Enslaved | Peoples of the Historical Slave Trade

Enslaved.org Recommended Practices for Historical Slavery Data

March 2, 2022

VERSION 1

By Catherine Foley and Alicia M. Sheill

DOI: 10.25971/5xt9-7x54

Introduction **3**

I. Terminology **5**

Dataset vs. Database 5

Metadata 6

Linked Open Data and Ontologies 6

II. Managing Digital Data **8**

Tooling for Data Collection 8

Digital Files 8

Referencing Digital Files Within a Dataset 9

Digital File Naming 9

File Formats 9

III. Recommended Practices for Historical Slavery Data **11**

Defining Metadata 11

Identifying Records 12

Extraction and Recording Methods 14

Applying Controlled Vocabularies 15

IV. Using Key Enslaved.org Concepts **17**

Source 17

Create Fields for Citations in Individual Records or Create a Source Record 18

Person 19

Names 19

Age and Age Category 20

Sex 20

Relationships 20

Birthdate and Death Date 21

Person Data Points at Specific Event Dates 21

Differentiate People with Unique Identifiers 23

Connection to Event 24

Event 24

Event Type 24

Single-Person Events and Group Events 24

Dates 25

Connection to Place 25

Place 26

Place Type 26

Coordinates 26

Contextualizing Place through “Located In...” Statements 26

Data connections 27

Acknowledgments **30**

Introduction

This document explains recommended practices for data collection, metadata creation, data extraction, and related workflows for *Enslaved: Peoples of the Historical Slave Trade* (Enslaved.org). Enslaved.org is a digital platform to learn about the lives of individuals who were enslaved, owned slaves, or participated in the historical slave trade. A primary goal of Enslaved.org is to record the names of enslaved women, men, and children of African descent.

This document is a general framework describing how information about the names and lives of enslaved people and historical slavery can be extracted from archival materials in a way that acknowledges the humanity of the people described in the documents and structured so that the data can be understood and used now and in the future.

It is understood that dataset creators extract information from records for their own purposes and projects, and thus this set of recommendations is not meant to be prescriptive. The guide has been written with two primary audiences in mind: 1) data collectors starting out in the field who would like to learn about recommended practices, and 2) seasoned researchers who would like to enhance their work to make it more shareable and usable by other scholars. It can also be helpful to those scholars, genealogists, and members of the general public who are doing research in the field of historical slavery. It can be particularly helpful to those institutions -- libraries, archives, museums, historical sites -- that would like to prepare contributions to Enslaved.org. Finally, it can help teach students data-informed historical methods.

These recommended practices describe various considerations for specific fields that dataset creators may capture during data collection. Enslaved.org maintains documentation defining records and fields that reflect the type of information that the site integrates in its discovery hub. These documents can be found at docs.enslaved.org/metadata. These are not a definitive set of fields for a historical slavery data, but a subset of fields used by Enslaved.org to build an interconnected system of tools to search, browse, visualize, and analyze disparate, previously siloed datasets in a single location. Enslaved.org developed the records, fields, and controlled vocabularies it uses to link together historical slavery data in collaboration with a team of historians of slavery.

Within this recommended practices document, Enslaved.org provides more elaborate explanations for fields and data extraction processes that have challenged previous contributors to Enslaved.org. This document and other Enslaved.org materials are guides, not restrticive blueprints, for designing new historical slavery datasets.

Before jumping into discussing recommendations, it will be helpful to review the ethics of data curation and key terms used in the document. Enslaved.org's Statement of Ethics articulates the importance of and need for an intentional approach to ethically handling historical materials that document the lives of enslaved women, children, men, and families. As the statement explains:

> Although the historical record and data-driven approaches to history have often rendered enslaved people as nameless, we aspire to use historical data collections to recover, aggregate, and make accessible the names and life stories of enslaved people. We recognize that primary source material reflects the perspectives and biases of the material's authors, reflecting the systems of power and racist ideologies of the periods in which they were written. Most often, they fail to acknowledge the humanity of enslaved people and instead commodify Black bodies and experiences. We are committed to identifying by name as many enslaved people as possible and to representing individual and collective experiences in an intentional, humane, and ethical frame. We seek out submissions that read against the grain of dehumanizing archival perspectives, pursue alternative sources, recognize their own positionality within larger systems of power, and support descendant communities in telling their own histories. We work collaboratively with researchers and descendant communities to continually develop and follow practices that respect the lives of enslaved people.[1]

Enslaved.org also encourages data collectors to make their data "FAIR data," i.e. data that is findable, accessible, interoperable, and reusable.[2] Data that is discoverable and open maximizes the value of the data by making it useful and usable to others. This document is intended to provide guidance to data creators about how to describe their data so that other users can locate and reuse the data for purposes not originally envisioned by the data creator. To learn more about FAIR principles for data management and stewardship, see: https://www.go-fair.org/fair-principles/.

As part of FAIR data principles and with the recognition that any digital project could go dark, including Enslaved.org, multiple layers of preservation are encouraged. In addition to making datasets finable and accessible in the Enslved.org hub, Enslaved.org uses Dataverse (https://support.dataverse.harvard.edu/), a general purpose data repository operated by Harvard University, where datasets are assigned a DOI, preserved, and made searchable alongside data produced and shared by researchers from many disciplines. Enslaved.org data is deposited in its *Journal of Slavery and Data Preservation* Dataverse (https://dataverse.harvard.edu/dataverse/jsdp) under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0). Contributors are also welcome to propose an alternative data repository; contributors have used the University of Pennsylvania Scholarly Commons, GitHub, Zenodo, and the UK Data Service, among others.

[1] https://enslaved.org/statementofEthics/

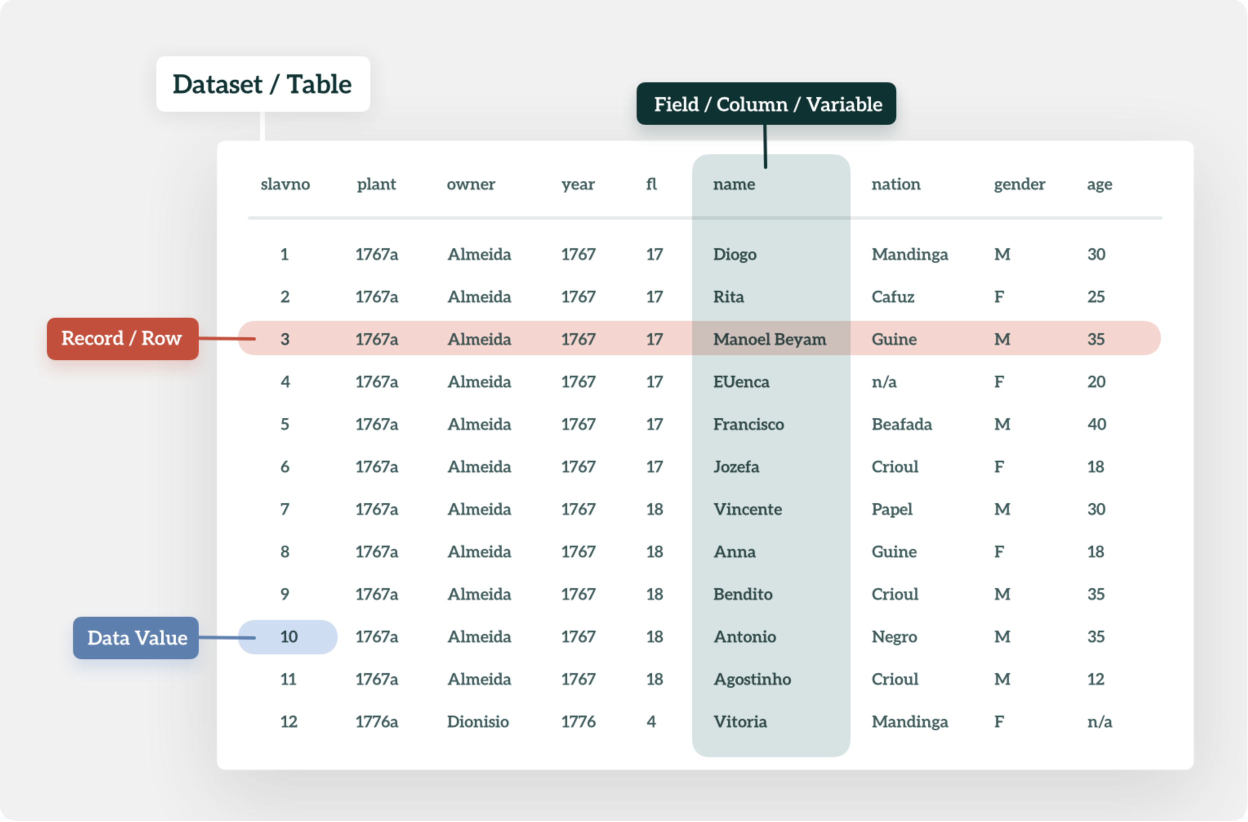
[2] Mark D. Wilkinson, Michael Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, et al. "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data* 3 (15 March 2016) 160018. doi:10.1038/sdata.2016.18.

