Folder Contents:

Diagrams:

* Various documents that depict the construction and intellectual reality of the metadata mapping and dataset construction.

Datasets:

* As exported from LC using JSON queries; CSV delimited version
  + These datasets can be shared/disclosed to the public
  + JSON Queries document enclosed
* Final NMAAHC dataset, normalized.
* Final combined dataset with LC & NMAAHC collection items.

Final Dataset:

* Final dataset that combines both LC and NMAAHC’s baseball collection items. This dataset includes normalized fields such as Object Type (format), as well as those data points that are permissible for publication. The column headings are the common field names determined to equally represent and describe the combined collection items.

Metadata Crosswalk/Map

* The Inside Baseball Crosswalk is an Excel spreadsheet that documents the metadata mapping across, Marc, JSON API, NMAAHC, DublinCore to the final common field names for these combined collection items. This document serves as the key or legend for understanding the evolution of data to its final state.
  + The individual institutions’ datasets have notable differences related to the differing cataloging processes between libraries and museums.
  + While Marc data was not directly exported, it is indirectly found in the JSON export, it provided a starting point for the metadata map and the uniting foundation for the other schemas incorporated.
  + DublinCore is provided as an additional schema within the map.

Translations:

* Object Type & Secondary Object Type:
  + Values found in each institutions’ multiple metadata fields recording format or medium were normalized into these two fields. Compromises were made by both institution and at both the high-level classification and specific object type category. For example, the high-level category for “baseball cards” within NMAAHC’s data is “Memorabilia and Ephemera” a term commonly found within museum cataloging classification. Users and researchers may not know to seek out this classification category in order to get to the lower-level term of “baseball cards.” Therefore, common names were inserted where logical. In this example, maintaining Memorabilia and Ephemera as a high-level classification found in the Object Type field unites LC collection items as well. The various lower-level and specific terms found within Memorabilia and Ephemera include not only baseball cards but also pennants, buttons, admission tickets, autographs and advertisements.
  + Another intensive example between Object Type and Secondary Object Type lies within the Media Arts-Photography term found in Object Type. While the Library’s cataloging is direct to the specific term, the Museum’s descriptive information works at various tiers to arrange the specific collection items, in this case within the genre of photography. The Library items that are from the photographic collection are placed under the high-level category of Media Arts-Photography, while the Museum’s items use the Library’s specific terms in various instances. One example is drawing up the Museum’s medium from “photomechanical print” to photographic print as the final Secondary Object Type value. This decision again references the need to utilize commonly understood terms that are more likely to be searched and understood by a general user or researcher of the data.
* Copyright
  + The translation for the copyright field was less intensive than the Object Type(s). More than not, the decision was already made for the various data stakeholders upon inspection of the data values for this information. The Library of Congress utilizes a simple but clearly understood formula of true/false to state the application of copyright restrictions. These values were translated to “Copyright Restricted” and “No Known Copyright Restrictions” for true and false, respectively. Both of these translated values were found within NMAAHC’s native copyright values, and because the Museum has multiple fields describing the copyright and restrictions for their objects, maintaining these values in the final dataset was the logical responsible decision made for the users and institutions.

Data Field Additions:

* Institution:
  + While the objects’ links included the institutions’ prefixes for their URLs (nmaahc.si.edu & loc.gov), there was only occasional reference to the institution in possession of the baseball collection item in the data as a whole. The addition of the “Institution” field was the outcome of this observation and could serve as a filter or search facet as well.
* Era:
  + Each institution had dates associated for a majority of their collection items, there was no common format to this date field that could readily unite the dates. The additional feature of circa dates throughout generated the outcome of Era field(s) to unite collection items by decade(s) and a date range.

Abundance of Fields with Same Names (e.g. Subject, Location):

* Subject:
  + Each institution collected data related to the general and specific subject matter of the item or object. However, each institution cataloged and thereby recorded this information in different and unique fields that resulted in a many-to-many mapping. Without considering these subjects or completing the metadata map in a linear fashion, similar subjects such as “segregation” or named individuals like “Jackie Robinson” would have not been united in the same field. Without parsing the multiples values in each institutions’ various subject fields and without cross mapping these distinct values to a common and united field, users of this data would be set-up to miss all the collection items related to the same subject matter. This end result would have left the institutions’ collections items just an independent as they were before any attempt at unification.
  + The result of this effort was the NMAAHC “Attributes-Object Type” values mapped and parsed to the Object Type fields, while other values also parsed out of the “Attributes” field included names, subjects, creators, publishers and manufacturers. While the Library’s subject headings are separated in the JSON raw export, they required extraction to pull out named individuals for distinction to align with the Museum’s “Constituents” field that records the passive or related subjects. The arrangement of fields in the final dataset represents the extent necessary to unite these collections within their subject matter.
* Creator/Publisher:
  + This field was multiplied in accordance with the multiple values presented within the Library’s raw data. The Museum only has one value mapped to the primary Creator/Publisher field and that value originates in the “Constituents” field as the active maker or “created by” value that followed this notation.
* Location:
  + Upon the LC JSON export, multiple fields titled Location\_001, Location\_002, etc. can be found in every collection item. While these location fields can be traced back to the Marc cataloging, the Museum employs a different cataloging approach. Within the Museum’s metadata the Geography field is an all encompassing and inclusive field to record at the lowest level a very specific street address and rising up to the region and country represented by the collection object. The Museum places this data into one field and required parsing to align to the LC data. The original Geography field is also maintained in this instance of many-to-many mapping to offer a check for users of the dataset and to ensure the accuracy of the parsed but corresponding data (the listed data in one field is followed consecutively and respectively by any corresponding location such as street or locale to city to state to country).
* Collection Title:
  + The Library’s cataloging procedures and digital collection organizational methodology records collection-level information such as “Branch Rickey Papers.” The Museum utilizes a similarly named field but not as a standard or regular field of information for an organizational or arrangement methodology for collection objects. Where present, NMAAHC’s collection title information is available, however this field in the final dataset is predominantly to offer more descriptive information related to LC collection items.

Other fields:

* Object Number:
  + Presented as the identification for NMAAHC collection object. There is no data point within LC information that corresponds. Note: one could consult the Digital ID link, which ends in the LC digital collection item ID but does not necessarily correspond to a catalog identification number as is the case with NMAAHC’s object numbers. NMAAHC’s object numbers can also be found in the Digital ID.