the differences, this incongruity led to considerable

² The “Corpus Inventory” records the team’s efforts to identify relevant titles, creating a list of books with minimal bibliographic metadata to be expanded in a future phase. The “Digitization Tracking Log” records a separate team’s efforts to obtain physical copies of titles in the corpus, scan them, generate OCR from the resultant PDF files, and add them to the BBIP digital corpus.

³ Some examples of differing element names used for “Author”: “Author”; “Author Name”; “Name as it appears”; “Author First name” / “Author Last name” [two elements]; “Author Name (Last name, first name or pen names)”; “Author Name (last, first)”. While each of these describes the same metadata point, the different field names require manually mapping the metadata.

duplication in the metadata when attempting to merge multiple sources, resulting in the need for a systematic and detailed approach to the process.

In this new approach, EDW started with the two largest metadata sets: Sheet #1 had 3,971 rows, and Sheet #2 had 6,695 rows (10,666 total, with expected duplication between them). On the initial merge of these two lists, there were 7,851 “unique” titles by 3,041 “unique” authors. After many hours of automated and manual cleanup using Open Refine and the pandas library in Python,⁴ the combined list was reduced to 6,602 unique titles by 2,489 unique authors. This revealed that, in addition to the 2,851 exact duplicate title entries between the two sheets, there were 1,255 effective duplicate title entries and 642 effective duplicate author entries that were entered more than once with (usually) small inconsistencies that made automated merging impossible.⁵ Similar complications occurred in other fields, such as publisher information. Although the goal of the metadata collection was to record the publication information from the first edition whenever possible, in many cases, these duplicate title entries were created from different editions with different publishers, dates, etc. In these cases, the earliest date was retained, with the alternate publication recorded in the Notes field to be reviewed in a later phase.

The remaining metadata sheets were smaller with varying numbers of elements to map. Some had only basic bibliographic metadata (title, author, publication date), while others contained over 20 relevant elements. EDW followed similar processes in Python and Open Refine to merge these into a single structured dataset following the New Schema and exported to an Excel workbook for easier review and use (“Workbook”). In its initial form, the Workbook recorded a single row for each title, with any metadata that could be mapped from the original source files.

Several factors influenced the decision to utilize a spreadsheet for recording metadata instead of a relational database like Microsoft Access or SQL. Primarily, the hands-on work of metadata creation is carried out by a rotating team of student workers, necessitating a data source that is intuitive and user-friendly in a collaborative environment. Additionally, HBW had already begun transitioning to Microsoft Teams for document storage and sharing, making an Excel spreadsheet a seamless fit within this established workflow. This approach facilitates ease of access and collaboration and allows for quick adjustments and updates as needed.

Although each of the elements in the New Schema had at least a few entries, many were sparsely populated. For example, although there were 6,602 titles, only 3,169 had a publisher and only 254 had an assigned genre. In addition, there were more than 200 titles that had been scanned more than once (often twice, but three times in some cases) and had duplicate scans with different ID numbers and separate entries in the source files. These duplicate records were merged in the Workbook, with the ID numbers recorded to avoid the appearance of missing items in the future.

⁴ pandas is a popular Python library used for data manipulation and analysis and was used here for merging data and identifying duplicates. Open Refine is an open-source tool commonly used for cleaning data and was used in this project to identify duplicates that couldn't be automatically merged.

⁵ “Exact duplicate” signifies the same entity referenced in multiple places with the exact same formatting, which can easily be computationally matched (e.g., “DuBois, W.E.B.” is identical to “DuBois, W.E.B.”). “Effective duplicate” signifies the same entity referenced in multiple places with different formatting, making computational matching challenging (e.g., “DuBois, W.E.B.” is formatted differently than “W.E.B. DuBois”).

A key takeaway from this entire process was that the lack of standardized metadata creation and a single place to record it resulted in many hours of unnecessary work (i.e., metadata cleanup and merging), duplicated effort (e.g., creating metadata for and/or scanning the same book multiple times), and reduced usefulness of the metadata (i.e., since there was not a single location to check for any given title). This led to the conclusion that in order to create a usable dataset for this corpus, we needed (1) a single location to record metadata for all titles, (2) a clear guide to what metadata to create and how to format individual elements, and (3) regular review of new entries to prevent duplications and inconsistencies from creeping back into the dataset. These goals led to the development of the metadata creation manual (“Manual”).

