Creating Cross-Agency Longitudinal Datasets for Education Research

|  |  |  |
| --- | --- | --- |
| Authors | | |
| Security | Linking | Cleaning |
| **Sarah Frazelle**  Senior Advisor for Research & Evaluation;  Education Northwest  sarah.frazelle@educationnorthwest.org | **Jared Silver**  Data Architect and Manager;  Center for Education Policy Research at Harvard University  jared\_silver@gse.harvard.edu | **Havala Hanson**  Senior Advisor for Research & Evaluation;  Education Northwest  havala.hanson@educationnorthwest.org |

**Information & Resources for DATA SECURITY**

Although laws vary across states and countries, some consistent best practices in data security include:

* Designate a team to coordinate policies and IT infrastructure
* Make sure there is a data sharing agreement in place for all student level data that is shared by an education agency.
* Only share student level data through a password protected and encrypted sites endorsed by your IT group and approved by legal counsel
* 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

Web Resources

Family Educational Rights and Privacy Act Regulations (FERPA) document *http://www2.ed.gov/policy/gen/reg/ferpa/index.html*

The U.S. Department of Education’s Privacy Technical Assistance Center (PTAC) offers up to date guidance and free training modules

*http://ptac.ed.gov/*

Guidance from the Data Quality Campaign on FERPA compliance from a policymaker’s view

*http://www.dataqualitycampaign.org/find-resources/complying-with-ferpa-and-other-federal-privacy-and-security-laws-and-maximizing-appropriate-data-use/*

Specific guidance for disclosure avoidance

*http://ptac.ed.gov/ptac-new-guidance-disclosure-avoidance-limiting-access-pii*

**Information & Resources for DATA LINKING**

Steps for designing your own identity resolution software

* Identify what data elements you have to match (SSN, name, address, etc)
* Define the Exact and “Fuzzy” matching criteria. Example fuzzy matching criteria:
  + 1-digit-off SSN: 123-45-6789 and 1**32**-45-6789 match
  + Birth Month/Day: 1/2/**91** and 1/2/**81** match
  + Phonetic First/Last Name: **Ch**eryl and **Sh**eryl match (and Sherryl and Sheryll…)
* “Weight” the criteria
  + “Best” matches get higher weight. I.e. Exact SSN match (100 pts) is better than a 1-Digit-Off match (50 pts) which is better than a Phonetic First Name match (5 points)
* Choose your “association threshold”
  + Identify the combinations of criteria you are comfortable with to determine two records represent the same person
  + I.e. Exact SSN (100 pts) and Phonetic Last Name (5 pts) combined can be considered the same person. Therefore, associate records when score >= 105
* Trial and error – adjust weights, adjust criteria, etc
* Always build in a mechanism to “de-associate” records that are *not* for the same person but *matched* (i.e. exceeded the threshold) and “manually associate” records that *are* for the same person but *didn’t* match

End-to-end identity resolution software solutions

* Google Refine
* Informatica Identity Resolution
* IBM Identity Insight
* Infoglide Identity Resolution Engine

**Resources for DATA CLEANING**

Information about common data cleaning problems found in education data and strategies to resolve them

*http://www.gse.harvard.edu/sdp/resources/toolkit.php*