**Proposed Nonprofit Organizational Meta-Data Standards**

*Draft January 2023*

The standardization of the following attributes would enhance data value across organizations. Data standards could be adopted internally or public assets like crosswalks could be produced and shared to converge on some common definitions and language.

For example, tax year vs filing year are often used interchangeably and cause confusion. The meta-data standard would define what is meant by “year of data” and entities can disclose whether their dataset is organized in the standard format (e.g. tax year) or an alternative format (e.g. filing year). At the very least the terminology would be defined.

The return on investment for time invested in these resources would be making nonprofit data more:

1. Valuable (support more insights)
2. Accurate and complete
3. Easy to use

It is a worthwhile endeavor if we can find a way for members to take the lead on one or two standards, which means that collectively each invests resources in one new data asset and has access to the dozen or so created through the group. The ROI on the time invested in a single asset is amplified ten-fold through the collaboration.

**Definitions**

Meta-data standards would be defined in a shared manual or research guide that is published on GitHub or a similar platform that would help manage discussions, issues, and edits.

Some of the value would come through shared language like the example of the term “year” above, and some of the value would come by making feature engineering explicit. When we talk about “grant-making organizations”, for example, what is the threshold that we use to categorize a nonprofit as a grant-making entity? It can be an arbitrary cut-off (more than 30% of expenses are grants) – it doesn’t have to be perfect, it just has to be explicit so that people know how derived fields on meta-data tables are produced.

**Meta-Data Assets**

An asset might be a specific data table described below which includes nonprofit data, external augmentation data like harmonized census tables, or custom ontologies like the Guidestar Philanthropy Classification System (PCS). In the future assets might also be tools, such as an API that takes nonprofit mission statements and returns predicted NTEE codes or takes the raw title of an employee and returns a standardized job category.

For some assets we might want to create and maintain one canonical version. For example, a table of NTEE codes associated with nonprofits. They are assigned by the IRS and there is no official mechanism to update codes. Several organizations have developed their own systems for updating codes, e.g.

<https://x4i.org/nonprofit-ntee-code-finder>

But that will result in orgs having different NTEE codes across platforms. A centralized process means all orgs are not spending time on the problem and all data is more consistent and fresh.

In other cases organizations might invest significant resources in gathering, curating, cleaning and augmenting data to the extent it becomes a marketable resource so it doesn’t make sense to share the data itself. In these cases would it be useful to have standards that simply describe the data structure – the tables, fields, and keys needed to support the standard? That way members can keep their actual data private but it would make it possible to define and describe the meta-data and easier for partners to share data and collaborate.

**Creation and Maintenance of Assets**

Given the collective action challenge of creating public goods two models make sense:

1. One central entity like a large company or a government agency creates all of the assets and shares them (google has done this with many open-source projects – share their own internal standards or libraries for others to use).
2. A set of organizations collaborates on the design of standards then agrees to each take the lead on one or two assets that are shared with the group, coordinating to avoid duplicative efforts.

Strategy 2 likely makes sense for this group.

**Organizational Meta-Data Tables**

**Organization Name**

We often need to identify an organization by its name, but organizations often have either (1) many names because of DBA status or use of acronyms, or (2) names are not unique because of federated structures. The names database would facilitate disambiguation purposes.

* Table of Nonprofit Names (one-to-many)
  + EIN
  + Year
  + Source of data
  + Organization name
  + Organization name type: official name or alternative doing-business-as (DBA) names or abbreviations
* Table of Entity Status (one-to-one?)
  + EIN
  + Entity status: a categorical variable indicating whether they are
    - (1) an independent entity,
    - (2) part of a federated structure (one central organization / EIN and group returns), or
    - (3) part of an organizational network (each chapter has a unique EIN but identical names like Ducks Unlimited, Habitat for Humanity, etc.)
  + Group exemption number / affiliation network name (if it exists)
* Types of standards
  + Shared disambiguation tables described above
  + Shared publicly? YES

**Organization Type**

A table that includes all attributes that describe an organization’s corporate form (corporation, association, trust, etc.), purpose, tax-exempt status, mission codes, etc.

It should be a table of more-or-less immutable organizational traits that can be easily added to any dataset with the EIN.

A lot of these are derived from existing data like 501c status (03, 04, etc), but they could be renamed or reformatted to make them easier to apply. Others would be derived attributes from defined standards. For example, how is a grant-making organization operationalized? Private foundations are easy, but there are also community foundations and other grant-making nonprofits. There is a code to flag community foundations but not necessarily grant-making nonprofits, which would be flagged by a high proportion of their resources/activities spent allocating grants.