I. Terminology

To better understand the following recommendations, it is helpful to consider key terms, their definitions, and their uses.

Dataset vs. Database

In the Enslaved.org context, **datasets** are most often tabular representations of information extracted from historical documents. In some instances, the **dataset** is stored in spreadsheets like Microsoft Excel or in a **table**. Irrespective of the carrier, the structured data at the center of Enslaved.org is digital, tabular, and arranged in a series of ordered columns and rows. Each **row** contains the same **columns** in the same order as all other **rows** (see Illustration 1: Dataset Components below). Every column, called a **field or variable**, contains a set of **data values** that are all of the same type. For example, the **dataset** could contain a **column** of names written out as text or a **column** of ages typed as numbers. The label or name of the **column** describes its contents. The **columns** define the structure of the **table**, while the **rows**, called **records**, contain the entries that make up the **dataset**.


Dataset / Table
Field / Column / Variable
Record / Row
Data Value

slavno	plant	owner	year	fl	name	nation	gender	age
1	1767a	Almeida	1767	17	Diogo	Mandinga	M	30
2	1767a	Almeida	1767	17	Rita	Cafuz	F	25
3	1767a	Almeida	1767	17	Manoel Beyam	Guine	M	35
4	1767a	Almeida	1767	17	EUenca	n/a	F	20
5	1767a	Almeida	1767	17	Francisco	Beafada	M	40
6	1767a	Almeida	1767	17	Jozefa	Crioul	F	18
7	1767a	Almeida	1767	18	Vincente	Papel	M	30
8	1767a	Almeida	1767	18	Anna	Guine	F	18
9	1767a	Almeida	1767	18	Bendito	Crioul	M	35
10	1767a	Almeida	1767	18	Antonio	Negro	M	35
11	1767a	Almeida	1767	18	Agostinho	Crioul	M	12
12	1776a	Dionisio	1776	4	Vitoria	Mandinga	F	n/a

Illustration 1: Dataset Components

Although a dataset could be integrated into a database, a database is more complex and is defined as "an organized collection of data, generally stored and accessed electronically from a computer system."[3] Notice that this specifies a "computer system," not just a "computer." This is because databases use database management systems or software to interact with users, other computer system applications, and the data itself. Oftentimes, researchers refer to their datasets as databases, but for our purposes one must be running database software to accurately describe one's dataset as a database. For most purposes when considering the process of extracting data from historical records, Enslaved.org recommends simply using the term "dataset." See the section on "Tooling for data collection" below for additional details.

Metadata

Metadata is information about the dataset describing its content, organizational structure, and characteristics, so that the data can be understood, stored, discovered, accessed, used, and reused.[4] In the context of Enslaved.org, metadata is robust documentation about the data so that it can be reused and interpreted for purposes not necessarily envisioned by the creator of the data. There are different types of metadata including (1) descriptive metadata, which describes information about a resource such as what it is and who created it; (2) structural metadata, which explains how a resource is organized and relationships between data elements; and (3) administrative metadata, which records provenance and rights information describing the origin and intellectual property of a resource. In short, metadata explains how resources or data are organized, used and managed.

Linked Open Data and Ontologies

Linked open data (LOD) is a method of publishing, sharing, and connecting structured data on the web. Data from different sources is standardized, aggregated, and formatted in such a way that it is machine readable and is structured by an underlying ontology. The Enslaved.org ontology defines concepts (e.g. Person, Event, Place, and Source) within the domain of historical slavery and relationships between these concepts. The abstract organization of the ontology is made practical in the Enslaved.org metadata documentation, which describes core data points expressed in our data model, for example Name, Sex, Occupation, and Freedom Status for people.

The key point here is that when LOD is generated from the original dataset, it stands as a separate dataset from the original. Any dataset added to Enslaved.org remains untouched and in its original deposited state within the data repository (Dataverse or another of the creator's

[3] https://en.wikipedia.org/wiki/Database

[4] William K. Michener, "Creating and Managing Metadata," in *Ecological Informatics: Data Management and Knowledge Discovery*, edited by Friedrich Recknagel and William K. Michener: 71 (Cham, Switzerland: Springer, 2018), https://doi.org/10.1007/978-3-319-59928-1_5; William K. Michener, James W. Brunt, John J. Helly, Thomas B. Kirchner, Susan G. Stafford, "Non-geospatial metadata for the ecological sciences," *Ecological Applications* 7 (1997): 331, https://esajournals.onlinelibrary.wiley.com/doi/epdf/10.1890/1051-0761%281997%29007%5B0330%3ANMFTES%5D2.0.CO%3B2.

choosing). The extracted LOD collected in the Enslaved.org hub helps to link the user both back to the original dataset and to other published datasets.

In very simple terms, an ontology defines the concepts, categories, vocabularies, and the relationships among them in a particular subject area or domain. For example, you can see the Enslaved.org ontology at https://docs.enslaved.org/ontology/. The ontology helps to make the LOD more machine readable and make the links and relationships between pieces of data. In short, computers do not understand what humans take for granted -- i.e., what is a baptism? who is involved? why is it done? -- thus the ontology helps to make searching and using the data more successful and effective.

Much more could be said on the topic of linked open data and ontologiest. For more details see the World Wide Web Consortium (W3C) (https://www.w3.org/wiki/LinkedData) and (https://www.w3.org/standards/semanticweb/ontology).

II. Managing Digital Data

Although many aspects of digital data management go beyond the purview of this document, below are several recommended practices that Enslaved.org have found to be useful to data collectors.

Tooling for Data Collection

When choosing a software application for data collection, what you pick is not as important as how comfortable you are with using it. If you have never used Microsoft Access and don't own a computer that runs it, then using it just because you heard it is database software is not setting yourself up for success. Since most dataset creators Enslaved.org encounters have single table spreadsheets, they often choose Google Sheets to collect, edit, and manipulate their data. Projects with multiple tables of data can use spreadsheets with multiple tabs/sheets within a single file. Often dataset creators leverage *"Identifiers"* to connect records and data in the different sheets.

Understanding the limitation of the software you choose is important to maintaining the integrity of your data. Google Sheets makes it easy to add new records and fields, but they also make it easy to delete said records and fields along with individual cells of data, so be careful of making unwanted deletions. Neither Google Sheets nor Microsoft Excel offers a particularly robust tracking system or history for edits of individual data points. Some general ability to roll back to previous dates and times (encompassing potentially many untracked changes) is available. Google Sheets is only available on the web, but also allows for multiple people to review, edit, and manipulate the data in tandem. Depending upon the nature of your data collaboration and internet access, this may or may not be appealing.

Many data collectors are familiar with Microsoft Excel, but there are several drawbacks to using it, described in the "File Formats" section below; finding an alternative to Excel could be especially important for projects working with sources in languages with special characters.

The most important thing to remember when choosing a data application is having the ability to export to a properly encoded, non-proprietary file format to ensure transferability of the data and longevity between platforms. Please see the "File Formats" section below.

Digital Files

Clearly organize your data file(s) in a single directory/folder on your computer or selected cloud storage system. If you keep the files within a series of directory folders, the exact path to the file should also need to be included as part of the file name within the data file.