Project Outcomes

Goal 1: Preparing the Metadata Workbook for Collaboration

Before turning the Workbook over to the Data Tank team that would work to collect the metadata, EDW made three additional changes to its initial form. First, to address the challenge of author-related metadata, which often involves multiple entries for a single author, all author-specific elements were moved to a separate sheet within the Workbook. This resulted in two primary spreadsheets: “Book” and “Author.” Metadata for each author is entered only one time. This ensures a single central location for adding or updating information, and it reduces duplication of effort and opportunities for error. The authors are linked to individual books using Excel’s “Data Validation” tool, which restricts the “Author” column in the Book sheet to names listed in the Author sheet. When a new author is added to the Workbook, the author must be added to the “Author” sheet first, and it can then be selected for the “Book” sheet. This approach simplifies the metadata and establishes a controlled vocabulary that can be easily edited as needed. An “Author ID” was established to assist in a future migration to a relational database, should the need for that arise.

To reduce variations and improve data retrieval for books with more than one author, three columns were created: (A) “Author Name,” (B) “Second Author,” and (C) “Additional Authors.” Columns A and B are individually linked to the Author sheet described previously. Column C is a free text field, allowing multiple authors to be included as needed. In a relational database, each author would be linked to the primary Author table, but a limitation of working in Excel renders this approach impractical: one book in the inventory has ten listed authors, which would require ten individual columns, decreasing the accessibility of the Workbook. If HBW were to move to a database in the future, entries in this column would need to be mapped. As column A has 6,586 entries; column B has 227 entries; and column C has only 88 entries, a potential manual mapping would be a feasible task.

Secondly, several metadata elements require a controlled vocabulary to limit options and ensure consistency throughout the dataset. As these option lists are generally small for this project’s needs, EDW opted to implement this vocabulary manually using the Data Validation feature described previously. A third sheet titled “Controlled_vocabularies” was added to the Workbook.

Specific columns from the Book and Author sheet were linked as appropriate, creating drop-down lists for ease of use.⁶

The goal of the metadata creation is to review each book in hand to record the most accurate information available. This process can be challenging in some instances because some titles are difficult to find in print and edition primacy might be unclear, among a host of other challenges. This may result in some level of incomplete or inaccurate metadata, which, in turn, could hinder the discoverability and accessibility of resources. This consideration led us to consult external metadata resources when available to ensure the most reliable and comprehensive results.

To facilitate easier access to reliable existing metadata sources, EDW used Application Programming Interface (API) tools to harvest targeted metadata from the Library of Congress, Wikidata, and Worldcat to supplement the BBIP metadata and to make the overall process easier for the Data Tank team.

- Library of Congress [LC]: LC ID, publisher, publisher location, date of publication, genre, subject headings, author's date of birth and/or death
- Wikidata: Wikidata ID, Wikipedia link, author gender, ethnic group, date of birth and/or death, place of birth
- Worldcat: OCLC number

Harvested values from these fields were added to the Workbook to serve as authoritative resources for the Data Tank to consult as needed when working on a given title or author's entry. The ID numbers will be retained in the dataset, and the other elements can be deleted after use.

Goal 2: Development of the Metadata Creation Manual

With the necessary previous steps completed, a primary goal is to ensure that those same issues do not begin to crop up again over time. To this end, EDW set about creating a manual that can be used for training and for reference regarding each element. [See Appendix D for the complete manual.]

The Manual was created as a Word document that includes each of the elements identified in the New Schema. The document is divided into separate sections for the "Book" and "Author" spreadsheets, with the order in the Manual following the order in the Workbook. The entry for each element includes a description of what metadata should go into that element. In most cases, there is a note on how the metadata should be formatted. To help reduce potential variations, examples are often included to illustrate the formatting, such as in the entry for "Publisher Country":

Publisher Country

- Only fill this out for non-US publishers. Spell out the country completely.
 - * Ex: Nigeria
- Leave blank for US-based publishers.

⁶ For example, "Press Type" is limited to only Academic, Commercial, Religious, or Self-published/Vanity.

Most entries are quite brief; however, they may be expanded when necessary to describe how to work with potential use cases and variations. For example, the “Title” element has numerous examples covering capitalization, subtitles, ellipses, and other considerations. The length of the entry, descriptions, and examples are reflective of the frequency and types of errors encountered during the previous metadata merging stage.

