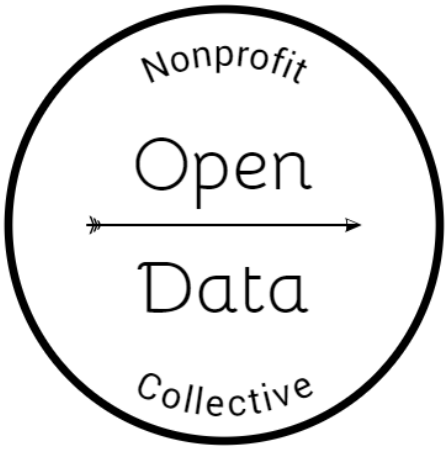
Instructions for Creating Concordance Files



Summer 2019

Instructions

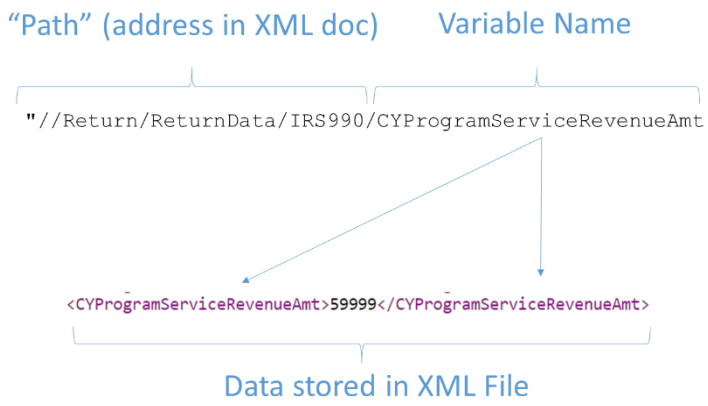
The IRS has released all 990 E-Filer return data as a collection of XML documents stored on an Amazon Web Server (AWS). The files currently contain no documentation and in a format that makes the data difficult to use. In order to make the data usable it needs to (1) be extracted from the XML document, and (2) documented.

To accomplish both of these objectives the Nonprofit Open Data Collective is creating a Master Concordance file that serves as a bridge between the current and the desired formats. Think about this file as the Rosetta Stone, as it allows us to translate data from complex XML files to flat spreadsheet tables (a relational database), and provides the documentation necessary to use the data for analysis.

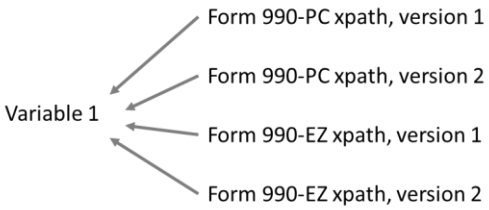
The Master Concordance file contains the following information:

- Variable name
- Variable definition
- The location of the field on the 990 Form or Schedule (for example, Part II, Line 3)
- The xpath

The "xpath" is just a fancy way to look data up on an XML form. It functions just like a directory path used to locate a file on your computer (for example, C:\Users\Documents\Recipes).



If it were as simple as matching a single variable to a single xpath and providing a definition this process would be fairly simple. There are two things that make it challenging: (1) the same variable has different xpaths on the 990-PC form and the 990-EZ form, and (2) the IRS has changed the e-filer forms multiple times resulting in many versions of a single xpath. The main purpose of the Master Concordance file is to identify all of the different versions of xpaths that correspond to a single variable.



As an example, on the Form 990-PC, Part IX, Line 4, Column A reports benefits paid to members. Similarly, Form 990-EZ, Part I, Line 11 reports the same field. The form was changed in 2013, so there are two versions of each xpath resulting in the following four possible locations of the same data:

(1) /Return/ReturnData/IRS990/BenefitsToMembers/Total
(2) /Return/ReturnData/IRS990/BenefitsPaidToOrForMembers/Total
(3) /Return/ReturnData/IRS990EZ/BenefitsPaidToOrForMembers
(4) /Return/ReturnData/IRS990EZ/BenefitsPaidToOrForMembersAmt

Your job will be to select a section of the 990 Form, then one at a time select each field (variable) in the section and identify all of the relevant xpaths.

