Broadcasting A/V Data Projector Design Schematic

Related documentation:

- 2-Data model

This document outlines the steps that have been or will be taken to create the 'projector' Airtable base for the Broadcasting A/V Data project. While this project builds on the work of Unlocking the Airwaves and thus uses data from its Airtable base, three other collections of broadcast radio programs are also used, meaning three other 'source' bases had to be created before the projector could come together. Accordingly, the first few sections of this document cover the data cleanup and extraction steps to create these other bases, as well as the synced views that have ultimately come to fill in the projector base.

Source bases

In the source bases, the table that ultimately is synced into the projector is the CPF Authorities table, where the authorities tied to each program are listed as the key field. A view specifically for syncing has been created in each source table and has been locked with the note "Locked for syncing with BAVD base" so that the records and fields in the projector are stable.

Airwaves Application Base

NOTE: This base is simultaneously being used as the data source for the Unlocking the Airwaves project, and is thus much more filled out. For the BAVD project, some of this existing expanded authorities data will be used, but for initially building the projector, only the Airwaves landing page URLs, Wikidata IDs, and SNAC IDs were synced in beyond the identifiers of linked content.

Original data format: 'CPF Authorities' table in 'Airwaves Application Database'
Data preprocessing and cleanup: None due to simultaneous use for Airwaves backend–additional cleanup will come in synced base so as not to cause disruptions on the Airwaves site
View synced fields in BA/VD projector base
Synced table title: NAEB CPF Authorities

KUOM (UMN) Metadata

Original data format: JSON file with metadata records for each program
Data preprocessing and cleanup:

- Using OpenRefine, parse JSON file to table
- Wrap in quotes values containing commas
- Join multi-valued cells using ", "

- Export table as CSV
- Set up Airtable base for program records, creating fields for all JSON fields that are needed. Use field descriptions to match exact field names from JSON file to human-readable field names in Airtable. (Can't just import CSV as new table due to file size limit–must create fields first then use CSV import app to import data)
- Duplicate creators and contributors fields and convert duplicate to linked fields, linking both to new table 'KUOM CPF Authorities'
- Rename linked fields in new table to indicate whether links are as Creator or Contributor
- Duplicate names column in CPF Authorities table and title it something like 'cleaned names', then export both columns of names as CSV
- In OpenRefine, clean up CPF names by first separating roles out to their own column with just roles, then clearing dates, whitespace, and other minor errors. Wrap any roles that contain commas in quotes, then export CSV
- Make new blank field in Airtable for CPF roles
- Re-import, matching original names to key field and putting cleaned names in cleaned name column. Duplicate original names field and title it 'Original CPF Name'
- Copy cleaned names over on top of key field (which currently has original names). Delete the cleaned names field once cleaned names are occupying key field
- Convert role field to multiple select type.
- Use a formula to wrap the values of 'Original CPF Name' in quotes, then convert to multiple select field type. Delete the plain text 'Original CPF Name' field and rename the multiple select field to the same name.
- Dedupe the KUOM CPF Authorities table with the Airtable dedupe app, doing all exact matches and some similar (if differences are clearly orthographic, such as capitalization or missing periods). During the dedupe, be sure and retain the linked programs attached to both authorities, as well as the 'Original CPF Name' and role field values from both authorities, in case there is need to undo a merge later on.

View synced fields in BA/VD projector base
Synced table title: KUOM CPF Authorities

NFCB (UMD) Metadata

Original data format: CSV of program records
Data preprocessing and cleanup:

- Import CSV as new table in Airtable (can do directly, without using CSV import app)
- Duplicate and copy around data so that identifiers are key field instead of handle URLs
- CPF authorities are separated by a semicolon and no space in this data–use formula field to replace all semicolons with "," (quotation mark, comma, quotation mark) and place quotation mark at beginning and end of entire string in cell. This should result in string that functionally has each CPF authority wrapped in quotes, and separated by commas, which Airtable will correctly read when converting to linked field. Do this on Contributor, Provider/Publisher, CorpSubject, and PersonalSubject, keeping both the original and formula fields. (Creator field is fully blank in this dataset)

