OES and SOC – It's complicated

The occupational coding structure used by the BLS is the Standard Occupational Classification (SOC). While several programs rely on the coding structure, the biggest and most used is the Occupational Employment Survey (OES). This survey is the best source of wage data at the occupational level, providing hourly and annual wages and employment counts annually and at the national, state, and regional level. Despite their intention to use SOC codes, the limitations of survey methods do result in some discrepancies between official SOC structure and the codes and titles under which OES data is published. Understanding what types of discrepancies exist and why can help troubleshoot errors when the new data is released.

Methodological reasons for OES-SOC discrepancies

Problem		
Ability to distinguish between two similar occupations with the survey questions	**Description**	The survey instrument sent to employers is a standard form meant to accommodate all industries and occupations. Titles used by employers, though, are not necessarily the same as SOC codes or any kind of industry standard. As a result, there are some types of occupations where you simply cannot identify the specific SOC code with any confidence and the OES program instead assigns them to a combination of codes (an occupation roll-up).
	Consequences	There are two kinds of roll-ups. There are the ones inherent to the SOC structure, like a 2-digit SOC, the built-in grouping of like occupations in the same left-most digits. There are also some custom OES roll-ups, designed to combine a subset of the detailed occupations within a standard roll-up. In the case of the first type, the results are that 1) you have to be prepared to use roll-up occupations so that you don't lose the most detailed level of information and 2) comparisons to other data sets can become complicated. In the case of the second type, the issues are the same but there are complications with documentation and labelling – OES-specific codes are often presented as SOC codes and then descriptive information can be missing and the relationships with other sources may not be defined.
R&D vs. non-R&D (or industry-specific occupations)	**Description**	OES publishes data under some unique codes that distinguish between R&D and non-R&D. The occupations are the same, but in certain industries they are substantially different both in assigned tasks and compensation. Breaking them out improves the data quality, but requires more occupational detail than SOC provides.
	Consequences	These codes are not SOC codes and are not documented as such. Descriptive information may be missing or confusing, and if they're being related to other data sources there can be confusion or problems with joins.

SOC structure revisions	**Description**	Every 10 years or so the SOC structure is reevaluated and restructured. This is necessary because new occupations come into existence and others change or disappear. Rather than making those revisions on a rolling basis, there are distinct versions that can be related back to one another through time. Unfortunately the OES program combines data from three years to produce the estimate of a single year. This is important for confidence in the estimates, but it complicates the change between SOC structures.
	Consequences	Because of the 3-year panel, during a switch between SOC structures OES uses non-standard codes to combine different occupations. These vary from year to year and are specific to the changes being made during that transition.

Implementation of SOC 2018 in OES

The approach to the change to SOC 2018 has been a gradual, multi-year process. In the first year (data year 2018, published in 2019), SOC 2010 was still the primary coding structure but 6 new hybrid codes were rolled out. In the second year (to be published in 2020), codes have largely been converted to 2018 but some of the old hybrid codes were preserved and new ones were created.

OES and the WID

In the WID many disparate data sources and programs are brought together under similar labeling structures. Those structures are often built off of and used to power applications that may relate data from different sources. As a result, there have to be common lookup tables and crosswalks to standardize the relationships between those data sources and preserve referential integrity. As a result, the fact that OES is SOC-based with some tweaks creates some problems, even more than if it were an entirely different coding structure.

Referential Integrity

Referential integrity refers to the structural checks built into the database to ensure that only good, meaningful information is added. The most common one that people encounter is a Key – Primary Keys apply to the same table to ensure that duplicate or competing data is not loaded (this is why the area and time period are always part of the key in a WID table – you are forced to replace old data so that edits and revisions can't get mixed up with the correct value). Foreign Keys point to other tables – they're meant to ensure you're not loading data assigned to an industry that doesn't exist or in an invalid area. That means that the table they're pointing to (the lookup table) has to be correct and complete. All the data tables that use occupation data point to the same occupation code lookup table, so if they're all using SOC 2018 the codes listed in the table as codetype 19 (SOC 2018) have to include every valid code that might be used in the data table. So although there are OES-specific codes that are not officially SOC, they have to be in that OCCCODES table as codetype 19, too, or it's not possible to load the complete data.

Sometimes people try to solve this problem by creating a different occupational type code for the non SOC-standard codes in OES. Usually, that makes a mess. The outputs we get from the BLS and LEWIS to load all have the same codetype for a given publication year, so you actually would have to identify the

ones that are different and change them. Plus, the person doing the data load would have to have a level of OES program expertise that very few people have. As a result, doing it that way would be error prone, unsustainably complicated, and confusing. Instead, we mark them in the SOCCODE table with an oesflag. OCCCODES is for referential integrity, while SOCCODES contains information about the occupations that is useful for analysis.

Titles

Another way that the OES codes may differ from standard SOC codes is in their titles. In some cases, there are either SOC or OES changes to the titles only. These aren't usually huge changes (SOC doesn't replace a code with a completely different occupation), but when states want to present data under the accurate title and it's different between OES and other programs changing the title to the OES version in the OCCCODES table may break the displays for other data sources. This is another reason for the SOCCODES table – it gives a more options for how to title occupations.