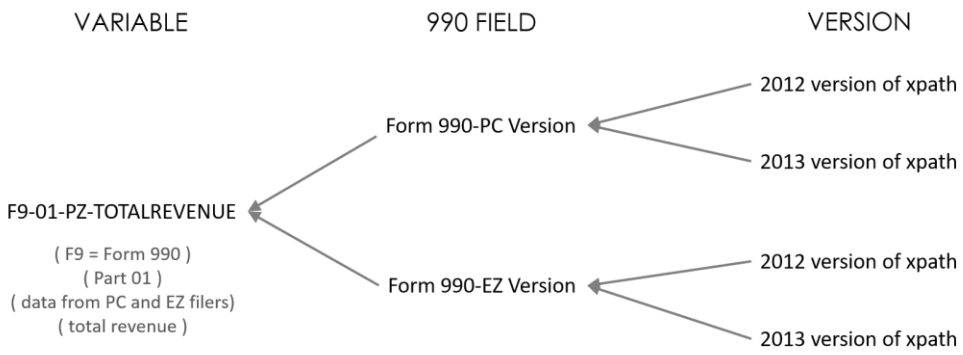
General workflow to enter 990 fields into the Master Concordance spreadsheet:

After your team has selection a section of the 990 Form or Schedule to work with,

(1) Select a specific field from the 990 form
(2) Identify the twin field on the 990-EZ form, if it exists (ignore this step for Schedules)
(3) Create a descriptive variable name (for example, TOTREV for total revenue)
(4) Append the appropriate prefix to the name (see below – e.g. F9-05)
(5) Create a clear definition of the variable using info from the 990 form and your knowledge of nonprofits
(6) Identify all of the relevant xpaths associated with the field
(7) Document the location(s) of the variable on the appropriate forms or schedules using a location code (see below)
(8) Record any relevant notes and questions about the process, and if you have questions flag the variable for further review by adding an “X” to the column labeled “flag for review”

In short:

1) Assign variable names to proper groups of xpaths.



2) Add meta-data (location codes, part, line number, scope).

3) Add database info: data type (simple), rdb_relationship, and rdb_table info.

4) Inspect data extracted using the group of xpaths associated with a variable and mark the variable as validated.

Fields:

FIELD	ADD?	DEFINITION
xpath		xpath
variable_name_old		old version (for reference only)
variable_name	X	variable name
variable_scope	X	Header or Signature (HD or SG), PC (only), EZ (only), PZ (PC+EZ filers), or PF. See scope rules on instructions
description		field description derived from the xsd schema files (update if it is not provided)
location_code_xsd		location codes derived from the xsd schema files
location_code_family		one location code for both 990-PC and 990-EZ versions to facilitate variable sorting (sorting by location_code_family should result in a variable order that matches the order on the 990-PC paper version)
location_code	X	human-created location codes
form	X	The data comes from which form or schedule: F990, SCHED-A, etc.
form_type	X	PC, EZ, or PF for F990 version. Use “SZ” for all schedules.
form_part	X	Field location on the 990 form or schedule by “part”.
form_line_number	X	Field location on the 990 form or schedule by line number (includes columns, section, etc.)
form_version		Derived from the xsd schemas
form_version_latest		The latest year in the list of versions
form_version_current		Is this the most recent version of the xpath? Only relevant for xpaths that are repeated.
data_type_xsd		custom data types listed on the schema
data_type_simple	X	simple version used the generate a relational database: text, numeric, checkbox, date
rdb_relationship	X	Cardinality of the variable: ONE-TO-ONE or ONE-TO-MANY
rdb_table	X	Name of the table that the variable belongs to (see instructions)
required		Is the field required to submit the return? From the xsd file, we might drop this because it's not clear how much the rules are enforced.
duplicated		Is the xpath duplicated in the concordance file? This only happens when the xpath remains the same across versions, but meta-data like the location changes.
production_rule	X	These are rules that should be applied to the data after scraping it in order to standardize or interpret it (for example true occurs as T on one version of the form, 1 on another).
validated	X	Has the VARIABLE been validated by collecting and examining data to examine the integrity of the xpath mapping? Indicate with an X.

New Variable Names

Many of the current names were computer-generated, and as a result very difficult to interpret. We want to make the names as easy to work with as possible.

Each variable name consists of a prefix of form-part abbreviates as 5 characters (e.g. F9-03 for Form 990 Part III), and a variable name. Confirm the prefix matches the other meta-data, then update the descriptive component.