Referencing Digital Files Within a Dataset

Many digital archive projects include related digital files, whether scans of original documents from an archive, photographs, or even audio and video files. It is important to manage these digital files related to the dataset. To accomplish this, include the exact filename in a column within your record. Persistent identifiers such as URIs or DOIs can also be included within a column of the dataset.

Digital File Naming

Although often overlooked, clearly naming digital files can make a huge difference in the ease of sharing and publishing project data on the web. It is strongly recommended to use basic characters (A-Z, a-z, 0-9) when generating file names for digital files. Enslaved.org advises against using 1) special characters (parentheses, commas, other punctuation, etc.), 2) spaces, and 3) diacritics (a sign added to a basic letter such as an accent or umlaut), as such characters can hinder or even stop the transfer of files between computers and can render incorrectly on the web if not encoded correctly. It is better to avoid using these characters altogether within file names.

File names, including extensions (e.g. .jpg, .gif, .txt, .mov, etc.) must be unique to an individual file, i.e. no two files can have the same name. Leveraging record identifiers to create a filename is a good strategy for creating unique filenames (see section on "Identifying Records" above). Capitalizations count in file naming and must be maintained within a filename but casing should not be used to create a unique file name.

File names can be versioned by appending a suffix to the end of the file name. This suffix could be an alphabetical letter or a more robust mechanism, like a date with a timestamp. It is recommended to use the format of YYYYMMDD for dates, so January 28, 2021 would be 20210128. Formatting dates in this method will allow easier sorting and analyzing of data and files.

Finally, filenames should include the extension of the file as it is an important part of the file that sometimes does not display depending upon the setup of one's computer. A jpg image file for my Newspaper Advertisement Event (referenced in the "Identifying Records" section) created in February 1835 could have the identifier EVE-ADV-0001-183502.jpg.

File Formats

It is important to have the ability to export your dataset to a properly encoded, non-proprietary file format to ensure transferability of the data and longevity between platforms. Microsoft Excel is an example of a proprietary, i.e. paid-for software (a user has to own said software to view xlsx files). Microsoft Excel, LibreOffice, Google Sheets and other software applications have the ability to export documents to open formats such as comma separated values (csv) or tab separated values (tsv), which are the recommended formats for Enslaved.org since they can be read by multiple programs. (Even if a data collector uses Excel to create a dataset, they can submit a csv or tsv export of their data instead of an xlsx file when it comes time for submission.) The ability to export your data into an open format like csv makes it easier to share your data, as well as integrate it into other systems, thus increasing the lifespan of the dataset. Using a file type like csv or tsv is also preferred because they maintain the data's encoding.

Failure to properly encode a dataset may lead to the loss of unique characters such as diacritics or glyphs, or dates found within the data. For most Roman language script datasets, the use of the Unicode Transformation Format 8 (UTF-8) encoding often provides a sufficient set of characters. Encodings are easy to include when saving or exporting data to open formats such as csv, but be aware that proprietary softwares like Microsoft Excel use their own set of encoding formats that may not be directly transferable to other systems. Microsoft Excel is also known to have difficulty with maintaining ISO date formats. Non-Roman scripts such as Chinese, Japanese, Korean, Cyrillic, Hebrew, and Arabic require additional encoding formats that should be considered and formatted properly for the data found within the dataset.

III. Recommended Practices for Historical Slavery Data

Although there are many ways to capture data related to the historical slave trade, Enslaved.org chooses to focus on People and the recreation of their lives by describing Events that included them, when and where these life events occurred, and what sources they are documented in. The inclusion of details pertaining to the historical documents that led to the creation of these records give integrity to the data that is collected. All of this information is captured within Enslaved.org's four record types: Person, Event, Place, and Source.

This section provides recommended practices for the collection and processing of tabular data for contributors to Enslaved.org. The recommendations include: **1)** defining metadata, **2)** identifying records, **3)** applying controlled vocabularies, **4)** extraction and recording methods. Suggestions regarding building connections across the data are discussed throughout.

Defining Metadata

Metadata is robust documentation about the dataset so that the data can be interpreted and reused for purposes not necessarily envisioned by the creator of the data. Metadata includes detailed descriptions about how the data is collected and processed into a dataset. Contributors to Enslaved.org describe this process with a description of their data project and an explanation of their methodology in the short data article that accompanies each dataset, published in the *Journal of Slavery and Data Preservation* (JSDP).

Metadata also includes detailed information about each field/column heading in the dataset. Enslaved.org and JSDP documents often refer to this as a field definitions table; in other contexts, you might see it referred to as a data dictionary or part of a codebook. In a separate tab/sheet in your dataset or in a distinct word processing document or spreadsheet, contributors should clearly list and write a brief definition explaining the fields or column headings used in the dataset. The definitions need to specify the exact meaning of the data recorded in the field in plain language. It is helpful not to simply repeat column headers in your definitions. Column field/column heading definitions should provide context about the extracted data so that someone who is unfamiliar with the original source document understands what the data is and how it was extracted. It may be obvious to the data collector that the Date column refers to the date when an enslaved person took flight and escaped bondage. But there are many potential dates that could be related to a fugitive's life experiences, for example, the date when another person placed an advertisement in a newspaper, or a date when the advertiser claimed to have last seen the enslaved person. These nuances should be clarified in the field/column heading definitions. **The most important thing a data collector can do is robustly document the data as it is being collected.**

Recommended practices documentation and definitions should attempt to answer these questions:

- What is the data you are capturing?
- How are you recording that data (this could include use of controlled vocabulary terms)?
- Are the terms pulled directly from the source document (extracted/verbatim), or did you interpret content in the source document in order to capture the data point? For example, did you impute Sex based on the gendered first name of a person, or was it explicitly noted in the document?
- Did you transform abbreviations used in the original document into a set of standard terms for titles or honorifics?
- Did you introduce an identifier for each person in the dataset, one that was not present in the source document?
- Are you recording additional information, such as the order in which data appeared in the original source (sort order)?
- What conventions did you use to convey ambiguous information in the original source? For example how did you record illegible letters or words using brackets, etc.?

See Enslaved.org’s metadata documentation (https://docs.enslaved.org/metadata/) for a sample of how contributors can describe their fields with meaningful definitions.

Identifying Records

Enslaved.org recommends every record within a dataset have a unique identifier. This is one of the most important things a data collector can do at the outset of a new data collection project. How individual projects implement identifiers will vary based on the goals of the research. Some projects attempt to unambiguously identify the same person across all of the sources analyzed. In other cases, the dataset creator might be more interested in precisely recording every detail of the original source without explicit attention to disambiguating the people described in the document.

A unique identifier is a string of characters and/or numbers that is unique to one single record and is not repeated throughout the entirety of the dataset unless it is to connect related records (more on this later in this document). If unique identifiers are not already present within the dataset, they can be included as a new field/column. The creation of unique identifiers can be as simple as numbering your rows starting at “1” or as complicated as adding prefixes before each number indicating data structure, i.e. differences between record types. For example, if I were to have Person (PER) 1 and Event (EVE) 1, I may want to use PER-00001 and EVE-00001 to differentiate between the two records. Enslaved.org recommends that data collectors use a minimum of 3 characters when identifying record types, i.e. EVE instead of just EV for Event. For numbers included in an identifier, Enslaved.org recommends always maintaining the same number of digits, so an identifier number beginning with 00001 would at most have 99,999 entries in that record type. The inclusion of unique identifiers can help with sorting and analyzing the data as it is developed.

If data collectors want to provide further indications of record differences within a record type, they could add Person Role, Event Type, or Place Type to the identifier. For example, if there were two types of events, e.g., Court Cases (CAS) and Newspaper Advertisements (ADV), our identifiers would be EVE-CAS-0001 for Court Case 1, and EVE-ADV-0001 for Newspaper Advertisement 1. (When creating linked open data, Enslaved.org staff will add a prefix to all records coming into its discovery hub to indicate the project that contributed the record. This can be a helpful tip for individual data collectors working on more than one data collection project.)

Unique identifiers are a powerful tool that can be very useful for data collectors (and for others who want to use the researcher's data). They can be used to make connections between data points and to other records within the dataset, and make the connections correctly and effectively. See the "Data Connections" section below for more on this topic.