Elements limited to a controlled vocabulary are noted along with the possible values. A separate section in the Manual briefly identifies these elements and notes that only a designated Data Tank staff member, most likely the team’s Coordinator, should update these values.

The Manual also includes brief sections on formatting author’s names to ensure consistency in data entry, including notes on entry for multiple authors, as described previously, and on entering special characters. In the case of HBW’s data, these are primarily letters with accents, such as é.

A limitation of working with Excel is the automatic formatting and encoding that can occur. A likely scenario is that Excel will change special characters to some sort of encoded string (for example, é can be encoded as 0xE9 , é , Ã© and more). Once this happens, it is very hard to undo, especially if the file is saved and then reopened. While this approach is not ideal, the recommendation is to use non-accented characters in their place (e.g., e instead of é). To track the correct spelling, the team member can use the Notes field to acknowledge the existence of these characters.

Example:

- Real name with correct spelling: Adébayò, Ayòbámi
- Excel’s formatting: Adébayò, Ayòbámi
- To avoid issues, enter this name as: Adebayo, Ayobami

Some elements are intended to be part of a secondary metadata creation process, and these are noted as such in the Manual.

The Manual is intended to be a living document that can be amended at any time as the need arises. For example, to address formatting questions as they arise, Data Tank, in consultation with EDW, will insert additional descriptions or instructions, clarify wording, or any number of other potential circumstances.

Goal 3: Training and Ongoing Quality Assurance

As a final step in ensuring the best possible outcome, select members of the HBW BBIP staff were trained on the Workbook and the Manual. Questions were addressed as needed, with some clarifications made to the Manual. These staff in turn trained students working with the Data Tank on the correct methods for recording metadata.

To facilitate multiple people accessing the shared Workbook, avoid duplication of effort, and build in a quality assurance process, three columns were added to each of the “Book” and “Author” sheets: “Claimed,” “Completed,” and “Checked by.” The person creating the metadata

for a given entry will put their name or initials in the “Claimed” column and make a note in the “Completed” column when their work is done. The Data Tank Coordinator or a designated Specialist team member will review the entries and formatting and enter their name in the “Checked by” column.

As the team works with these resources, questions or concerns are recorded either using the Comments feature of MS Word within the Manual directly or in a separate text document to be addressed by the Data Tank Coordinator. When appropriate, the Manual will be updated to reflect any new changes or decisions made as a result in consultation with Metadata Librarian consultant EDW and Director AH.

As previously mentioned, in addition to the metadata creation team Data Tank, BBIP has a separate team responsible for building the corpus and digitizing the books. This team tracks its progress using separate Excel spreadsheets. To ensure metadata is transferred to the Workbook accurately and without duplication, BBIP staff regularly review the Corpus Inventory and the Digitization Tracking Log. New entries are compared with existing records in the Workbook to avoid duplication, and basic metadata is added as needed. In this way, HBW can maintain the integrity of the Workbook as the primary source for BBIP metadata.

Ongoing metadata collection will require additional external funding to support the Data Tank in carrying out the detailed, research-intensive work to undertake the New Schema’s multiple phases. Once HBW’s metadata collection as well as digitization goals are achieved to complete its corpus, which consists of nearly 7,000 titles of Black literature published from 1853-2023, it will be available for scholarly research. It will also be accessible to educators, students, and the broader public for their various interests and use. Aligned with its mission of recovery, preservation, and promotion, HBW hopes this effort will foster future literary scholarship, digital humanities studies, and sustained engagement with the literary tradition.

Appendix

A: BBIP Metadata Schema (2017)

Below are all metadata elements for the original BBIP metadata schema, developed in 2016-2017 as part of an NEH grant. Full schema also included category of metadata, description of field, type of field (e.g., free text, controlled vocabulary), notes on formatting, and Boolean indicating if the element is repeatable.