Variable Prefixes

Variables names appear like this: F9_01_REV_CONTR_GRANTS_PY

Each variable name will contain a four-character prefix that contains information about the form and part where it originates. They serve two purposes.

- Naturally group variables by topics (parts on the form).
- Assure that variable names are unique and unambiguous (if you have a variable called "name" or "address" but no context, it's not clear what they are).

In order to create a prefix, apply the following rules:

- Assign a two-character code for the FORM or SCHEDULE using "F9" for Form 990 or 990-EZ, and "SA", "SB", "SC" for Schedule A, Schedule B, Schedule C, etc.
- Reference the Part and Line of the 990 where the variable is located.
    - Use "00" (zero-zero) for header info.
    - Use "01" (zero-one) for Part I, "02" (zero-two) for Part II, etc.
- The prefix is the topic group for the variables. Map the variables with an EZ scope to topics on the 990-PC form.

For example:

> F9-03- *refers to a variable found on Form 990, and it is located on Part III of the form.*
>
> SD-06- *refers to a variable found on Scheduled D on Part VI of the form.*

The variable name can contain **no more than 32 characters** including the underscores. Statistics programs cannot use cases with longer names.

Variable Names

When naming variables try to lead with the noun and follow with the adjective or qualifier. This helps to naturally group variables. Prefixes are omitted in these examples for simplicity:

Bad:

- grants
- net_fundraisers
- rent

Good:

- rev_grants
- rev_fundraisers_net
- rev_rent versus exp_rent

OK:

- street
- city
- zip

Better:

- addr_street
- addr_city
- addr_zip

Abbreviations:

Keep track of abbreviations so they can be applied consistently by multiple coders.

Original	**Abbreviation**
Total	TOT
Address	ADDR
Foreign	FRGN
Country	CNTR
ZIP Code (US)	ZIP_US
Postal Code (FOREIGN)	ZIP_FRGN
United States	US
SIGNATURE	SIGNTR
PREPARER	PREP
CURRENT	CURR
CONTRIBUTIONS	CONTR
Beginning of year	BOY
End of year	EOY
accounting	ACC
financial statement	FINSTAT
controlled entity	CE
Line 1, Line 2, etc	L1, L2
Officer	OFF

Principal	PRIN
Business	BIZ
Member	MEMB
Description	DESC
Receipts	RCPT
Revenue	REV
Total	TOT

Updated abbreviations are located on THIS SPREADSHEET.

Location Codes:

The location code indicates where the fields are located on the Form 990 or Schedule. Use the following conventions for location codes – this will ensure that they are machine-readable and can be **easily sorted and searched**.

FORM 990:

F990-PC-PART-03
F990-PC-PART-11-LINE-04
F990-EZ-PART-03-SECTION-A-LINE-09A
F990-EZ-PART-11-SECTION-B-LINE-04

SCHEDULES:

SCHED-A-PART-03-LINE-12B
SCHED-B-PART-02-SECTION-B-LINE-04

Note: If the variable scope is PZ, then document both the form 990-PC location and the 990-EZ location for the code. These are used to generate the data dictionaries, and will guide the reader to the proper variable locations on the forms.

- Specify the relevant Form or Schedule using “F990” or “SCHED-X”.
    - If the scope is EZ use “F990-EZ-“, if the scope is PC or PZ use “F990-“.
- Specify the location on the form using PART-##. Use 01, 02, up to 09 for numbers less than 10.
    - Use PART-00 (zero-zero) for any variables that occur in headers or footers, not numbered sections.
- If the form part is split into sections, include as SECTION-X.
- If a line exists, record as LINE-01.
- If a column exists, append it to the line number: LINE-01a.
- If there is a column but no line number use COL-A, COL-B, etc. For example, we would not reference line numbers from the table located at Part-07-LINE-01a because these are repeated items in a one-to-many table (DTK name, DTK title, etc.). But variables are organized by columns. So their locations codes would be something like:

Name of employee:	F990-PC-PART-07-LINE-01A-COL-A
Hours worked per week:	F990-PC-PART-07-LINE-01A-COL-B

When in doubt, select location codes that sort the variables by the order they appear on the forms. Because forms are not consistent there is no perfect way to do this, but try to be consistent.

