990 XML: UNBREAKING TAX FILINGS,

SOMEWHAT, IN A BROKEN WORLD

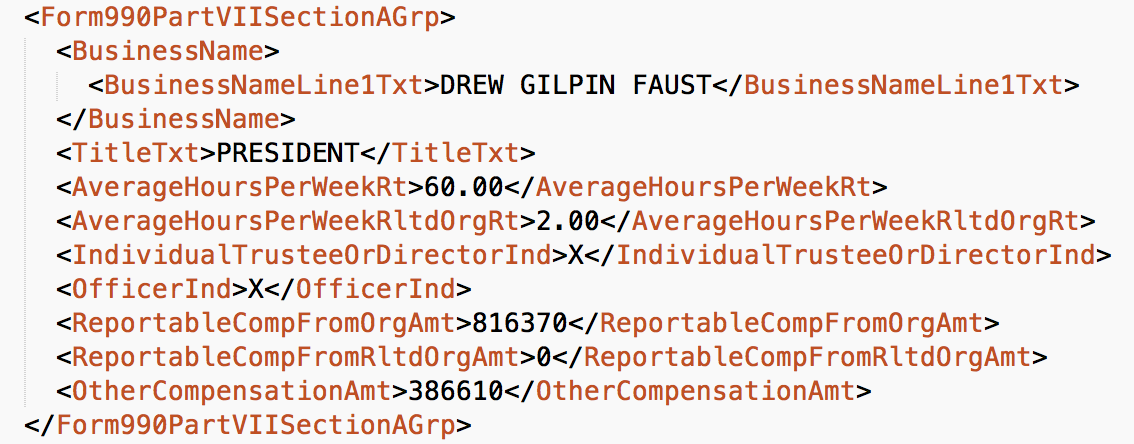


Image source: Trustees of Harvard University, 2016 filing, (via Sublime Text)

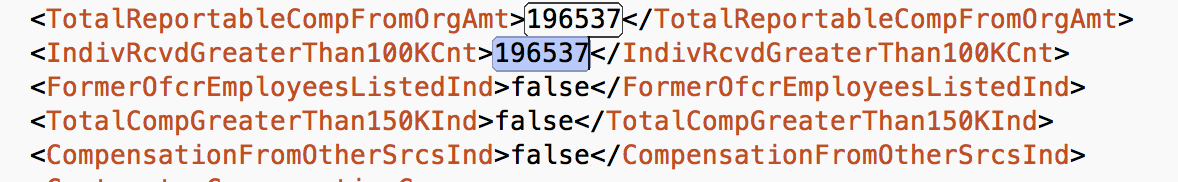
SCOPE OF VALIDATION: [ Source, Jesse Lecy, last call in September] The aim is to reproduce the variables, as they appear in the xml filings (regardless as to whether they are correct in the traditional sense, see above).

TYPES OF ERRORS: It is necessary to be exceptionally precise while we’re talking about errors. What types of errors do we need to be able to catch?

TYPES OF ERROR-CHECKING TESTS: If an xpath is “wrong” for a subset of filings, then there may be a statistical difference in the values of the “wrong” values to the “right ones” (assuming the “right” and “wrong” values come from columns that are statistically distinguishable. Extra credit: given any two randomly transposed variables, what is the chance they will be statistically distinguishable?). What algorithm would you run, and above what tolerance would you flag an “anomaly” for human review? Are there judgements that can be made without a human involved?

IRS PROCESSING ISSUES: Some schedule B numerical fields appears as “RESTRICTED”; SSN values detected in free text fields are replaced as “XXX-XX-XXXX”. The latter is a field-lengthening requirement--that is, the IRS will replace a nine-digit series of numbers with an eleven character “censored” version of it, breaking the length of free string fields. There are doubtless other issues that occur between filing submission and irs release (maybe Ben Getson knows about these?). Both of these issues break IRS type definitions, though aren’t that hard to hack around.

NUMERICAL ISSUES: The xml is validated against the xsd types (which include regexes at some point) but lots of data is just wrong. The “NATIONAL WATERWAYS CONFERENCE INC”, FY2014, below, seemingly reports they employ 196,537 at a salary of $100,000 a year or more (though that value looks like it was somehow copied from the line above?).



LOW TECH VALIDATION: [Source: Jacob]

Have folks look at the output of a standard format of the processed xml data versus the IRS’s pdfs and/or/ the source xml. Pick a variety of forms. Publish the “verified” data in a code-independent format, with the full xpath, so that anyone can use it to test against that filing.

Preference is to double enter forms to make sure they are right--this is super tedious.

This could be done separate from a “filing” to get all the variables, it’s just that there’s a time cost to looking up each filing for someone to compare to.

PHYSICAL ISSUES:

How many humans? Located where? Doing what? How many bathrooms? How much coffee?