Another key to good data practices is to consider the full lifecycle of data curation. Enslaved .org's four identifier field types (Person Identifier, Event Identifier, Place Identifier, and Source Identifier) support immediate and long-term data management, including augmenting and updating records. These fields uniquely reference every single person, event, place, or source record in a dataset. Existing only once in a dataset, this identifier field is used to disambiguate one record from all others in a dataset.[5] As such, it allows a dataset creator to make a record for an event. For example, the Trans-Atlantic Slave Trade Database assigned each ship voyage, the primary unit of analysis in the project, a unique identification number. The project used this identifier to unambiguously distinguish a single voyage from all others that may have the same ship name, voyage duration or time frame, or ports of call. As data in a voyage record expands or gets modified over time, the individual VOYAGEID facilitates precise updating of the correct record. Unique identifiers are a powerful tool to reduce data duplication and potential recording errors.

Manual creation and maintenance of identifiers is challenging, especially without built in error-checking. The onus is on the data collector to review the identifiers and guarantee the uniqueness of each one. Although onerous, many creators consider it valuable work to make data more shareable and usable.

When combining all of these methods, human readable identifiers will be created that give data management staff the ability to tell 1) which project contributed the record, 2) the overarching record type of Person, Place, Event, or Source, and sometimes 3) more specific information about the record given the complexity of the project such as Person Role, Event Type, or Place Type.

To read more about how other organizations implement identifiers, see Leigh Dodds, "How do

[5] Patricia Harpring, *Introduction to Controlled Vocabularies* (Los Angeles: Getty Publications, 2010), 237. https://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/index.html.

different communities create unique identifiers?" Lost Boy blog, https://blog.ldodds.com/2020/04/14/how-do-different-communities-create-unique-identifiers/ or Victor Chircu, "7 Strategies for Assigning Ids," Simple Oriented Architecture blog, https://www.simpleorientedarchitecture.com/7-strategies-for-assigning-ids/. We also recommend JSDP editor Daryle Williams's discussion in "Free Africans and Concessionaires, Rio de Janeiro, 1860," *Journal of Slavery and Data Preservation* 2, no. 1 (2021), https://doi.org/10.25971/s1bj-6242.

Extraction and Recording Methods

Depending on the dataset creator's research interests, the focus of the project, and the nature of the source materials analyzed, a dataset creator will likely make decisions about how to extract and record data. Possibilities include 1) faithfully recording all of the data in a source document (**verbatim capture**); 2) inferring additional information from the source material (**interpreting or imputing data**); and/or 3) translating some or all data from the source document into another language (**translation**). There is no best or right way to make these decisions. Often data collectors will implement more than one strategy when extracting and recording (i.e., a data collector may use one data value to record a verbatim term in one column, translate the term in another column, and/or interpret the term in a final column. The selection of how to extract and record is completely dependent on the dataset creator's research goals. **What is important is that the data collector *thoroughly documents* the fields and the controlled vocabularies they are using to structure their dataset at the outset of a project**. This documentation should include names and definitions of the fields, stating if each field appeared in the source (verbatim, inferred, and/or translated), and how the data was recorded within the field. See the "Defining metadata" section above.

An example will illustrate the differences between these approaches and the value of careful documentation. One data collector studying post-mortem inventories of slaveholders' possessions in Brazil created a dataset with a blended approach. The column headings for the core part of this dataset were derived from relatively standardized paragraph descriptions of each enslaved person listed in the inventory. The inventory takers included the same types of information for each person in the same order across all inventories. The data collector broke these categories into column headings, which they labeled or named in English even though the inventories were in Portuguese. For many of the fields, the data collector translated into English the data included in the Portuguese descriptions: M (for male) and F (for female) were recorded for male and female in Portuguese. Some data was inferred. When Sex was not explicitly recorded in the document, for example, the data collector looked to other gendered words or to the names of the enslaved to determine Sex. Finally, the dataset creator added an entirely new column found nowhere in the archival document to record their analysis or interpretation of the "nation" recorded for each enslaved person. The dataset creator defined a standard list of geographic regions into which they placed the various listed nations, thus identifying the place of origin for each enslaved person, which was their research focus.

From this example, it becomes clear that coding data (as in the M for male and F for female example above) or using other forms of shorthand can lead to confusion for people trying to read and analyze the data, if the data collector does not document or define the codes. This is especially true for datasets that are not in the same language as the archival documents or that use jargon that the data collector added to the dataset, hence the necessity of documenting one's choices in the field definitions table/data dictionary and the Methodology section of *Journal of Slavery and Data Preservation* data articles.

Applying Controlled Vocabularies

Controlled vocabularies are used to standardize data values within like fields and provide consistency across a dataset.[6] These carefully selected and defined lists of words and phrases provide a consistent way to categorize data values within a field so that similar data in disparate datasets can be searched effectively. For example, when describing Sex within the Enslaved.org controlled vocabulary, the terms "Female," "Male," and "Intersex" are used. These controlled terms standardize words that encompass this category in historical documents such as woman, man, lady, gentleman, boy, girl, mujer, hombre, niña, niño, or abbreviations like M and F used in some datasets. Aligning with Enslaved.org's controlled vocabularies is recommended, but not required.[7] As in the following example, the key is both to establish a standardized process for extracting and recording the data for each field and to document the process for each field. Although controlled vocabularies encourage shared language, the current list of terms will not be sufficient to accommodate the specifics of every historical example; use whatever terms you think are best and after initial submission, talk to Enslaved.org staff about the possibility of adding these terms to the controlled vocabularies to better represent your evidence.

Pro-tip: **Specify** and **define** terms appropriate for a category. These terms can be based on the data in the historical document but standardized so that they can be consistently used across records in the dataset. You might consider listing specific terms that you mapped to a single term and/or how you determined the appropriate controlled vocabulary to apply based on other aspects of the dataset. For the Sex example, you could specify that a historical ledger included a heading named "Gender" and recorded that data as "M" or "F" but that you decided to record the data as Male (for "M") and Female (for "F"). Or that the Sex column you included in the dataset did *not* appear in the original document but is data you inferred (or interpreted/imputed) based on other information in the source document, for example from the first or given name of a person or gendered phrases in a description field–for example "gentleman," "boy," "hombre," and "niño," which you coded as Male, while "woman," "lady," "girl," "mujer," and "niña" was categorized

[6] Six Person fields use a controlled vocabulary: Sex, Age Category, Occupation, Relationships, Person Status (freedom status), and Role within Event. Event has one controlled vocabulary that categories the Type of slavery/slave trade event, for example a manumission, marriage, or sale. The Place Type vocabulary defines terms such as county or parish, domicile, maroon community, plantation, port, etc. Definitions for all controlled vocabulary terms are located at https://docs.enslaved.org/controlledVocabulary/.

[7] Enslaved.org continues to expand and refine the controlled vocabularies as contributors submit new data.

as Female. Maintain documentation about your methods, terms, and controlled vocabularies in a separate tab/sheet within your data collection or in ancillary files. New terms can be added over the course of the data collection as the need arises. Be sure to update your documentation as you make decisions to add or modify terms. Clear delineation of controlled vocabulary terms allows you and anyone else assisting with the dataset to consistently use terms across all records during the entire collection period, even when your memory fades about the significance of including or editing terms. See the “Defining Metadata” section above for further information on the importance of documentation.

IV. Using Key Enslaved.org Concepts

This section will describe recommended practices for using Enslaved.org's four core concepts: Source, Person, Event, and Place.

Source

On the most basic level, a source is something that contains historical information about a topic under investigation.[8] Primary sources could include documents, ledgers, municipal records, diaries, letters, newspaper articles, artifacts, etc. containing historical material or data. Researchers value being able to examine sources consulted by other researchers to verify the authenticity and veracity of the source, to learn more about a topic on which the source reports, to analyze the provenance of the data, and to see for themselves if they agree with another researcher's interpretation of the same material. Providing a comprehensive **citation** for each source used in the development of a dataset is a straightforward way to document the material behind the dataset. A robust citation should allow users to go directly to the original material to examine the quality of the extracted data, to dig deeper, or conduct parallel research on aspects of the source not included in the dataset.

