IDENTITY MATCHING AT ERDC

John Sabel, ERDC
Research Coordination Committee
September 6, 2013
Identity Matching at ERDC

• Identities come in from more than two sources
• Matching builds across time, not a one-time process
• Each source of data has their own set of matching rules due to differences in data elements
P20W Data Warehouse Project
Conceptual Data Flow

- **Staging**
  - Data is organized in a manner similar to the source; contains PII

- **MDM**
  - Identity Matching, Linking, and Cohort Management Processes
  - P20W ID is assigned using PII data kept outside the system

- **P20W Data Model**
  - Data is organized into common tables; No PII

- **Data Warehouse**
  - Data is organized optimally for analysis and reporting; No PII

- **Output for users**
  - Data Marts
  - Feedback Reports
  - Research Briefs
  - Data Sets

Output has no PII or P20W ID, Research ID only
Why Informatica?

• Their MDM Hub provides a central repository of identifiers (e.g. names, DOBs, SSNs) over time for every source.
• Provides deterministic and probabilistic matching tools for matching source data.
• Has a mechanism for assigning P20IDs (Person IDs) to collections of data.
• Preserves histories of merges. Has mechanism for unmerging data.
• Other non-identity matching criteria for using Informatica: PowerCenter dataflow programming language, Metadata Manager, etc.
Overview of Identity Matching Process

• Data moves from source files to Stage database. Identifiers are cleaned and standardized. Tokens are assigned (more on tokens later) to each record.
  • Names standardized using a set of rigorous set of business rules incorporated into an Informatica “mapplet.”

• Data moves from Stage database, to MDM Hub.

• In MDM Hub, new data (tokens) are merged with existing data.
  • If a person exists already in data warehouse, their new data (tokens) are assigned their existing P20ID.
  • New people get new P20IDs. Their new data (tokens) are assigned to their data (tokens).
Matching Process: Automerge First

• Data is first merged using “Automerge” rules.
• Automerge rules are conservative rules for merging identities, merging P20IDs.
• As ERDC has implemented them, they largely deterministic, similar to using SQL to merge data.
• False negative rates not too important. Rather, Automerge rules are designed to ensure a very, very low false positive rates.
# Automerge Sample Rule Set

- Rule set for automerging data:

<table>
<thead>
<tr>
<th>Rule Number</th>
<th>Search Level</th>
<th>Matching Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Typical</td>
<td>Gender, DOB, Complete Person Name (fuzzy)</td>
</tr>
<tr>
<td>2</td>
<td>Typical</td>
<td>Complete Person Name (fuzzy), DOB, College and Student ID</td>
</tr>
<tr>
<td>3*</td>
<td>Typical</td>
<td>Gender, DOB, First Name, Complete Person Name (fuzzy)</td>
</tr>
</tbody>
</table>

* As of 9/6/2013, rule still undergoing validation.
Matching Process: Manual Merge Last

- Data is next subjected to a manual merge rule set.
- Manual merge rule sets are looser. They are designed to so that the false negative rate is low. False positives are not much of a concern (since potential matches will be manually reviewed).
- Manual merge rule sets creates a manual review table of pairs of identities.
- This table is brought into Excel, and a human evaluates each pair of identities to determine if there is a match or not.*
- Matches are uploaded to the MDM hub, and results merged.

* ERDC custom process. Informatica comes with a tool for side-by-side evaluation of match pairs, but side-by-side comparisons do not scale well.
Manual Merge Sample Rule Set

- Rule set for creating potential match pairs:

<table>
<thead>
<tr>
<th>Rule Number</th>
<th>Search Level</th>
<th>Matching Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Typical</td>
<td>Complete Person Name (fuzzy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>College and Student ID</td>
</tr>
<tr>
<td>2</td>
<td>Typical</td>
<td>Complete Person Name (fuzzy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSN</td>
</tr>
<tr>
<td>3</td>
<td>Typical</td>
<td>Complete Person Name (fuzzy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DOB</td>
</tr>
</tbody>
</table>
### Manual Review in Excel Example