* One master one-to-one table with current traits for each org that can be used to join these mostly immutable org attributes to any dataset
* Table fields (not exhaustive)
  + **EIN**
  + **DATE**
  + **CURRENT (Y/N)**
  + Corporate form
  + Year of formation
  + 501c type
  + IRS ruledate
  + State of legal domicile
  + Common purpose code: service provider, advocacy, technical assistance, professional association, fundraising (monetary or non-monetary support), etc
  + Grant-making status: private foundation, public foundation, public charity that makes grants, etc
  + Affiliation status (e.g. single organization monetary support, foundation serving another org, etc.)
  + ~~NTEE codes~~
  + ~~PCS codes~~
  + ~~Tax exempt purpose codes~~
* Historic changes table – a more compact table that records changes since they would not be frequent and the use case for old values is less clear
  + EIN
  + Date
  + Variable name
  + Old value
  + New value
  + Source of change
* Type of standard
  + One current and cumulative table of attributes organized by EIN with only the current org traits
  + Cumulative meaning it contains EINs of closed entities that would no longer be in the BMF or Pub 78 but would appear in many longitudinal datasets
  + Historic changes to traits could be stored in a separate table
  + Shared public asset: YES

**Organization Location**

A set of location assets that recognize the multi-dimensionality of location and also provide crosswalks of common geographic units to make translation easier.

* Table type: one org to many locations
* Table fields
  + EIN
  + Date
  + Source?
  + Current?
  + Org address raw
  + Org address clean (building number, street, city, state, zip)
  + *Type of address? (HQ, Field, Accountant?)*
  + Lat-Lon geocode
  + Census tract FIPS
  + ZIP code
  + County FIPS
  + Metro ID
  + State FIPS
  + Address type: HQ, field office, service location
  + Location type: urban, suburban, rural
  + Source of the data and date (addresses change)
* Type of standards
  + Asset 1: standardized crosswalk of location definitions
    - Tracts, zip code, counties, metros and location type need to be harmonized to a specific year (the most recent files available) since geographic units change frequently
    - This asset should include shapefiles corresponding to the crosswalk year (typically harmonized to the most recent decennial census files)
    - Shared public asset? YES
  + Asset 2: census tables harmonized to the year of Asset 1
    - Harmonization of census data requires apportionment to reconcile changes in geographic unit definitions over time – for example, population count or unemployment measures at the tract level with data from 1990, 2000, 2010, and 2020 all harmonized to 2020 tract definitions
    - Harmonized files are necessary for longitudinal analysis but hard to produce and maintain – different files are needed for each geographic unit
    - See: <https://www.brown.edu/academics/spatial-structures-in-social-sciences/ltdb-following-neighborhoods-over-time>
    - Shared public asset? YES
  + Asset 3: disambiguation database of nonprofit addresses described above
    - Would require aggregating records from the same form across time, or from multiple forms and databases
    - Would be the basis for better matching processes that typically require organization names and addresses
    - Shared public asset: NO because of potential abuse by marketing agencies

**Organization Mission & Industry Taxonomies**

Tables that contain nonprofit missions, activity descriptions, and activity taxonomies

* Table 01: Nonprofit Mission and Activity Descriptions (one-to-may)
  + EIN
  + Year
  + Mission statement
  + Source of info
  + *Current/best?*
* ~~Table 02: Program Service Accomplishments (one-to-many)~~
  + ~~EIN~~
  + ~~Year~~
  + ~~Program service accomplishments~~
  + ~~Source of information~~
* Table 02: Mission Taxonomies (one-to-many)
  + EIN
  + Year updated
  + Taxonomy type: NTEE, PCS, tax exempt purpose, etc
  + Taxonomy value/code
  + Primary category (Y/N) - if an org has multiple codes within a single taxonomy
  + Source (IRS, update from org, re-code, predicted value)
  + Confidence?
* Taxonomy Crosswalk
  + Crosswalk of NTEE, PCS, NAICS, and any other taxonomies that have a one-to-one mapping and can be easily merged to existing data when one code is present
* Type of Standards:
  + Asset 1: one table of current NTEE/PCS codes for all nonprofits (active and closed)
  + Asset 2: crosswalk of mission codes

**Additional Meta-Data Standards**

The list above is by no means exhaustive but can be treated as a jumping-off point.