AMNH Library
Filenames for EAC-CPF entity records
June 24, 2015

Background
In 2012, the Research Library of the American Museum of Natural History (AMNH) was awarded a grant from The Andrew W. Mellon Foundation, through a program administered by the Council on Library and Information Resources’ (CLIR) Cataloging Hidden Collections. The AMNH Archive Project will produce in-depth descriptions, called finding aids, for major archival collections relating to Museum expeditions. The project will also result in brief histories of these expeditions and biographies of those who participated in them.

These histories and biographies, called entity records or entities, are encoded in the descriptive standard EAC-CPF, Encoded Archival Context – Corporate bodies, Persons, and Families. Currently the project has produced 2,466+ personal name entity records and 703 expedition entity records. Of these, about 80-100 are described in rich detail. The remaining records have minimal level description – birth and death dates with brief notes about affiliation to the museum for persons, places and personnel along with purpose of research for scientific expeditions.

The data for these records were recorded in spreadsheets and converted to EAC-CPF xml using macros and Python scripting (written by Nick Krabbenhoeft). The xml files will be imported into xEAC, an open-source application for creating, managing and publishing EAC records.

In addition to describing expeditions and people, minimal entity records were created for AMNH itself, Departments, Permanent Halls and Temporary Expeditions. Spreadsheets were customized for each set with data fields mapping to EAC-CPF. Conversions will be made to create additional xml files.

Note: The terms filename and identifier are used interchangeably in this document because the filename for entities is derived from the record’s unique ID.

Filenames Structure for EAC Records
Master spreadsheets were created for each category of entity record:

- Personal Names
- Expeditions
- Departments
- Permanent Halls
- Temporary Exhibitions

With the exception of Permanent Halls\(^1\), each record corresponds to a single line in its respective spreadsheet file. The descriptive fields vary from one data set to another. However each set includes a field for a unique identifier (eaccpf:recordId).

\(^1\) Unlike the other data sets, Permanent Halls consists largely of variant names spanning the early 1900s through the present. Each name corresponds to a building number in the museum. Several spreadsheets were created to represent each floor, with separate sheets for each section on its respective floor. Raw data was taken from in Valerie Thaler’s spread sheet of Hall names found in AMNH publications. More information about this her work can be found at: http://images.library.amnh.org/hiddencollections/resources/amnh-permanent-halls/
The unique ID string consists of an alphanumeric sequence that represents the authoring institution (amnh), the type of entity (p for person or c for corporate body), an underscore followed by a numeric code indicating the category (person vs. expedition vs. museum department, etc.), and finally a six-digit numeric value assigned in the order they appear in their master spreadsheet beginning with 000001.

The resulting identifier for a person entity record looks like this: amnhp_1000001
The museum as an entity has the identifier: amnhc_0000001

The chart below breaks down the components in their respective categories:

<table>
<thead>
<tr>
<th>Types of entity records at the AMNH</th>
<th>alpha code</th>
<th>numeric string (7 digits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Museum as a whole</td>
<td>amnhc_</td>
<td>0000001</td>
</tr>
<tr>
<td>Personal names</td>
<td>amnhp_</td>
<td>1000001</td>
</tr>
<tr>
<td>Expedition names (corp body)</td>
<td>amnhc_</td>
<td>2000001</td>
</tr>
<tr>
<td>Department names (corp body)</td>
<td>amnhc_</td>
<td>3000001</td>
</tr>
<tr>
<td>Hall names (corp body)</td>
<td>amnhc_</td>
<td>4000001</td>
</tr>
<tr>
<td>Temp exhibits (corp body)</td>
<td>amnhc_</td>
<td>5000001</td>
</tr>
</tbody>
</table>

Here is the rationale for this particular design:

**Why “amnh” in the filename?**
The AMNH Library anticipates the publication and use of entity records as linked data. The records will have URIs locally assigned, and they will also be submitted to an international authority cooperation, SNAC. Embedding the institution name into the identifier immediately ties the record to the author, AMNH. This is important for referencing within an aggregated description, such as SNAC (rdf:sameAs). It will also immediately identify the museum should external descriptions link to our records.

**Is “c” or “p” necessary?**
No, but again, thinking about outside users and possibly file management, we thought it would be helpful to embed this information right into the filename. Corporate body and Persons are general enough categories to persist even if the encoding metadata standard (EAC-CPF) changes in the future.

**The underscore**
Not semantically meaningful, the underscore breaks up the alphanumeric code in a way that a person can easily read the name. *Is that the letter “O” or a zero?* It may also be useful when running batch actions -- the files can be isolated in a few different ways.

**Adapting a naming convention from the AMNH Library negatives**
Similar to the way the library negatives were numerically coded by size, the different categories of entities were assigned a number by type of entity as it pertains to the museum. In the same way that the initial “3” in negative number 314328 represents an 8 x 10 negative, the initial “3” in amnhc_3000012 represents an AMNH Department.
Why stop at six digits?
Seven digits altogether make up the numeric string, leaving a six-digit expansion for the sequential number assignments. If we ever have more than 999,999 in any one category, we may need to rethink this structure. For now, projecting under a million entities per category was a safe move.

How do you assign new numbers? And, how can you prevent the duplication of a unique identifier?
As mentioned, the data was initially captured into spreadsheets with each line of the spreadsheet assigned to a single entity (except Permanent Halls). First the alpha code was designated to the sheet, followed by the underscore. The first entry was assigned “…000001”. From there the remaining cells in the ID column were filled sequentially using Excel: Fill Series command. New records are added to the end of the sheet (or sorted accordingly) and are assigned the next sequential number.

While there is no guarantee that there will not be an occasional duplication, keeping the data in sequence by record ID will help to prevent dupes. There is an Excel spreadsheet that tracks the IDs from all the master spreadsheet files. It is (and should be) used in tandem with xEAC to assign IDs in the future. xEAC will overwrite a file with the same name. Special attention should be given when adding this data into the application. Keeping a separate file log of assigned IDs will be useful for this purpose as well as clearing up accidents. *Uh-oh! Who was amnhp_1002563? I meant to type 3563!*

Legacy Excel files
Once imported into xEAC, the original spreadsheets should be archived and treated as legacy documents. The entity records should be edited and managed in xEAC and the legacy spreadsheet data can be used as reference.