- BBIP ID
- Title
- Author - Name
- Publisher
- Publisher Location
- Date of Publication
- Press type (Academic or Popular)
- Word count (Connor)
- Illustrations/Photographs
- Presence of Preface / Introduction
- Author of Preface/Introduction
- Author - Pseudonym
- Author - Gender
- Author - Ethnicity
- Author - Education
- Author - Birthplace
- Age of author when book published
- Alternative Career of Author
- Location of Action in Book
- Genre / Theme
- Tone / Literary Movement / Style
- Era in Novel
- Narrative Voice
- Rhetoric/Linguistic Features/Sentence Length
- Social/Racial Emphasis
- Vernacular
- Presence of Music
- Style of Music
- Presence of Violence
- Nature of Violence
- Presence of Verse/Verse Elements
- Presence of Speculative/Supernatural Elements
- Presence of Autobiographical Elements (family chronicle/slave narrative tradition/etc.)
- Profession of Protagonist
- Protagonist Ethnicity
- Protagonist Sexual Orientation
- Class of Protagonist
- Awards Won by Novel
- Awards Won By Author
- Genre and Year of Previous Novel by Author
- Recently Discovered Author
- Literary Predecessors/Successors
- Book Review
- Pioneering Theme/Contribution in Black Writing
- Popularity in Decade of Publication (Popular/Unpopular/Critically Acclaimed/Bestseller)
- Additional Notes / Keywords/Sources

B: BBIP Metadata Schema (2021)

All metadata elements for the 2021 BBIP metadata schema, developed by HBW stakeholders and partners, including faculty, staff, and students.

- BBIP ID
- Author Name
- Author Gender
- Author Race
- Author Nationality
- Author Birth Year
- Original Publication Title
- Publication Decade
- Original Publication Year
- Original Publisher Type
- Genre 1
- Genre 2
- Fiction Type
- Literary Movement/Era
- Main Protagonist Gender
- Novel Setting
- Author Wikipedia Page
- Novel Wikipedia Page
- Author Wikipedia Page Description
- Novel Wikipedia Page Description
- Library of Congress
- Google Scholar Citations
- Subject/Topic

C: BBIP Metadata Schema (2024)

Categorized element names in BBIP's 2024 Schema. Full list of elements along with descriptions and formatting notes are identified in the Metadata Creation Manual [See Appendix E].

Administrative

- Unique ID
- References
- Notes

Bibliographic

- Title
- Author
- Publisher
- Publisher Location
- Date of Publication
- Press Type
- Presence of Illustrations / Photographs
- Name of Illustrator
- Presence of Preface / Introduction
- Author of Preface / Introduction

Historical / Contextual

- Literary Movement/Era
- Awards Won by Novel and/or Author
- Rediscovered Novel

Content

- Word Count
- Genre/Theme
- Narrative Voice
- Usage of Vernacular [linguistic]
- Presence of Music, Violence, and Religion [topics]

Author-related

- Pseudonym
- Gender
- Ethnicity
- Birthplace (city, state, country)
- Birth Year and Death Year

External metadata sources

- Library of Congress ID
- WorldCat ID
- Author Wikipedia URL/Wikidata ID

Appendix D: Data Tank Student Contributors (2023-2024)

University of Kansas:

Abisola Akinsiku

Evan Barton

Tristan Brothers

Maddi Brown

Jade Harrison

Jackson Hoffmann

Meleah Perez

Krisdapa Sirasudhi

Selamawit Yemata

BBIP Metadata Creation Manual

Metadata for the BBIP corpus should be collected in a single Excel workbook, which will be stored in the BBIP Teams site:

Files / General / BBIP Metadata / BBIP_metadata.xlsx

There are three sheets in the workbook: (1) “Book”, (2) “Author”, and (3) “Controlled vocabularies”. Each of these will be covered in this manual.

1. “Book” contains metadata about individual books.
2. “Author” contains all the metadata about authors.
 - Note that the “Author” field in the “Book” sheet will pull from a list of existing names in the “Author” sheet.
3. “Controlled vocabularies” will contain some values used in the other two sheets.
 - **Please consult with a Data Tank Coordinator before updating this sheet.**

This document contains guidelines for creating and formatting the metadata for each field. Attention to detail is very important – it is better to take your time and be precise, rather than go quickly and include typos or other errors in the data. If you have questions or are uncertain about the values to enter, please consult with a Data Tank Coordinator.

Note: If you are unable to fill out any field for any reason, just leave it blank. Do not include other text, such as “unknown” or “N/A”, etc. If you need to draw attention to something for future review, please add a comment to the “Notes” field.

