Tool L: Analyzing Data From Multiple Sources

When data analysts obtain data from multiple sources, they may have work to do to link different data files so that you can analyze them together. For example, to calculate student enrollments in arts courses, they may need two data files:

- A file from the Student Information System containing unique student IDs, the ID of the school each student attends, each student's grade level and demographic data about each student. In this case, the field containing student IDs would be the key field. Each student ID would appear only once, but the same school IDs would frequently reappear.
- A file from the Course Information System containing unique course IDs and IDs of students who take those courses. In this case, the field containing course IDs would be the key field. Each course ID would appear only once, but the same student ID will appear several times throughout the data file if that student has taken several different courses.

In the best-case scenario, both files use the same system of unique student IDs, which makes it possible to link them for analysis. By matching student records to one another, data analysts can analyze student enrollments by school and demographic group, for example.

What do data analysts do when they receive multiple data files on the same students or entities that don't share fields such as unique student IDs? If they receive information from two different state agencies, for example, they might find that one dataset uses internal ID numbers for students while the other uses students' Social Security numbers.

In cases like these, analysts should:

- Determine if one agency can create a new file using the same identification key employed by the other agency — by substituting Social Security numbers for internal student ID numbers, for example.
- If neither organization can make that accommodation, analysts may have to match the data sets with one another by finding other ways of linking them. For example, they could use student names and birthdates to link students in one file to students in the other. While some students may share the exact same names and birthdates, that problem would not be common.