- Duplicate formula fields and convert them to linked fields, linking all four to new table 'NFCB CPF Authorities' and renaming them to indicate which kind of relationship the link is (Contributor, Provider/Publisher, CorpSubject, PersonalSubject)
- On CPF Authorities table, use formula field to find any names with a leading space and fix them. Then, use formula fields to remove dates from the ends of names (can also use OR)
- Create dupe of cleaned up names that is wrapped in quotes, then convert that to multiple select field type. Name the field 'Original CPF Name'
- Use dedupe app to dedupe all exact matches and any similar matches where the differences are orthographic. At this point, leave any similar matches for which you might have to do research to determine if they are true dupes. Keep the original names by preserving both records' original CPF name value in that field

View synced fields in BA/VD projector base
Synced table title: NFCB CPF Authorities

WHA (UW-Madison) Metadata

Original data format: MODS XML program records
Data preprocessing and cleanup:

- With developer assistance, parse needed fields from XML records to CSV
- Import the CSV into Airtable, keeping all fields as single line text/long text to start
- Use a formula field to construct the stable URL for each program record
- Use a formula field to strip the prefix off of the identifiers for the records, then duplicate the original key field that contains the prefixes; finally, copy the stripped identifiers into the key field–these are now the identifiers
- Duplicate the CPF Authorities and CPF Subjects fields, then convert the duplicates to linked field, linking them to new CPF Authorities table and adding LINK to the field name
- Convert the Date created, Topical subjects, CPF subjects, Geographic subjects, and Series fields to multiple select type
- In the CPF Authorities table, use a formula field to create a copy of the key field wrapped in quotes
- Create 2 new multiple select fields, one for CPF creator/contributor links and one for CPF subject links–these are to preserve the original CPF name string for reference
- Filter the view to only authorities that are linked as a CPF role (not subject), then copy the full CPF string with quotes over into the new multiple select field for creator/contributor links
- Filter the view to only authorities that are linked as a CPF subject, then copy the full CPF string with quotes over into the new multiple select field for CPF subjects
- Use formula fields to extract the roles and LOC URIs off of the CPF strings in the key field, storing roles in a Role multiple select field and LOC URIs in a single line text field (unless there has been an error, even after deduping you should only have one or zero LOC URIs per CPF)
- Once the key field only contains names, remove any lingering leading or trailing spaces and dedupe on exact matches, making sure to preserve all links to programs, roles, and full CPF strings from all the merged duplicates

View synced fields in BA/VD projector base
Synced table title: WHA CPF Authorities

Ingesting Data Updates

General categories and workflow

1. Category 1 (most complex to update): Programs metadata updates that include changes to CPF authorities
    - This category includes any situation where the creators/contributors/CPF subjects have been altered, corrected, added to, or pared down in the program records from any of the 4 source collections. This would also include any updates where entirely new program records are added.
2. Category 2 (somewhat complex): Programs metadata updates that does NOT include changes to CPF authorities
    - This category may include changes to subject terms, series information, or other descriptive metadata. This is only complex in that pieces of descriptive metadata may be used to organize or enhance CPF data, so we don't want to import updates that wipe out useful metadata unless it was incorrect in the previous version of the data.
    - Depending on what descriptive metadata fields were changed (if known), this type of update can either be CSV imported directly on top of the old data, or by adding a duplicate field of the original metadata for comparison.
    - For example, if the series title field was updated, it might be prudent to create a duplicate of that field, then CSV import the new data onto the duplicate field to see what changed, to ensure that nothing important was lost. However, if the updated field is something like broadcast date, it is not necessarily to compare the new and old data.
3. Category 3 (simple): Programs metadata updates that only change URLs or add entirely new metadata fields without altering old ones.
    - This category may include new stable/persistent URLs for programs, or additional metadata fields that have been generated (such as stable URLs for radio series).
    - This kind of update can just be CSV imported into the source base

Workflow:

- Do any preprocessing of the data like was done for initial ingest (assuming data takes same format as initial ingest)
- Ingest new programs data into source base
- Extract new authorities and disambiguate/merge in differences between old and new in source base
- Link new authorities to combined list of BAVD CPFs
- Disambiguate/merge in new authorities in BAVD CPFs
- Update Crit-Index and Crit-Rec-Enhance lists
- Import updated program display criteria fields back to source bases

- Reconcile new BAVD authorities to Wikidata and SNAC

Note: In the course of the BA/VD project, updates were only received from the University of Minnesota team, and so only a workflow for updating the KUOM collection exists.