Part VII Compensation of Officers, Directors, Trustees, Key Employees, Highest Compensated Employees, and Independent Contractors

Check if Schedule O contains a response or note to any line in this Part VII . . . . . ☐

Section A. Officers, Directors, Trustees, Key Employees, and Highest Compensated Employees

1a Complete this table for all persons required to be listed. Report compensation for the calendar year ending with or within the organization's tax year.

- List all of the organization's **current** officers, directors, trustees (whether individuals or organizations), regardless of amount of compensation. Enter -0- in columns (D), (E), and (F) if no compensation was paid.
- List all of the organization's **current** key employees, if any. See instructions for definition of "key employee."
- List the organization's five **current** highest compensated employees (other than an officer, director, trustee, or key employee) who received reportable compensation (Box 5 of Form W-2 and/or Box 7 of Form 1099-MISC) of more than $100,000 from the organization and any related organizations.
- List all of the organization's **former** officers, key employees, and highest compensated employees who received more than $100,000 of reportable compensation from the organization and any related organizations.
- List all of the organization's **former directors or trustees** that received, in the capacity as a former director or trustee of the organization, more than $10,000 of reportable compensation from the organization and any related organizations.

List persons in the following order: individual trustees or directors; institutional trustees; officers; key employees; highest compensated employees; and former such persons.

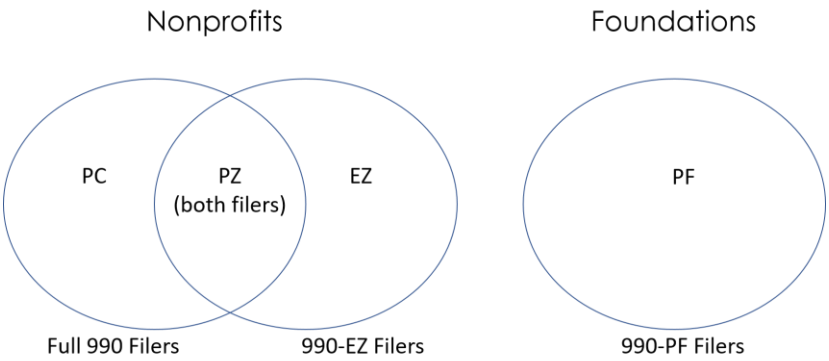
☐ Check this box if neither the organization nor any related organization compensated any current officer, director, or trustee.

(A) Name and Title	(B) Average hours per week (list any hours for related organizations below dotted line)	(C) Position (do not check more than one box, unless person is both an officer and a director/trustee)						(D) Reportable compensation from the organization (W-2/1099-MISC)	(E) Reportable compensation from related organizations (W-2/1099-MISC)	(F) Estimated amount of other compensation from the organization and related organizations
		Individual trustee or director	Institutional trustee	Officer	Key employee	Highest compensated employee	Former			
(1)		☐	☐	☐	☐	☐	☐			
(2)		☐	☐	☐	☐	☐	☐			