As researchers know, citing historical sources of the data is one of the most important things to document. This need is intensified with a dataset. Every row or record within a dataset should include a reference to the source(s) of the data in that row/record. This allows the data creator and future users to know the provenance of the data and gives credibility to the overarching dataset. How a dataset creator records this source information can take one of two paths: a full citation for each source used in the creation of a single record within a field in that record, or separate **Source Records** that connect/link the source to every record derived from that source through the use of a Source Identifier. The advantages and challenges of these approaches are discussed in the following section. Irrespective of the approach, records in a dataset should all point to a complete citation for the source, one that is precise enough to allow an individual to find the original document, whether it be in a library, archive, museum, personal collection, website, etc.

Pro tip for archival sources: Include as many of the following pieces of information as possible to create a **full citation** for each archival source used in the dataset. A dataset creator can also collect this information in a series of fields in the dataset, rather than a single field.

- Name of repository
- Collection name
- Collection number

[8] https://www.historyskills.com/source-criticism/analysis/source-kind-and-type/ <Accessed 18 February 2021>

- Box number
- Folder
- File
- Page Number

A suggested organization for this is:
Title or description of item. Date of the item. Title of the manuscript collection. Number of the item. Name of the repository, Location of the repository.

The JSDP data article that accompanies each dataset will require that Sources be listed in the bibliography format of the *Chicago Manual of Style* (https://www.chicagomanualofstyle.org/tools_citationguide/citation-guide-1.html); they can be cited in the same way within the dataset using the guidelines for Manuscript Collections or the relevant source types.

For example,

- The Mayor's Register of Free Blacks in the City of New Orleans from 1840 to 1864. Volume 4, Book 1, Page 3. Louisiana Division/City Archives and Special Collections at the New Orleans Public Library.
- Inventories of slaveholders' possessions (1767-1831). Uncataloged box 1767. Arquivo Judiciário do Estado do Maranhão, São Luis, Brazil.
- *Maryland Gazette,* June 4, 1770, p. 6.

Create Fields for Citations in Individual Records or Create a Source Record

We recommend using one of the approaches described below to document sources within a dataset, not in a separate listing (file) alone. Connecting records in a dataset to the source(s) from which the data was derived will allow others to locate the source, verify the data, and examine the specific original material informing each point of data. Similar to scientists who release experimental data collected during their research to facilitate the reproducibility of research results, dataset creators extracting data from historical documents must provide robust documentation so that the data can be verified through comparison with the original document.

The first way to connect source information to records derived from a source document is to include Source fields within the dataset that would allow a user of the data to form a *full* citation for the source material. This row-by-row reference embeds explicit source data in every record. The provenance of the extracted data can be traced to the original material from a single field in each record. This may lead to redundant source data in every record; this is fine, as it will be clear what materials were used to create each and every record in the dataset without having to consult ancillary documentation files or records. Sensical abbreviations can be used so long as they are clearly documented in the field definitions table.

Alternatively, a dataset creator can create a separate **Source Record** to capture information about each of the historical documents used to create the dataset. The Source record would include the same handful of fields describing the nature, location, and provenance of the historical documents. The **Source Record** would additionally have a unique identifier for each source. This identifier would be added into a Source field of every record derived from that source. **An advantage of this method is that the identifier becomes a shorthand way (i.e. through the source identifier) to reference the source in every record without duplicating the same lengthy citation in numerous records.** Duplicate and redundant data can introduce errors and make it challenging to implement a small change to the citation across all of the records with that citation. An edit to a single Source record would be available to all other records that reference that record via the Source Identifier. More granular citation references like page number, box, or folder could also be included in other columns within the related record (i.e. in the Person, Event, or Place) if that level of specificity is desired.

For dataset creators who elect to make a separate Source record for each source used in the development of the dataset, Enslaved.org recommends that they use the following columns/fields to record the citation information for each source:

- Source Identifier: Create a unique reference for each source in the collection. (See the section on *"Identifiers"* for suggestions about how to create identifiers for your Source records.) Source identifiers indicate the sources from which Person records were extracted or where data about Events included in the dataset were derived. (Learn more about the importance of unique identifiers for all types of records in the *"Data Connections"* section of this document.)
- Document Type: Record the general category describing the historical document from which the data was extracted. Enslaved.org uses a controlled vocabulary to classify a wide variety of historical document types. This data conveys the nature of the source materials in ways that make it easier to search for and browse across similar datasets and records.

Person

In the Enslaved.org context, a **Person has biographical and identifying information about an individual, organization, or group of people acting as a single entity who participated or were involved in the historic slave trade**. The project promotes the use of controlled vocabularies and connections to events to show these people within an historical context. In doing so, it both keeps the ethical focus on people and makes finding and learning about their lives possible on Enslaved.org.

Names

As good practice, a dataset creator should consider including additional name columns to provide more granular representations of the data, like Surname, Given Name, Alternate Name, Honorifics, etc. This can improve searching and findability of the data within the dataset, along with indicating naming conventions in various cultures. Names can also change over time, so making a decision on which column of data contains the preferred name to reference should be included within the documentation. For example, Samuel Ajayi Crowther was born with the name “Ajayi” and later in life used the name “Samuel Crowther.” A general name column could contain the full name “Samuel Ajayi Crowther” and be used for a general reference to this person without giving preference to one name over another. Given Name could list Samuel and Surname could include both “Crowther” and “Ajayi.”

Enslaved.org uses “Unnamed” for any records where a person’s name is not identified. Using this approach will make it easier for reviewing the data (did I accidentally delete that name because the cell is blank?), but it will also make it easier to display the person record in any subsequently created data interfaces and further humanize the person by indicating that although this person’s name is unidentified, the person still existed and is important to record.

Age and Age Category

When recording details about a person’s age, consistency continues to be key. It is important not to conflate various types of information into a single column. When using numerals, have a column for Age that exclusively includes numerals. When using an Age Category like Infant, Child, or Adult, data collectors should log these in a separate column and stick to the controlled vocabulary that have been adopted and documented.

Sex

When logging the field Sex, it is critical to stay with the controlled vocabulary so that Sex and Age are not conflated. Again, the term “boy” can conflate the sex of Male with the Age Category of Child. While a data collector may want to include the verbatim term “boy” in a column of its own so as to capture how the information was recorded within the original document, Enslaved.org does recommend the data collector standardize and interpret/impute further information into additional columns. Having standardized language goes a long way toward allowing integration and subsequent discoverability of data when it is added to other platforms or used for research by others.

Relationships

Recording relationships between individuals within a dataset means capturing two related data points and attaching them to each individual record.

Examples of reciprocal relationship types include:

- Parent ⇔ Child
- Sibling ⇔ Sibling
- Spouse ⇔ Spouse
- Grandparent ⇔ Grandchild
- Aunt/Uncle ⇔ Niece/Nephew
- Cousin ⇔ Cousin
- Godparent ⇔ Godchild

To avoid confusion, Enslaved.org recommends leveraging the unique identifiers already created to make these connections and using a delimiter, i.e. any character not otherwise included within the data to differentiate between the identifier and the relationship data points. Relationship types should be explained in the data documentation.

For example, the record for Josefina, who has person identifier PER-000,1 could have a cell within a relationships column that contains the following relationships and delimiters:

PER-0002 | Parent; PER-0003 | Sibling; PER-0004 | Child

The pipe “|” denotes a connection between the unique identifier and a relationship type, while the semicolon “;” denotes a new set of relationship data points. Consistent use of any characters not found elsewhere within the dataset may be used in place of the pipe and semicolon.

In this example:

- PER-0002 is Josefina’s parent
- PER-0003 is Josefina’s sibling
- PER-0004 is Josefina’s child

The use of unique identifiers, consistent delimiters, and controlled vocabularies (all described within your documentation) allows the data to be easily interpreted by others examining the data.

This method is useful for computational consumption, but it can make the data much less human readable. People often solve this with a two-columns strategy: relationship text in one column and relationship id in a separate column afterward.