Table of Contents

<i>Formatting notes</i>	2
Formatting names	2
Special characters	2
<i>Book sheet</i>	3
<i>Author sheet</i>	8
<i>Controlled vocabularies sheet</i>	11
<i>Updating existing entries</i>	12
Book metadata	12
Author metadata	12
Important note on Author metadata	12

Formatting notes

Formatting names

There are several fields which may include personal names (author, illustrator, etc.). Use the following formatting in all cases.

- The individual’s full name, including the middle name and/or initial if known.
 - Lastname, Firstname Middle
- If there are multiple initials, do not include a space between them.
 - Ex: Mason, J.D.
 - Not: Mason, J. D.
- If the individual has a suffix as part of their name (“Jr.”, “Sr.”, etc.). format as Last, First Middle Suffix.
 - Ex: Powell, Adam Clayton Sr.
- If there are multiple authors, format them all the same way, separated by a semi-colon and a space: ‘; ’
 - Ex: Yona; Newton, Myeisha; Aryande, Shaunn; Jaye, Jammie
 - Do not enter multiple names this way: “Youmans, Asha and Allison Frank”
 - Note that the Author Name field in the Author sheet should never include more than one name.

Special characters

Unfortunately, Excel does not handle special characters well. In the case of the BBIP data, these are primarily letters with accents, such as é.

Basically, what will happen is that, at some point, Excel will change these characters to some sort of encoded string (for example, é can be encoded as 0xE9 , é , Ã© and more). Once this happens, it is very hard to undo, especially if the file is saved and then reopened.

While this is not ideal, the recommendation is to use non-accented characters in their place (e.g, e instead of é). If you want to track the correct spelling, you can use the “Notes” field to at least acknowledge the existence of these characters.

For all fields, if there is a hyphen or dash printed, always use a hyphen (the short line located to the right of the number row on the keyboard).

Example:

- Real name with correct spelling: Adébáyò, Ayòbámi
- Excel’s mangling of this: Ad√©b√°y·ªçÃÄ, Ay·ªçÃÄb√°mi
- To avoid, enter this name as: Adebayo, Ayobami

Book sheet

1. BBIP ID

- Unique ID number for each individual book.
 - * This number should correspond to the filenames of all files related to the book (e.g., scans, OCR, HTML, etc.).
 - * This number should always be included in any exported spreadsheets, e.g., distributed for analysis, projects, additional metadata creation, etc.
- Ex: 600001452

2. Title

- Full title of the book as it appears on the title page¹, including subtitle.
 - * Separate the main title from the subtitle with a colon and a space: ‘: ’
 - * Use title case, where all major words in the title are capitalized, and all other words (articles, prepositions, conjunctions) are lowercase. The first word of the title and subtitle should always be capitalized.
 - Ex: To Paris with Love: A Family Business Novel
 - * Include all words, such as “A”, “An”, etc. Include them in the order that they appear on the title page.
 - Ex: The Miseducation of Obi Ifeanyi
 - NOT: Miseducation of Obi Ifeanyi (The)
 - * If the title page includes the phrase “A novel”, include that as the subtitle.
 - Ex: Hush Harbor: A Novel
 - * If the subtitle begins with “Or”, capitalize “Or” and follow with a comma and space, then the remainder of the subtitle.
 - Ex: Out of the Darkness: Or, Diabolism and Destiny
 - * If the book is part of a series that is not distinctly part of the title, the series can be included in parentheses after the title. If there are multiple books in the series, all titles should be formatted the same way.
 - Ex: Far Beyond the Stars (Star Trek Deep Space Nine)
 - NOT: Star Trek Deep Space Nine: Far Beyond the Stars
 - * If the title has an ellipses in the middle, include a space after the ellipses but not before.
 - Ex: Ask Me No Questions... I'll Tell You No Lies
 - * If the title as printed on the title page uses an ampersand instead of the word “and”, use “&”. If the word “and” is printed, use “and”.
 - Ex: Whiskey & Ribbons

¹ The title page can be found at the beginning of the book. It usually has the title, author, and publisher listed. If the page has just the title, it’s probably not the title page (this is known as the half-title page) – check other pages to see if there is a different one that has all items listed. The copyright page will be found usually just after the title page, and has lots of technical details about the book, printing, editions, etc. When entered metadata, if there is a difference between the formatting on the title page and the copyright page, use the formatting as found on the copyright page.