KUOM (UMN) Metadata

To update program records when there is new/modified CPF data:

- Using OpenRefine, parse JSON file to table
- Wrap in quotes values containing commas
- Join multi-valued cells using ", "
- Export table as CSV
- CSV import new metadata ***other than CPFs*** into programs table, merging on program ID–updating the creators/contributors is more complicated, so do not import data in those fields in the first round of updates
- Temporarily create a copy of the Creators and Contributors text fields in the program tables, and name them "Creators update" and Contributors update"
- CSV import the updates to the Creators and Contributor fields into the copied "update" fields–this will allow us to isolate the program records with changes to creators and contributors
- Create two formula fields, one to compare old and new creators fields, and one to compare old and new contributor fields–use the formula {OLD VERSION OF FIELD}={NEW VERSION OF FIELD} (0 means the fields are different, 1 means they are identical)
- Filter the view in Airtable where either the compare creators field = 0 or the compare contributors fields = 0
- Copy the creators/contributors from the text update field into the creators/contributors LINK field
- Isolate and re-merge any duplicates that were introduced/reintroduced from the creation of new linked records–here is one way to do that:
    - In the KUOM CPF Authorities table, filter to look only at CPF records where number of creator and contributor links is 0–these are authorities whose names likely were fully corrected due to typos in the original data, therefore these out-of-date authorities no longer link to anything
    - Search for matches for each cell in this list against the list of newly created CPFs from the updated data import, and when found, mark them in a new field called "Update reconciliation"
    - Dedupe the combined list of new and old authorities that refer to the same CPF, merging into the OLD record, but keeping the NEW form of the name and other metadata (like linked programs)--this will maintain existing connections in the BAVD CPF Authorities base
    - Next, isolate any new records created by the updated data import with parentheses in the name, and preserve the role and original names in the CPF Role and Original CPF Name field

  - Then, create a formula field that cleans the parentheses role off the name, and copy the cleaned names over into the primary key (KUOM CPF) field
  - Dedupe the whole KUOM CPF Authorities table on exact matches only, merging into the OLD record (but keeping any and all new data)
- Once all new creators/contributors have been added to the linked field, copy the contents of the Creators/Contributors update fields to the originals, as that is now the "canonical" data that UMN stores
- Isolate any CPF authorities in the KUOM base that now have no links to any programs and delete them. You'll also need to delete them from the BAVD CPF Authorities table in the projector base
- In the BAVD Projector base, isolate the CPF authorities not yet linked to the BAVD CPF Authorities table, wrap them in quotes, then link them to the BAVD CPF Authorities table
- Isolate and merge any duplicates that exist in the BAVD CPF Authorities table as a result of adding new records to the table–here is one way to do that:
  - Filter out any records that are already tagged "Has distinct exact match" or "Has distinct similar match", then dedupe the whole BAVD CPF Authorities table on fuzzy matches, paying attention only to dedupe matches where one record was created recently–those would be the ones you need to potentially dedupe
  - Merge into the old record when deduping (to keep record history)
  - Isolate remaining records in BAVD CPF Authorities with a date of record creation matching when the updated CPF data was added and use ctrl+F to search for matches manually against the list of the rest of BAVD CPFs, then dedupe any matches [note: only recommended if remaining list of new CPFs is small, less than 200 or so records; otherwise, use OpenRefine cluster and edit]
- Mark any CPF records that now fall into Crit-Index and/or Crit-Rec-Enhance as such
- Reconcile any new Crit-Rec-Enhance records to Wikidata and SNAC

Projector base

The purpose of syncing several collection-based bases into one CPF-centered projector base is to minimize the amount of data that must be regularly managed and changed during the project. This multi-base system also allows us to refer back to how the data, especially names of authorities, originally appeared when we received it, making it easier to undo mistakes and sort out cases of distinct authorities with the same name. Additionally, since we receive data as program records but are not dealing directly with the program records in this project, storing those records in other bases cuts down on unnecessary metadata being presented to us (which will also hopefully improve the base's load time and speed).

Overall structure

Referring back to the source bases: Each table that is synced to a source base has an 'Open source record' button that allows for easier reference back to the source base when needed (e.g., to examine the linked programs, or track down a potential mistaken merge of authorities).