Birthdate and Death Date

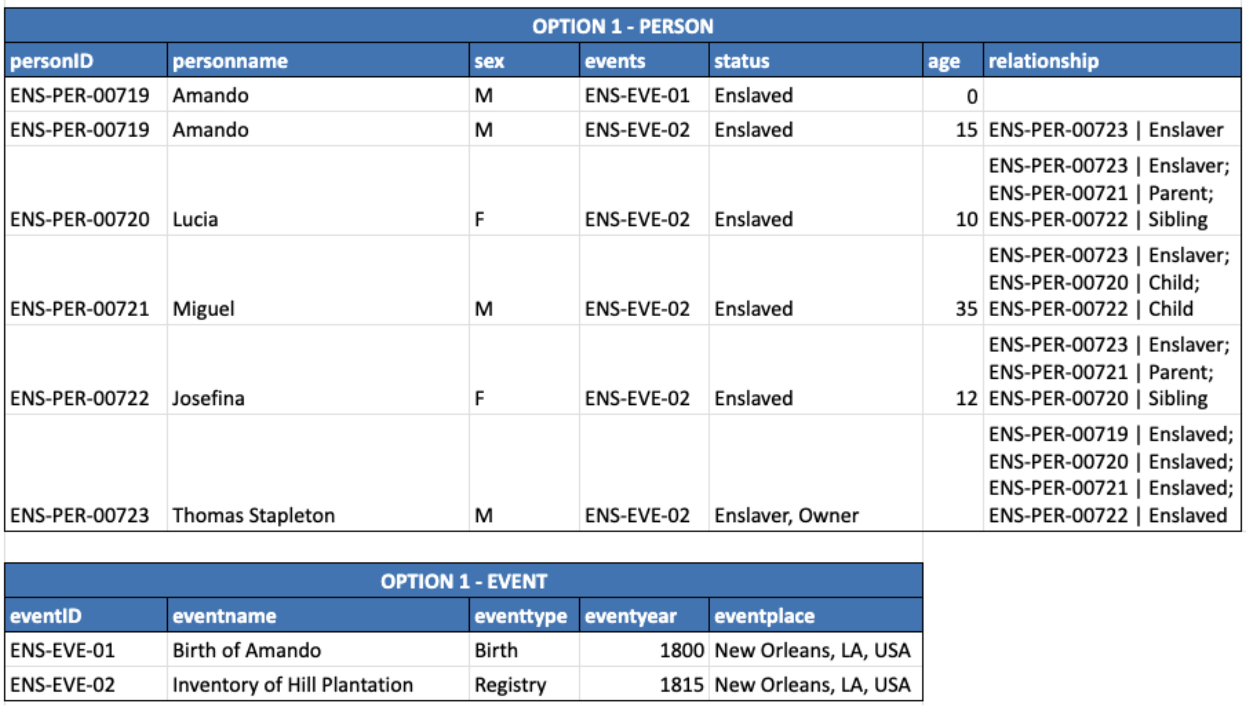
If the only Event information gathered about a person is their birthdate and death date, it makes sense to include these dates as part of the Person record, as these data points are specific to the human experience. However, if other information about the birth or death events has been collected, such as a description, place, other participants in the event, etc., Enslaved.org recommends that the dataset creator create separate Event records for birth and death.

Person Data Points at Specific Event Dates

Person information collected at a specific date can be represented in one of two ways; this is especially relevant for data points like Age or even Occupation that will change over the course of a person’s lifetime.

While Option 1 is the more commonly used for most datasets where multiple people participate in a single event (i.e. a group event), Option 2 may be more practical for datasets where individuals have multiple events, i.e., non-group events.

<u>Option 1</u>: The Person data is included within the Person record and linked to the Event through the Event Identifier. This data would be included within 2 tables or tabs within the same spreadsheet and are shown together below for comparative purposes.



OPTION 1 - PERSON						
personID	**personname**	**sex**	**events**	**status**	**age**	**relationship**
ENS-PER-00719	Amando	M	ENS-EVE-01	Enslaved	0	
ENS-PER-00719	Amando	M	ENS-EVE-02	Enslaved	15	ENS-PER-00723 | Enslaver

ENS-PER-00720	Lucia	F	ENS-EVE-02	Enslaved	10	ENS-PER-00723 | Enslaver;
						ENS-PER-00721 | Parent;
						ENS-PER-00722 | Sibling

ENS-PER-00721	Miguel	M	ENS-EVE-02	Enslaved	35	ENS-PER-00723 | Enslaver;
						ENS-PER-00720 | Child;
						ENS-PER-00722 | Child

ENS-PER-00722	Josefina	F	ENS-EVE-02	Enslaved	12	ENS-PER-00723 | Enslaver;
						ENS-PER-00721 | Parent;
						ENS-PER-00720 | Sibling

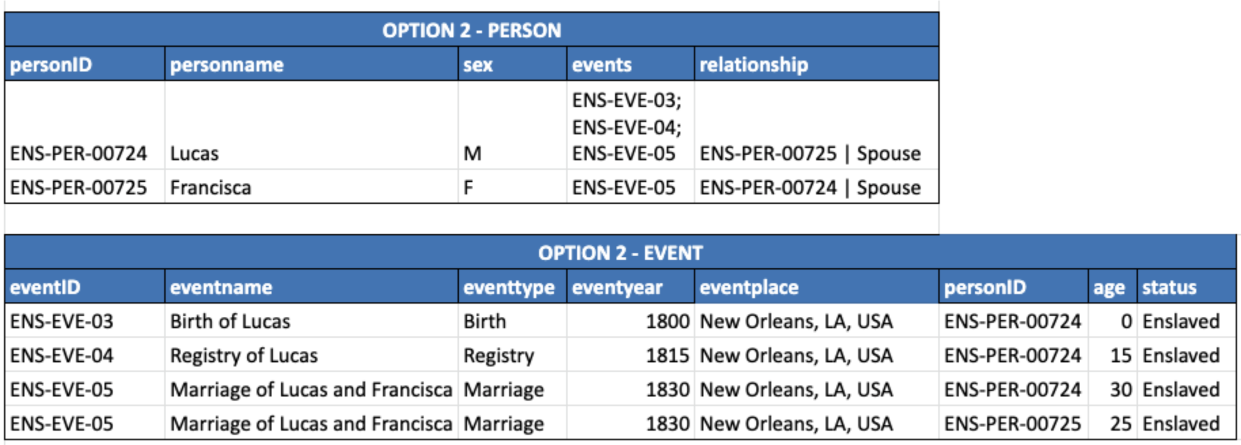
ENS-PER-00723	Thomas Stapleton	M	ENS-EVE-02	Enslaver, Owner		ENS-PER-00719 | Enslaved;
						ENS-PER-00720 | Enslaved;
						ENS-PER-00721 | Enslaved;
						ENS-PER-00722 | Enslaved

OPTION 1 - EVENT				
eventID	**eventname**	**eventtype**	**eventyear**	**eventplace**
ENS-EVE-01	Birth of Amando	Birth	1800	New Orleans, LA, USA
ENS-EVE-02	Inventory of Hill Plantation	Registry	1815	New Orleans, LA, USA

Illustration 2: Option 1 for recording person data points at specific Event dates

In the example above (Option 1), data for Amando, including his Gender (**M**), Freedom Status (**Enslaved**), and Age (**15**), are recorded in the Person record. Amando's Person record also includes a reference to an event, a **Registry**, through the Event identifier (**ENS-EVE-02**). More granular information about the Registry, including the name (**Inventory of Hill Plantation**), data about when (**1815**) and where (**New Orleans**) the inventory was taken, are represented in the associated Event record. Additional Event records could be included for Amando by duplicating his personID and connecting other Event records (**ENS-EVE-01).**

<u>Option 2</u>: The Person data is included directly within the Event record and linked back to the Person through a listing of the Person Identifier.