3. Author

- This will be a pre-filled dropdown list in the “Book” sheet. Always check the list first to see if the author has already been entered. If you need to add a new author, see the next section on the “Author” sheet.
 - * If you are entering an author’s name and find an error with the existing value, you will need to make changes in the “Author” sheet. *Please see the note on updating existing entries at the end of this document.*
- If there is more than one named author, only include one name in this field. Additional names will go into subsequent columns.

4. Second Author

- Same formatting as “Author” above.
- This column is for books with more than one named author. If there are more than two named authors, only include one name in this field. Additional names will go into the following column.

5. Additional Authors

- This column is for books with more than two named authors.
- Format all names as described in the “Formatting names” section above. Separate multiple names with a semi-colon and a space: ‘; ’
 - * Ex: King-Gamble, Marcia; Mason, Felicia

6. Publisher

- Enter the publisher’s name as it appears on the title page, including punctuation and contractions. Differences in the publisher’s name can reflect legally or historically different entities, and specificity can make a difference.
 - * Ex: “Penguin Books” is a different entity than “Penguin Group”

7. Publisher Location

- Enter the publisher’s city and state as listed on the title page.
 - * Spell out the city completely.
 - * Abbreviate US states in the same field (two letters, no period).
 - Ex: New York, NY ; Philadelphia, PA
 - * Do not include street location, even if included on the title page.
- If multiple locations are listed, use the first one in the list.
- If the listed publisher location is a US city, and no state is listed, please include the state when known.
 - * Ex: title page states “New York”. Entry should read “New York, NY”

8. Publisher Country

- Only fill this out for non-US publishers. Spell out the country completely.
 - * Ex: Nigeria
- Leave blank for US-based publishers.

9. Date of Publication

- List the earliest date found on the copyright page as a 4-digit year.
 - * Example: Copyright page states “First published in 1973. This edition published in 2004.” – Date entered should be “1973”.
 - * Example: Copyright page states “©1923. First Penguin edition 1945.” Date entered should be “1923”.
- Do not include a month, even if one is listed.

10. Press Type

- This will be a selection from a pre-filled drop-down list. The values are stored in the “Controlled vocabularies” sheet.
- Values include: Academic, Commercial, Religious, Self-published/Vanity

11. Illustrations/Photographs

- This will be a selection from a pre-filled drop-down list. The values are stored in the “Controlled vocabularies” sheet.
- Values include: Illustrations, Photographs, Both

12. Illustrator Name

- Include if known, following the name formatting described in the “Formatting names” section (above).
- Note that you do not need to add these names to the “Author” sheet.

13. Presence of Preface/Introduction/Author’s Note

- Yes/No

14. Author of Preface/Introduction/Author’s Note

- Include if known, following the name formatting described in the “Formatting names” section (above).
- Note that you do not need to add these names to the “Author” sheet.

15. Library of Congress entry

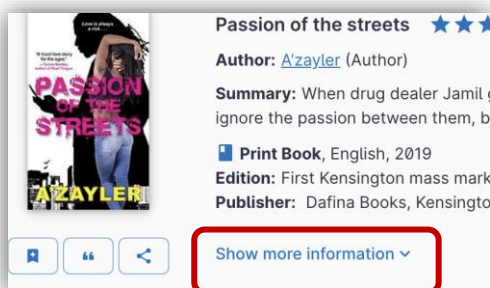
- ID number of the book's entry in the Library of Congress
- This field may already be filled out based on computational matching.
- Start at this search page: <https://catalog.loc.gov/vwebv/searchBrowse?editSearchId=E>
 - * Search by book title
 - For better results, search by the main title in quotes (omit the subtitle)
 - * Confirm that you have an exact match. From the metadata on the page, copy the LCCN value (not the URL) (located roughly halfway down the page):

Browse by shelf order	PS3601.Z39
LCCN	2018288068
Dewey class no.	813/.6

- Note that if you have questions about other book metadata (spelling, title, publication date, etc.), this is a reliable source to confirm.