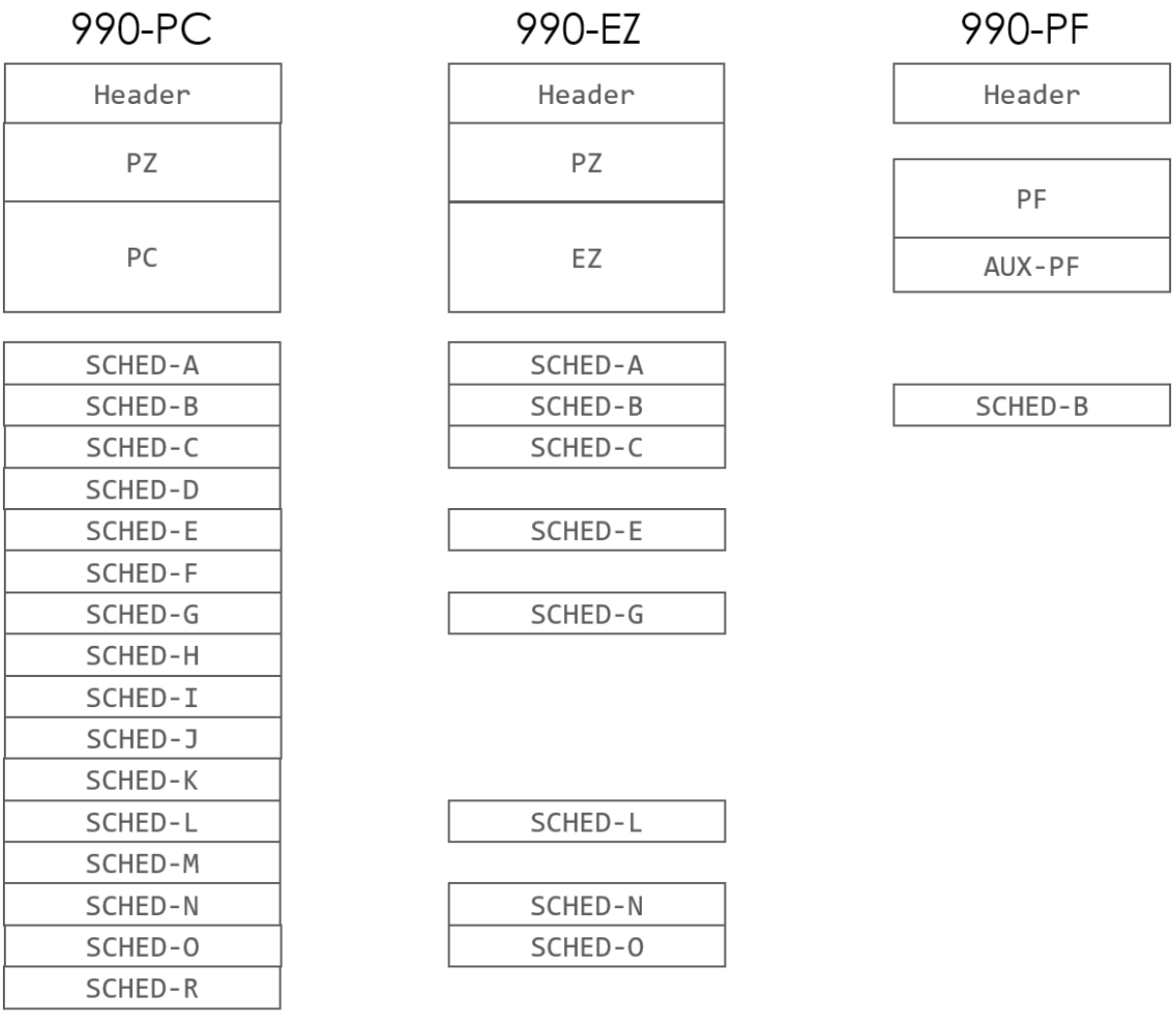
Variable Scope

Scope codes describe which forms and schedules contain which variables:

HD = Header and Signature (all filers)
PZ = 990 + 990-EZ Filers
PC = Only 990 Filers
EZ = Only 990-EZ Filers
PF = Only 990-PF Filers



VARIABLE SCOPE



Relational Database Schema Conventions for the Master Concordance File

Schema Fields on the Concordance File:

These three fields currently on the concordance file will be used to document the RDB structure:

data_type (xsd and simple)
rdb_relationship (one-to-one or one-to-many)
rdb_table

Conventions:

Data Types

The data_type_xsd and data_type_simple capture data types of the fields on the forms. The xsd schema version retains the meta-data from schemas and the simple version reduces types to:

- numeric
- text
- checkbox
- date

RDB Relationship:

Code each field as either “ONE” or “MANY” representing the relationship between a single tax return and possible values of the variable (one-to-one or one-to-many). For example, each filing contains one EIN. A single filing will include multiple board members.

RDB Table naming conventions:

RDB tables are organized according to forms, parts, and table as follows:

```
F9-P03-T00-DESCRIPTIVENAME  (Form 990, Part III)
SA-P07-T04-DESCRIPTIVENAME  (Schedule A, Part VII)
```

Form/Schedule– Name of Part – Name of Table – Table Num.

Where **TABLE-00 contain all one-to-one relationships** from the form or schedule, and TABLE-01 or higher contain one-to-many relationships. For example:

F9-P03-T00-MISSION
F9-P03-T01-PROGRAMS

Forms are split into parts, so there is a TB-00 for each part, and if relational tables exist they will start at TB-01, TB-02, etc. Some one-to-one fields might be repeated across tables.