OPTION 2 - PERSON

personID	personname	sex	events	relationship

ENS-PER-00724	Lucas	M	ENS-EVE-03;
			ENS-EVE-04;
			ENS-EVE-05
				ENS-PER-00725 | Spouse

ENS-PER-00725	Francisca	F	ENS-EVE-05	ENS-PER-00724 | Spouse

OPTION 2 - EVENT

eventID	eventname	eventtype	eventyear	eventplace	personID	age	status
ENS-EVE-03	Birth of Lucas	Birth	1800	New Orleans, LA, USA	ENS-PER-00724	0	Enslaved
ENS-EVE-04	Registry of Lucas	Registry	1815	New Orleans, LA, USA	ENS-PER-00724	15	Enslaved
ENS-EVE-05	Marriage of Lucas and Francisca	Marriage	1830	New Orleans, LA, USA	ENS-PER-00724	30	Enslaved
ENS-EVE-05	Marriage of Lucas and Francisca	Marriage	1830	New Orleans, LA, USA	ENS-PER-00725	25	Enslaved

Illustration 3: Option 2 for recording person data points at specific Event dates

In the Option 2 example above, Lucas participates in *multiple* Events, including a Birth event (**ENS-EVE-03**), a **Registry** event (**ENS-EVE-04**), and a **Marriage** event (**ENS-EVE-05**). In each event, which happens on different dates, Lucas has a different age (**0** at birth, **15** when he was registered in the Registry, and **30** years old when he married Francisca). Therefore, capturing this person data in the event record (instead of in the person record) allows one to collect how old Lucas was at different life events. This can be accomplished in a similar way using Option 1 by repeating the Person information (like Name, Sex, and Person Identifier) over and over again in multiple rows.

While no one way is right or wrong, the decision should be made based on what information the dataset creator is prioritizing in representing historical data. In general, when recording a *single* event for a person, including the person data within the Person record (Option 1) probably makes the most sense. When collecting *multiple* Events about a person that have multiple data points within each event, Option 2 should be considered.

Differentiate People with Unique Identifiers

Using a unique Person Identifier goes a long way toward aiding in the creation of a single record or series of linked records leveraging a single identifier for a Person. It is important that identifiers be applied to all Person records within a dataset. While the goal of a data collector is often to record the names of enslaved people, they often capture the names of the person's enslaver(s), but only identify the enslaver as a datapoint tied to the enslaved individual. This leaves enslavers without their own identifiers, but adding identifiers for them is beneficial as well. With the enslaver data embedded in another person's record, it is difficult to answer, *How many people did Thomas Stapleton enslave?* An identifier must be specified for each listing of Thomas Stapleton who is known to be the same person. If an identifier is not specified and present, it cannot know if the enslaver Thomas Stapleton named in one enslaved person's record is the same Thomas Stapleton logged in another record. The same concept applies to any named participants within an Event (Notaries, Judges, Buyers, Sellers, Shippers, Concessionaires, etc.). In short, use identifiers to connect the same person to multiple Events and People.

Connection to Event

There are various ways to connect Person data to an Event record, as discussed in "Person data points at specific Event dates" above. What is most important is to make sure that the Person data is linked to or references any related Events. Listing the related Event Identifiers within the Person record is often the most simple and straightforward way to make sure that the Person data is tied to a place at a specific period in time as Event records typically include both temporal (time) and geographic (place) data points (please see Illustrations 2 and 3 above, along with "Connection to Place" section under Event below).

Event

In the Enslaved.org context, an **Event is information about a single incident or occurrence that led to the recording of information**. The project promotes the use of controlled vocabularies and connections to place and time to locate these events within an historical context and to make them easier to search and browse across datasets. The full list of event type controlled vocabularies can be found at https://docs.enslaved.org.

Event Type

The purpose of the Event Type field is to categorize events based on their overarching characteristics or outcome. Often this field is absent from datasets because information about the type of event is implied by the type of document from which the data is extracted. When made explicit in Enslaved.org, event type allows the project to group together all records by type of event irrespective of originating dataset or contributor. Therefore, users can see all of the baptisms records for enslaved people, for example, from anywhere in Enslaved.org's corpus of datasets. Enslaved.org encourages all dataset creators to consider including the Event field and using Enslaved.org controlled vocabulary terms for Event Types.

Single-Person Events and Group Events

It can be challenging to decide how many Event records to create for various event types. A single-person event is an Event where only one person is known to have participated. As discussed previously, it may be problematic to create an event for a birth or death when one only has a date. If this is the case, many data collectors include this information in the Person record instead. When Enslaved.org integrates birth and death dates into its discovery hub, it creates an event record for each Event and connects the Person to that Event. If a data collector has Date and Place information for an Event or any other additional Event information, Enslaved.org recommends creating a separate Event record with an Event Identifier.

Group events are any Events that clearly include multiple participants, like marriages, plantation inventories, voyages, etc. Details about the event can be recorded one time in an Event record and then be referenced (via an Event Identifier) in all of the Person records of people who participated in the event. In the case of a marriage, the Event Identifier for the marriage would be added to the Person records for the bride, groom, priest/clerk/officiant, and any other event participant. (See the section on “Identifying Records” for suggestions about how to create identifiers for your Event records. Learn more about the importance of unique identifiers for all types of records in the “Data Connections” section of this document.)

Dates

Date fields are used to express when an event happened. Use one date field for events that occured at a single point in time. For events like a voyage that occur over a period of time, Enslaved.org recommends using two date fields, one for the date when the event began (Start Date) and a second for the date when the event ended (End Date).

Date data should be formatted consistently within a dataset. There are several published date formats standards you might consider implementing, including ISO 8601-1, W3C Note on Date and Time Formats, and the Extended Date/Time Format Specification. Enslaved.org recommends using a string of YYYY-MM and YYYY for incomplete dates and YYYY-MM-DD for full dates; MM-DD-YYYY and DD-MM-YYYY are not recommended. It is vitally important to explain date format in the datasets’s field definition table.

Date fields are a good example of when not to mix data types within a field. Dates should *not* include non-date data, for example “approximately” or “circa” or “c.” For approximate or ambiguous dates, you might consider one of the following approaches.

- Option 1: Add another field for recording terms that qualify the date, including circa or approximately for ambiguous dates in the date column. Use the date qualifier field for even more vague or imprecise dates, for example to note when an event happened “before” or “after” a date or even a year. Events that are described as having happened “Before” and “After” other events with specific dates are commonly found in oral testimony or life histories/biographies, personal narratives, and testimony.
- Option 2: Add new date fields to capture data for each ambiguous type of date, for example Circa Event Date or Occured Before Date. As long as the exact meaning of the data included in the field is documented, this approach is reasonable. A downside is that it increases the number of fields in the dataset, which can become unwieldy and cause confusion about which field is the appropriate one for a specific event.

Connection to Place

There are various ways to make the Event data connect to a Place record, as mentioned in the “Person data points at specific Event dates,” “Birthdate and Death Date,” and “Connection to Event” sections above. What is most important for Event is to make sure that the Event data is connected to any related Place records. Listing the related Place Identifiers (described in the next section) within the Event record is often the most straightforward way to make sure that the Event data is tied to a Place or Places. If an event happened in more than one place -- like a ship voyage that began in one port and traveled to several other ports before reaching its destination -- include multiple Place Identifiers in the Place field of the Event record, one for each of the pertinent places. Document and use a set of standard delimiters between the multiple Place Identifiers. (Learn more about delimiters in the “Data Connections” section and see an example of their implementation in the Person record “Relationships” section.)

Place

In the context of Enslaved.org, **a Place is information about a location where an Event occurred**. Using controlled vocabularies and connections to date to locate these places within an historical context and to make them easy to search and browse on Enslaved.org.

Place Type

The purpose of the Place Type field is to categorize a place based on its overarching characteristics to promote the findability of similarly named places. Often this field is absent from datasets because information about the type of place is not explicitly expressed within an original source document or the whole dataset refers to a single place type. Including Place Type allows Enslaved.org to group together all records by type of place, for example, a court, irrespective of originating dataset or contributor. Therefore, users can see all of the court records for enslaved people from anywhere in the Enslaved.org discovery hub. More information on the controlled vocabularies relating to Place Type can be found at https://docs.enslaved.org.

Coordinates

The inclusion of geospatial coordinates, if knowable, is also useful data to add to your Place records, though not required. Geospatial coordinates not only allow for better definition of your Place record, they also allow Enslaved.org to link to other Place databases like GeoNames[9] and Wikidata.[10] These linked open data points provide further information about places included in your dataset. Inclusion of these identifiers can be as simple as adding an additional column for each with the connected identifier. For the example of New Orleans, the Wikidata identifier https://www.wikidata.org/wiki/Q34404 would be Q34404, and the GeoName identifier https://www.geonames.org/4335045/ would be 4335045; data collectors should be sure to document which method is being used. This concept can also be applied to any connections to other Person or Event records found within Wikibase.