Creating the combined authorities list: The authorities from each of the four synced tables will be combined together after deduping and cleanup by linking to a single fifth table called 'BAVD CPF Authorities'. In this table, the overlapping authorities that appear in multiple collections will be identified and merged together.

Extending the combined authorities: Once a cleaned, deduped, combined list of authorities is available, this list will be reconciled against Wikidata and SNAC and extended with data from these sources (which will also highlight the authorities needing enhancement). The data pulled in from Wikidata and SNAC will be stored in their own tables in the base, linked to the BAVD CPF Authorities via their Wikidata IDs and SNAC IDs.

Phased deduping and merging

CPF authorities are being deduped in several phases, as a way of maintaining a trail of breadcrumbs should two authorities be marked as duplicates, only to later find out that they are distinct. Additionally, should any of the underlying collections data change in the future (e.g., we get an updated dataset from any of the holding institutions, this phased approach should make it easier to add any new authorities that appear or merge together authorities that are newly represented in multiple collections.

Initial cleanup/deduping: Initial data cleanup occurs in the four source bases, rather than in the combined projector base. This round of cleaning includes only basic things like standardizing name formats so that exact matches can be identified. For example, in the KUOM data, each authority has their role in that radio program specified after their name in parentheses. Initial cleaning removed these and separated them out to their own column so that any names that overlap across multiple collections will be automatically identified. The NFCB data similarly includes dates of existence after many authority names–these were stripped off so that the names alone remain. Then, the Airtable deduping app was used to merge the now-exactly matching names, along with any similar matches that were only varied by capitalization or punctuation. Any other similar but non-exact matches were left alone for now. The "original" name forms for each authority are preserved in a multiple select field in the source base. NOTE: For the Airwaves source base, the initial cleanup of stripping dates does not occur, because the Airwaves site uses names with the existence dates attached.

Second-phase deduping within collections: The second phase of deduping occurs within each source data table in the projector base, and looks for similar and fuzzy matches of authorities within each collection individually to create the most pared-down list possible of authorities before the four authorities lists are merged together. For this deduping, existence dates will also be stripped off the Airwaves CPFs, in preparation for merging all four lists together. This dedupe will occur on a copy of the key field of each source table in the projector base, which will ultimately become what links to the BAVD CPF Authorities table. This dedupe uses OpenRefine's cluster and edit feature.

General deduping of BAVD CPF Authorities: The final, and most involved deduping will occur after the four source data CPF authority lists have been cleaned, internally-deduped, and linked to a combined BAVD CPF Authorities table. Upon initial linking, any exact matches across the four collections should be automatically merged together. Before additional deduping occurs, these should be manually inspected to ensure that there is no question that the exact matches refer to the same entity (e.g., that there aren't two different "Smith, John"s that need to be disambiguated. Once the exact matches are confirmed, Airtable's deduping app and OpenRefine's cluster and edit can be used to find similar and fuzzy matches, deduping them once it is confirmed that they refer to the same entity. This will be one of the most time-consuming tasks, but is necessary to find the true overlaps across collections.

Data reconciliation and extension

Applying criteria for reconciliation and enhancement: Once a fully deduped and merged list of authorities exist, those which do not meet the criteria for reconciliation and/or enhancement should be either removed or marked insignificant. Any entity found across multiple collections is immediately significant, and meets the criteria for reconciliation (crit-rec). For entities only found in one collection, we will need to determine a standard by which we decide if the entity meets crit-rec, probably related to the number of radio programs attached to the entity. All the rest (likely a couple thousand) will be marked insignificant, so that attention is only paid to the subset of authorities for which there is overlap or are individually significant enough. Ideally, this will be only a couple hundred entities.

Reconciliation: The list entities meeting crit-rec will be reconciled against both SNAC and Wikidata, using information from the four collections to disambiguate and make correct matches against both sources. Any and all entities with matches in either SNAC or Wikidata will be extended by pulling data directly (Wikidata) or getting the SNAC staff to provide us the appropriate constellations.

Enhancement: From the entities that meet crit-rec, a smaller set will fall under the criteria for enhancement (crit-enhance). These entities will have SNAC and Wikidata records created if no match currently exists or otherwise have their Wikidata and SNAC records enhanced by the BAVD team. This enhancement work will first take place in Airtable, then be submitted to Wikidata and/or SNAC manually and with OpenRefine.