Tables for the main 990 form:

FULL-NAME	DESCRIPTION	RDB	EZ-SECTION
F9-P00-T00-HEADER	Return Header	one	header
F9-P02-T00-SIGNATURE	Signature Block	one	sig block
F9-P01-T00-SUMMARY	Summary	one	part-01
F9-P03-T00-MISSION	Mission and Program Changes	one	part-03
F9-P03-T01-PROGRAMS	Program Service Accomplishments	many	part-03
F9-P04-T00-REQUIRED-SCHEDULES	Checklist of Required Schedules	one	part-05?
F9-P05-T00-OTHER-IRS-FILING	Statements Regarding Other IRS Filings and Tax Compliance	one	part-05?
F9-P06-T00-GOVERNANCE	Governance, Management, and Disclosure	one	part-05?
F9-P07-T00-DIR-TRUST-KEY	Officers, Directors, Trustees, Key Employees, Highest Compensated Employees, and Independent Contractors	one	part-06
F9-P07-T01-COMPENSATION	Section A: Compensation of Officers, Directors, Trustees, Key Employees, Highest Compensated Employees	many	part-04;06
F9-P07-T02-CONTRACTORS	Section B: Compensation of Independent Contractors	many	part-06
F9-P08-T00-REVENUE	Statement of Revenue	one	part-01
F9-P09-T00-EXPENSES	Statement of Functional Expenses	one	part-01
F9-P10-T00-BALANCE-SHEET	Balance Sheet	one	part-01
F9-P11-T00-ASSETS	Reconciliation of Net Assets	one	part-02
F9-P12-T00-FINANCIAL-REPORTING	Financial Statements and Reporting	one	header

At the end of each form and schedule there is always a supplemental information section. These will all be designated as **TABLE-99** since they all have a one-to-many structure and are always at the end of each form, and they will all have the same name SUPPLEMENTAL-INFO. So they differ only in the form/schedule prefix.

XX-XXX-T99-SUPPLEMENT-INFO (general)
SA-P06-T99-SUPPLEMENT-INFO (specific)

Primary Keys *(these are added at the programming stage)*

The Document Locator Number (DLN) serves as the official primary key for joining tables.

These variables will be included in all of the tables.

- **DLN**
- EIN
- org name
- filing year
- submission date

Note EIN-filing year will not be unique if a nonprofit appends a filing.

The Validation Process

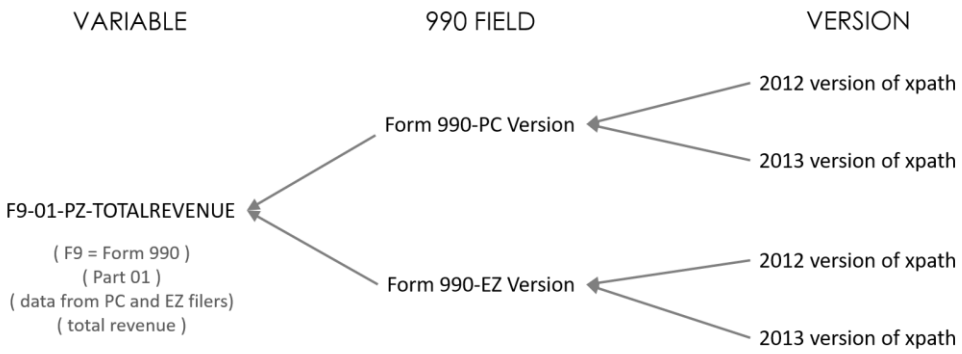
Once you have completed the validation and documentation of a variable, mark each case with an X in the validation column at the far right.

We are looking for three primary types of errors:

(1) An xpath is incorrectly associated with a variable. The particular xpath indexes data that is distinct from the rest of the data in the variable (for example, four xpaths that index total revenue and one that indexes total assets).

(2) All xpaths that index the same concept are split between two or more variables and need to be merged. For example, conceptually the Total Revenue Current Year on the form 990 and Total Revenue on the form 990-EZ are indexing the same data since the 990-EZ form does not split revenue into current and prior years, only current.

(3) Meta-data is incorrect. Update data types, and if Xpaths are available on date of event, update location codes as necessary.



OUTSIDE SCOPE OF VALIDATION:

We do not need to make sure the data in the XML documents is correct. We only need to make sure that we are correctly extracting data from the XML format.