16. WorldCat entry

- Direct link to the book's WorldCat ID (OCLC)
- This field may already be filled out based on computational matching.
- Start at this search page: <https://search.worldcat.org/>
 - * Search by title (use quotes for better results)
 - * Optional: Use the filter on the left to limit to “Print Books”
 - * Confirm that you have an exact match. Click “Show more information”



- From the metadata that appears, copy the “OCLC Number / Unique Identifier:” (not the URL)
- Note that this number also appears in the URL, so you could just grab the number from there.
 - * Ex: <https://search.worldcat.org/title/1078689597>
- Note that this is not a reliable source about specific book metadata (spelling, title, publication date, etc.).²

² OCLC is a great resource, but entries are often added by individual library catalogers and tend to be edition-specific and not necessarily reflective of first edition publication information.

17. Genre/theme

- One or more keywords describing the genre and/or theme of the book.
- This will be a selection from a pre-filled drop-down list. The values are stored in the “Controlled vocabularies” sheet.

18. Literary movement/Era

- Literary movement that the book is a part of.
- This will be a selection from a pre-filled drop-down list. The values are stored in the “Controlled vocabularies” sheet.

19. Narrative Voice

- This will be a selection from a pre-filled drop-down list. The values are stored in the “Controlled vocabularies” sheet.
- Values include: First person, Second person, Third person, Mixed

20. Use of Vernacular

- A brief standardized description of the use of vernacular in the book.
- This will be part of a secondary data collection project.

21. Presence of Music

- A brief standardized description of the presence of music in the book.
- This will be part of a secondary data collection project.

22. Presence of Violence

- A brief standardized description of the presence of violence in the book.
- This will be part of a secondary data collection project.

23. Presence of Religion

- A brief standardized description of the presence of religion in the book.
- This will be part of a secondary data collection project.

24. Awards won by novel

- Enter as “Name of Award (Date as 4-digit year)”
 - * Example: Booker Prize (1987)
- For nominations, include as “Name of Award (Date as 4-digit year) [type of nomination]”
 - * *Note the parentheses for the date and the square brackets for the type*
 - * Example: Booker Prize (1987) [short list]
 - * Example: Book of the Year (1965) [nominee]
- Can include inclusions on lists in this section in the same format.
 - * Example: 100 Notable Books of 2020 - The New York Times (2020)
- Separate multiple values with a semi-colon and a space: ‘; ’

25. Rediscovered novel

- This is a specific category that indicates the book was included in the original HBW list of titles.
- This will be part of a secondary data collection project.

26. References

- Optional: Include any external sources consulted during the creation of the metadata.
- Separate multiple references with a semi-colon and a space: ‘; ’

27. Notes

- Optional: Include any pertinent notes about the book, metadata, process, questions that you have, etc.

Author sheet

1. Author ID

- Numeric value that will be essential for future data migrations, such as to a relational database format.
 - * This number should always be included in any exported spreadsheets, e.g., distributed for analysis, projects, additional metadata creation, etc.
- Following the same format as the Book ID – 9 digits, starting the values with a ‘9’ for differentiating from Book IDs.
 - * Ex: 900001452

2. Author

- The author’s full name. Include as much as is known, following the name formatting described in the “Formatting names” section (above).

3. Author Pseudonym

- Any known pseudonyms used by the author. Format following the name formatting described in the “Formatting names” section (above).
 - * Separate multiple values with a semi-colon and a space: ‘; ’

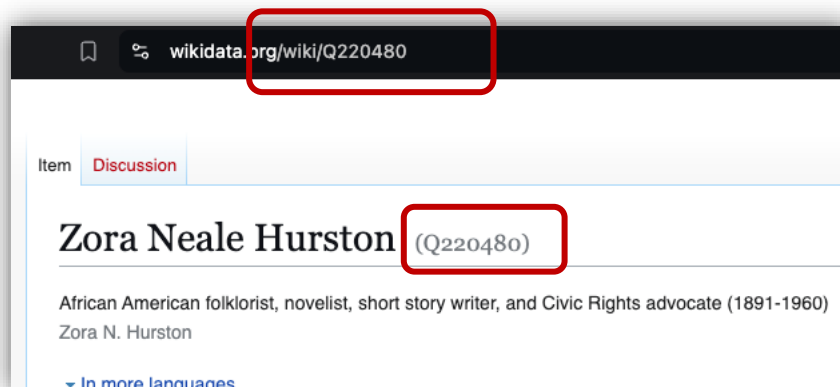
4. Author Gender

- This will be a selection from a pre-filled drop-down list. The values are stored in the “Controlled vocabularies” sheet.
- Values include: Male, Female, Non-binary

5. Author Ethnicity

- A very high standard of proof is needed for this field. In general, this means (1) the author claims it themselves, or (2) it is widely agreed on by scholars.