Although a different title may not substantively affect the meaning of the data presented, setting up appropriate business rules for how to deal with the discrepancies and identify problems can actually be more difficult. You won't have a foreign key violation with the wrong title – you'll get a phone call from someone looking at an application and have to work backward through every transformation. Defining the meaning and use for the various code title fields and sticking with it from year to year – codetitle in OCCCODES and soctitle and soctitlel in SOCCODE – can save a lot of trouble.

Crosswalks

Often states want to relate data between programs. That's one of the major advantages of the WID – once the work to cleanup and load all the different program data into a structure that shares common definitions has happened, it's much easier to pull out many different data elements based on a common characteristic. When comparing through time, though, sometimes you need a structure in place for making that comparison. So, to compare SOC 2010 wages to SOC 2018 wages it would be necessary to know if a code has been changed or split or merged. There's room in the WID structure for those crosswalks, but because they're not structurally needed they're often not loaded, updated, or maintained until someone has a use for them.

Relating data between programs may also require a crosswalk or some kind of business logic. Even if most SOC codes are common between them (Projections and OES, for example) some are different but still possible to relate. So applications either need to be able to make imprecise comparisons through a crosswalk and use adequate descriptive content to explain the mismatch to users or they need to be able to deal with the potential for unmatched (null) data. Many times this need is very application-specific and might have to be created or updated manually.

Recommended Procedures

There are many different uses for the WID and depending on the needs of applications or analysts, different problems can rise to the forefront. While it's not possible to anticipate and solve all those problems, there are some broad approaches that may prevent frustration.

2018 OES Data

In 2019 OES released the 2018 vintage data. This is the version that was largely SOC 2010 based, but included additional hybrid codes and specialty roll-ups. States loaded this data mostly back in May,

ideally with codetype 14 and they should have been able to preserve referential integrity by adding a few occupations to the OCCCODES table. Available for download in the WID format for various tables here: http://data.widcenter.org/download/OES%202017%20changes/

2019 OES Data

The 2019 OES data will be released around May of 2020. In broad strokes, though, there are two approaches to preparing the WID for this new data load.

SOC is king approach:	**OES is best approach:**
Overview	
Build your lookup table content by combining SOC 2018 values, 2019 hybrid codes, and OES special aggregations from two separate data files.	Load a complete lookup table generated from the OES publication occupations.
Best for…	
This makes it easiest to preserve official SOC codes and titles and would be ideal for states that are using several SOC based data sources because it loads a complete SOC structure as well as the OES additions. This would also be the easier approach if a state has already loaded SOC 2018 data – because the codes are added in separate files you can avoid duplication.	This is easiest for states that use OES data but no other occupation data and that don't need to relate different data sources.
Considerations	
The SOC 2018 content linked to below includes all levels of detail, not just the detailed occupations that OES uses for data collection. Standard roll-ups don't have to be added separately, but it does mean that there may be codes in the OCCCODES table for which there are no data records, depending on suppressions and what aggregations are published. Depending on application needs, it may also be necessary or desirable to load the OES titles for codes and update a title field with the official OES code so it can be presented consistently with the BLS publication.	Because in some cases OES publishes to a SOC aggregation (5-digits, instead of the full 6), you will lose some detailed industries and some mid-level aggregations (3- and 4-digit roll-ups) doing it this way, in addition to using titles from the OES program instead of standard SOC. This file is generated from the national publication. While in general the national estimates have less suppression than local areas and this list is likely to work for preserving referential integrity, some states publish specific detail or different roll-ups. Usually, that's output from LEWIS. If your state has a special use case for OES there may need to be additional values added to the SOCCODE and OCCCODES tables. Another hazard is that needs change and if you or a future analyst goes to the SOCCODE or OCCCODES table expecting to get a complete and accurate SOC 2018 code structure they'll run into problems. Documenting the original source and intention of the contents of those tables is recommended.
Note – in the WID 2.8 release, all title fields were lengthened. That's the kind of change that can be missed during the upgrade process, but if there are problems getting the titles into the local version of the SOCCODE or OCCCODES table, check to make sure the field length is correct.	

Links to resources:

SOC 2018	OCCCODES http://data.widcenter.org/download/soc2018/occcodes.csv
	SOCCODE http://data.widcenter.org/download/soc2018/soccode.csv
OES 2018 hybrid codes	OCCCODES http://data.widcenter.org/download/OES%202017%20changes/occcodes2017add.csv
	SOCCODE http://data.widcenter.org/download/OES%202017%20changes/soccode2017add.csv
OES 2019 hybrid codes	OCCCODES http://data.widcenter.org/download/OES2019/occcodes2019add.csv
	SOCCODE http://data.widcenter.org/download/OES2019/soccode2019add.csv
All OES 2019 codes	OCCCODES http://data.widcenter.org/download/OES2019/OESocccodes2019.csv
	SOCCODE http://data.widcenter.org/download/OES2019/OESocccodes2019.csv