**Pairwise evaluation of matches:**

<table>
<thead>
<tr>
<th>Match</th>
<th>P20ID</th>
<th>Class</th>
<th>LastName</th>
<th>FirstName</th>
<th>MiddleName</th>
<th>DOB</th>
<th>College ID</th>
<th>SSN</th>
<th>GENDER</th>
<th>COUNTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1392</td>
<td>1</td>
<td>OVAK</td>
<td>STEPHEN</td>
<td>J</td>
<td>9/21/1987</td>
<td>111</td>
<td>55555555</td>
<td>M</td>
<td>Grays Harbor</td>
</tr>
<tr>
<td></td>
<td>92827</td>
<td>1</td>
<td>OVAK</td>
<td>STEPHEN</td>
<td>J</td>
<td>9/21/1987</td>
<td>111</td>
<td>55555555</td>
<td>M</td>
<td>Jefferson</td>
</tr>
<tr>
<td>0</td>
<td>27284</td>
<td>2</td>
<td>I Li</td>
<td>LISA</td>
<td>M</td>
<td>9/14/1989</td>
<td>234</td>
<td></td>
<td>F</td>
<td>Douglas</td>
</tr>
<tr>
<td></td>
<td>43323</td>
<td>2</td>
<td>I Li</td>
<td>LISA</td>
<td>M</td>
<td>9/14/1989</td>
<td>233</td>
<td></td>
<td>F</td>
<td>Douglas</td>
</tr>
<tr>
<td>1</td>
<td>767</td>
<td>3</td>
<td>COOK</td>
<td>ALICE</td>
<td>M</td>
<td>12/13/1990</td>
<td>123</td>
<td>53333333</td>
<td>F</td>
<td>Jefferson</td>
</tr>
<tr>
<td></td>
<td>28579</td>
<td>3</td>
<td>COOK</td>
<td>ALICE</td>
<td>M</td>
<td>12/13/1990</td>
<td>123</td>
<td></td>
<td>F</td>
<td>Jefferson</td>
</tr>
<tr>
<td>0</td>
<td>767</td>
<td>4</td>
<td>COOK</td>
<td>ALICE</td>
<td>M</td>
<td>12/13/1990</td>
<td>123</td>
<td>53333333</td>
<td>F</td>
<td>Jefferson</td>
</tr>
<tr>
<td></td>
<td>46342</td>
<td>4</td>
<td>COOK</td>
<td>ALICE</td>
<td>LUDMILA</td>
<td>12/13/1990</td>
<td>124</td>
<td></td>
<td>F</td>
<td>Jefferson</td>
</tr>
</tbody>
</table>
Merging People, Easy; Unmerging, Hard

• Identities are easy to merge.
  • For example, to merge information for Jane Smith with Jan Smith just assign all instances of $P20ID_{Jane\ Smith}$ to $PersonID_{Jan\ Smith}$.

• Unmerging is hard, even impossible.
  • Say Jane and Jan are now found to really be two people. Now their data is comingled. How can you tease their data apart?

• How then can a data warehouse unmerge these two identities? Use tokens!
Unmerging: Made Possible Using Tokens

- Tokens are a set of identifiers that clumps data into the big lumps where *each token is guaranteed to only be associated with only one person*.

- K12 Token components example:
  - K12 School District ID: 17892
  - K12 Student ID: 0014353
  - Name: Jane Smith
  - DOB: 1/5/1995

- P20IDs are composed of one to many Tokens. Merging and unmerging Tokens into existing and new P20IDs is what the MDM Hub does.
Unmerging Example, the Problem

• ERDC initially thought that Jane Smith and Jan Smith were one and the same, so they were merged under P20ID = 354:

<table>
<thead>
<tr>
<th>P20ID</th>
<th>Source</th>
<th>TokenID Definition*</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>354</td>
<td>OSPI</td>
<td>Jane-Smith</td>
<td>1/5/1995</td>
</tr>
<tr>
<td>354</td>
<td>OSPI</td>
<td>Jane-Smith</td>
<td>1/5/1995</td>
</tr>
<tr>
<td>354</td>
<td>SBCTC</td>
<td>Jan-Smith</td>
<td>1/5/1995</td>
</tr>
</tbody>
</table>

* Actual TokenIDs in the data warehouse are surrogate keys. These surrogate keys are tied to the TokenID definitions that exist only in the MDM Hub.

• But later, ERDC realized that Jane Smith and Jan Smith were not one and the same.
Unmerging Example, the Solution

• Using tokens, it is straightforward to separate all the data associated with Jane Smith and assign a new P20ID to that data:

<table>
<thead>
<tr>
<th>P20ID</th>
<th>Source</th>
<th>TokenID Definition</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>354</td>
<td>OSPI</td>
<td>Jane-Smith</td>
<td>1/5/1995</td>
</tr>
<tr>
<td>354</td>
<td>OSPI</td>
<td>Jane-Smith</td>
<td>1/5/1995</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P20ID</th>
<th>Source</th>
<th>TokenID Definition</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>57289</td>
<td>SBCTC</td>
<td>Jan-Smith</td>
<td>1/5/1995</td>
</tr>
</tbody>
</table>

• Jane Smith and Jan Smith are now unmerged.
Tokens in the MDM Hub Database

e.g. Jane Smith, 1/1/1995, 111, W4543935
Other Sources of Data

- Marriage and Divorce data from DOH
- Name change data from AOC
- Death data from DOH and SSA
Questions?
Contact Us

Washington Education Research & Data Center
www.erdc.wa.gov

John Sabel john.sabel@ofm.wa.gov