6. Author Birthplace (city and state)
 - If known, the city and state where the author was born.
 - * Spell out the city completely
 - * Abbreviate US states in the same field
 - Ex: New York, NY ; Philadelphia, PA
7. Author Birthplace (country)
 - Spell out country name completely.
 - Ex: Nigeria, United States
8. Author birth year
 - Enter as a 4-digit year. Do not include month or day.
9. Author death year
 - Enter as a 4-digit year. Do not include month or day.
10. Author Wikidata ID
 - Wikidata ID for the author's reference page
 - This field may already be filled out based on computational matching.
 - Wikidata can be a rich source for specific metadata points (e.g., birth and death years, birthplace, etc.)
 - Start at this page and search by name: <https://www.wikidata.org/>
 - * The ID number begins with a "Q" and appears next to the name. It also forms the direct URL for the page in this format: <https://www.wikidata.org/wiki/Q#####>



11. Author Wikipedia page
 - Enter the full URL for the author's Wikipedia page.
 - This field may already be filled out based on computational matching.
 - Ex: https://en.wikipedia.org/wiki/Toni_Morrison

12. Awards won by author

- Enter as “Name of Award (Date as 4-digit year)”
 - * Example: Booker Prize (1987)
- For nominations, include as “Name of Award (Date as 4-digit year) [type of nomination]”
 - * *Note the parentheses for the date and the square brackets for the type*
 - * Example: Booker Prize (1987) [short list] [long list]
 - * Example: New Author of the Year (1965) [nominee]
- Can include inclusions on lists in this section in the same format.
 - * Example: Best New Authors of 2020 - The New York Times (2020)
- Separate multiple values with a semi-colon and a space: ‘; ’

13. References

- Optional: Include any external sources consulted during the creation of the metadata.

14. Notes

- Optional: Include any pertinent notes about the book, metadata, process, questions that you have, etc.

Controlled vocabularies sheet

A controlled vocabulary is simply a pre-selected list of values that can be used for a specific field. This sheet has the values for the following fields on the “Book” and “Author” sheets.

Due mainly to limitations of working in Excel in a Teams environment, changes made to the “Controlled vocabularies” sheet are not automatically reflected in the other sheets. If you make changes to the values in the “Controlled vocabularies” sheet, you must also manually make those changes to the corresponding fields in the other sheets.

Please consult with a Data Tank Coordinator before updating values in this sheet.

- Books
 - Press type
 - Illustrations / Photographs
 - Narrative Voice

- Author
 - Author gender

Updating existing entries

As you work through the metadata collection and cleanup, you will likely find mistakes or duplicate entries.

Book metadata

If you find duplicate entries for the same title, please perform the following checklist:

1. Determine which entry should be retained.
2. Copy all metadata from other duplicate rows into this row.
 - Handling conflicting metadata between entries:
 - * If the data source has an earlier publication date, this data source should be considered the de facto source.
 - * If data cannot be merged, metadata fields should be entered as multiple values separated by a semi-colon and a space: ‘; ’
 - **Make sure the BBIP ID numbers are all reflected in the row you are keeping.**
 - * These numbers are separated by a pipe symbol with no spaces ‘|’
3. Update the “duplicates_removed” column to reflect the number of rows you are removing.
4. Delete any unneeded rows.

Author metadata

If you find duplicate entries for the same author, please perform the following checklist:

1. Determine which entry should be retained.
2. Copy all metadata from other duplicate rows into this row.
 - Handling conflicting metadata between entries:
 - * If data cannot be merged, metadata fields should be entered as multiple values separated by a semi-colon and a space: ‘; ’
 - Make sure the Author ID numbers are all reflected in the row you are keeping.
 - * These numbers are separated by a pipe symbol with no spaces: ‘|’
[The “pipe” symbol is usually located just above the “Enter”/”Return” key on the keyboard.]
3. Delete any unneeded rows.

Important note on Author metadata

Due mainly to limitations of working in Excel in a Teams environment, changes made to the Author sheet are not automatically reflected to existing entries in the Books sheet.

If you make changes to the “Author” field in the “Author” sheet, you must also manually make those changes to the “Author” field in the “Books” sheet.