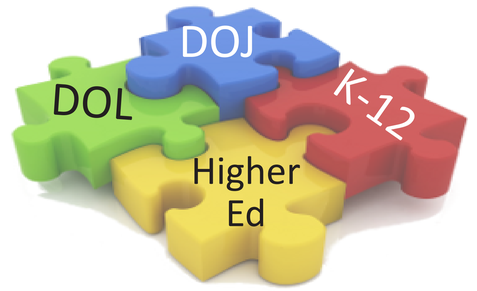
Creating Cross - Agency Longitudinal Datasets for Education Research



Securing

SECURITY HELPS:

Protect families from embarrassing disclosures, discrimination, differential treatment, and potential threats to family & job security

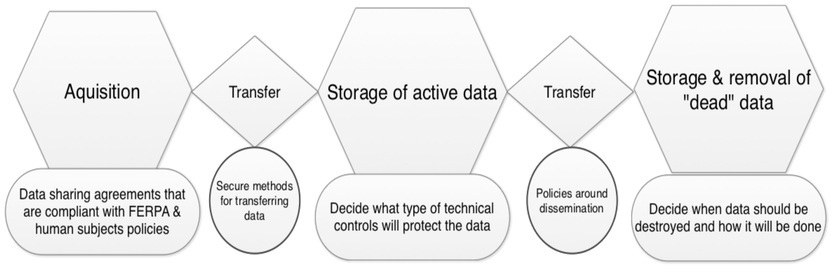
PERSONALLY IDENTIFIABLE INFORMATION (PII) is:

Names, ID numbers, and information that can be used to trace an individual's identity directly or indirectly

TWO SIDES OF SECURING PII DATA INCLUDE:

- Policies & Procedures (data sharing agreements)
- Technical controls (encryption, firewalls, etc.)

THE DATA SHARING PROCESS



Establish data flow policies and procedures. Your technical controls, like encryption and firewalls, are only as strong as how well everyone uses them. Make sure to plan for security around every step of the data sharing process!

BEST PRACTICES IN DATA SECURITY:

- Designate a team to coordinate policies and IT infrastructure
- Make sure there is a data sharing agreement in place
- Only share student level data through a password protected and encrypted sites
- Never email files containing student level information
- Keep printed reports stored in a secure locked location
- Always double check your policies with your legal counsel, laws vary across states and countries

DATA SHARING AGREEMENTS SHOULD INCLUDE:

- The organization or individuals who will have access to the data
- Limitations on use of the PII, including restrictions such as other datasets that will be linked to
- Penalties for inappropriate disclosure
- Terms for data destruction
- All applicable legal requirements to comply with
- A plan in accordance with State and Federal laws for responding to a data breach

Linking

IDENTITY RESOLUTION

1. Define Exact and Fuzzy matching criteria
2. Weight the criteria
3. Choose association threshhold

TRY EXACT MATCHING ON DATA ELEMENTS

- First name
- Last name
- Birth date
- Mailing address
- Gender
- Race/Ethnicity
- SSN

TRY FUZZY MATCHING ON DATA ELEMENTS

- Phonetic first name
- Phonetic last name
- Birth day/Year
- Birth month/Year
- Birth month/Day
- City and Postal code
- "1-digit-off" SSN

"FUZZY" NAME MATCHING

Phonetic Recoding (Steven and Stephen both "STFN")
Distance Algorithms (Dickson and Dixon = .83)
Combine Phonetic Recoding with Distance Algorithms for powerful matching

Coding Accuracy Support System (CASS)

Helps standardize mailing addresses

SOFTWARE:

LinkPlus, Google Refine, Informatice Identity Resolution, IBM Identity Insight, Infoglide Identity Resolution Engine

Cleaning

K-12

Common Problem	Suggested Workaround
Time-invariant variables not consistent across time	Take the modal value; then take the most recent value
Test score ceiling or floor effects	Use scaled scores instead of raw scores
Students in sample take different tests	Standardize within grade and school year
Students have multiple scores for the same test in the same year	Take the score from the first administration; if two scores on the same day, take the highest or average them
Missing test scores	Impute a prior year's test score or use multiple imputation
Core courses are not identified	Use course subjects and course names; reference codebooks if available
Students from different districts and/or states have different grading periods and/or grading scales	Use formula to standardize grading scales. Numeric grading scales are preferred for operations. Determine a "final grade" for each student based on grades for each term, weighted according to the number of terms

HIGHER EDUCATION

Common Problem	Suggested Workaround
Multiple IDs for the same student	Use data linking strategies
Identifying first year of enrollment after high school graduation	When possible, combine data from high school transcripts, K-12 data, and/or the National Student Clearinghouse. If not possible, take the first time the student attempted credits
Identifying developmental and dual enrollment courses	Be careful when interpreting variable names as "dual enrollment" could indicate courses delivered at a high school campus, but not courses taken by high school students at college campuses. Check course names against course numbers and subject codes to verify developmental and core courses

HUMAN RESOURCES

Common Problem	Suggested Workaround
Inconsistencies with years of experience	Make sure that teachers have only one novice year. Assume one year of experience is awarded for each year in which an individual has a job code of teacher and is assigned to students. Sometimes teachers have fewer years of experience than they did in prior years. Replace experience values with equal or higher values (depending on job code status) than the previous year. Sometimes teachers gain more years of experience than is possible during a span of time. Award the maximum possible, but no more. Fix later observations accordingly. Do not award or subtract experience for gaps in the data. The individual could have taught somewhere else or could have taken a break from teaching. Since it is impossible to tell, trust values that have the same value as the last year before the gap or the same value as the last year before the gap plus the number of years in the gap.
Inconsistencies with years of experience	It is important to know whether an individual is teaching in a given school year to troubleshoot inconsistencies in years of experience and to select samples for analyses. Code individuals as teachers if one of their job codes is teacher. When possible, assign job codes that indicate permanent roles over temporary ones and give preference to principals, academic positions, specialists, and counselors over job roles like educational assistant, coach, etc.
Multiple job codes and school codes in the same year	Standardize within grade and school year