Contextualizing Place through “Located In...” Statements

Enslaved.org recommends that you create a Place record for the most granular level of place data in your dataset. If, for example, a contributor wanted to state that a baptism or wedding took place at a specific church, the dataset creator would make a record for that church, using a unique identifier and a series of additional columns to indicate where that church was located.

[9] https://www.geonames.org/

[10] https://www.wikidata.org/

This would take the form of “located in” columns for place data, for example: located in Oglethorpe (ward), Savannah (city), Georgia (state), United States (country). The dataset would include columns for each of these data points, i.e. ward, city, state, country, etc.

Data connections

The use of consistently implemented unique identifiers (see “Identifying Records” section above) not only allows for precise reference to an individual record; it also allows for concrete linking of records to one another, even if they are in different datasets.

The “Relationships” section above provides a tangible example of why and how to use unique identifiers along with standardized delimiters (see the consistent use of delimiters “|” and “;” in the example below).

To restate the relationship example, the record for Josefina, who has person identifier PER-0001 could have a cell within a relationships column that contains the following relationships:

PER-0002 | Parent; PER-0003 | Sibling; PER-0004 | Child

In this example:

- PER-00719 is Josefina’s spouse
- PER-00721 is Josefina’s parent
- PER-00722 is Josefina’s sibling

The pipe “|” denotes a connection between the unique identifier and a relationship type, while the semicolon “;” denotes a new set of relationship data points.

This method can be put into practice for any type of data connections throughout your dataset. For Enslaved.org, data connections can be made from Person to Event, Event to Person, and Event to Place; all data points are connected to Source.

Illustration 3 above (see “Person data points at specific Event dates”) shows how to connect a Person record to multiple Events records by consistently using Event Identifiers and a delimiter to specify the boundary between identifiers.

In the illustration, Lucas, who has person identifier ENS-PER-0074, participated in three events, which are recorded in the following manner in the Event column of the Person record:

ENS-EVE-03; ENS-EVE-04; ENS-EVE-05

In this example

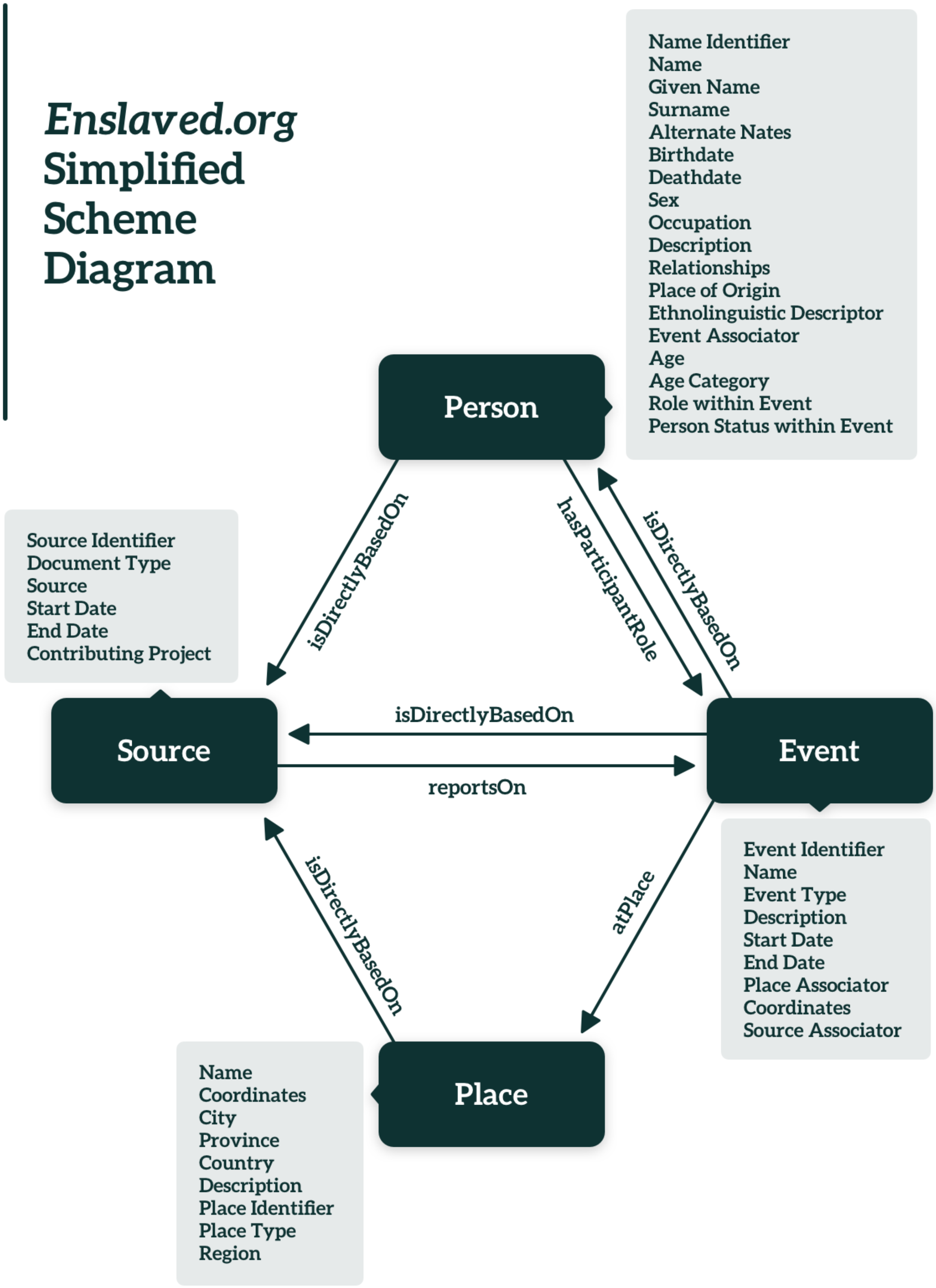
ENS-EVE-03 is Lucas’s Birth event
ENS-EVE-04 is a Registry event that identified Lucas as an enslaved person

ENS-EVE-05 is Lucas’s Marriage to Francisca

These three events are included in Lucas’s Person record using the unique identifier for each Event separated by a semicolon “;” and a space. Consistently using a delimiter like a semicolon followed by a space distinguishes in a standardized way where one Event data point ends and the next one begins. This can make it easier for the data collector to see that the appropriate data is recorded and will also allow for the data to be more easily parsed in the future.

The following illustration shows how Enslaved.org makes connections between the different datasets.

The arrows in the illustration indicate a connection between two different datasets. There are arrows pointing from person, event, and place to source, indicating that the information about people, places, and events come from a source. This source could be the same for all records in a project or different for the records in different datasets. There are also arrows pointing from person to event and from event to place. These arrows signify that people participate in events and that events occur at certain places. While there is no arrow pointing directly from person to place, these two types of data can be connected through an event. In other words, a person participated in an event that happened at a specific place.



Enslaved.org Simplified Scheme Diagram

Person

Name Identifier
Name
Given Name
Surname
Alternate Nates
Birthdate
Deathdate
Sex
Occupation
Description
Relationships
Place of Origin
Ethnolinguistic Descriptor
Event Associator
Age
Age Category
Role within Event
Person Status within Event

Source

Source Identifier
Document Type
Source
Start Date
End Date
Contributing Project

Event

Event Identifier
Name
Event Type
Description
Start Date
End Date
Place Associator
Coordinates
Source Associator

Place

Name
Coordinates
City
Province
Country
Description
Place Identifier
Place Type
Region

Illustration 4: The Scheme Diagram for Enslaved.org.

Acknowledgments

Thank you to Kristina Poznan, Sharon Leon, Ryan Carty, Austin Truchan, Marisol Fila, Daryle Williams, Walter Hawthorne, and Dean Rehberger for their contributions to this guide. We also thank the contributors to Enslaved.org to date, whose dataset contributions have informed our practices and helped hone our recommendations.