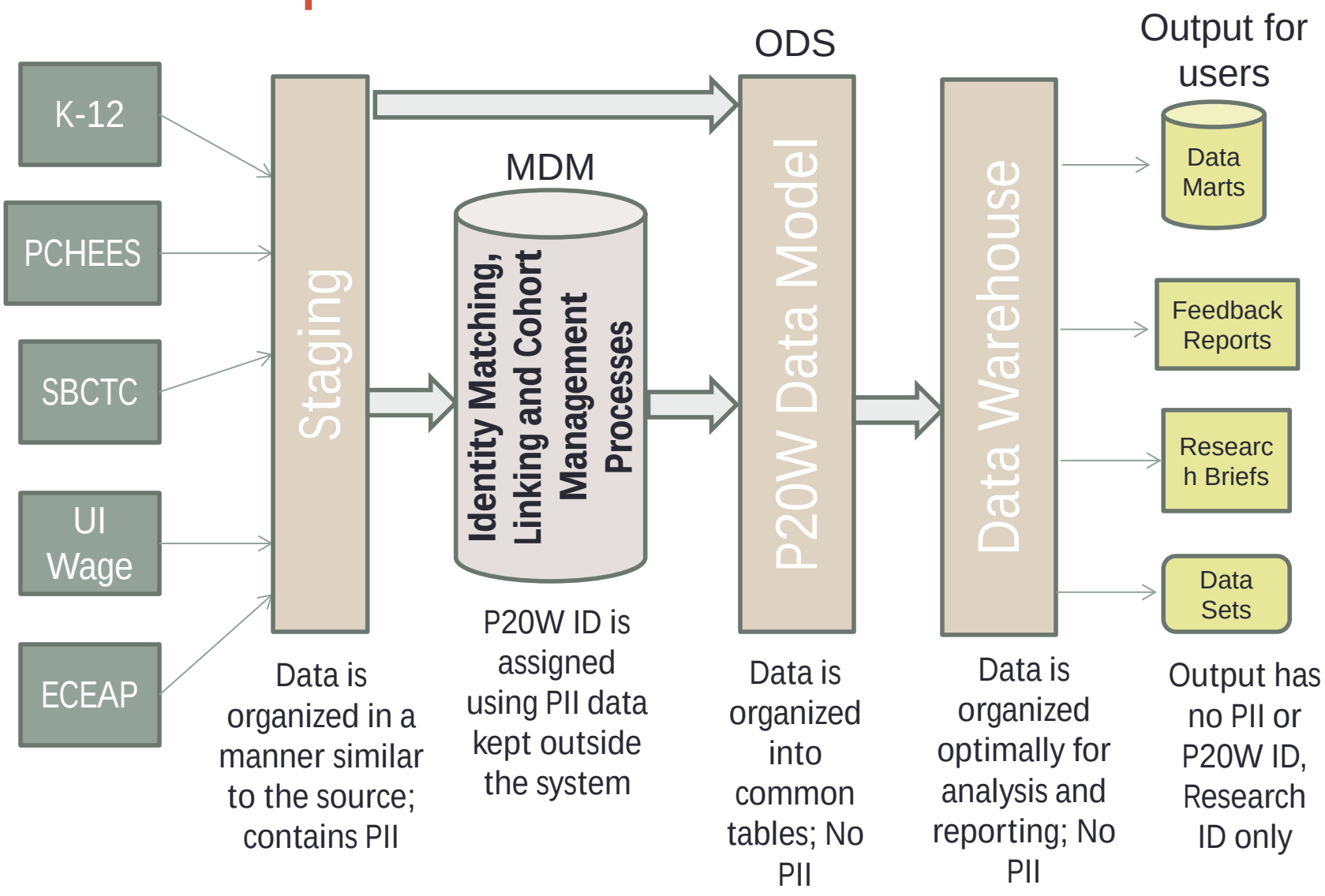
IDENTITY MATCHING AT ERDC

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



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:

Rule Number	Search Level	Matching Columns
1	Typical	Gender
		DOB
		Complete Person Name (fuzzy)
		SSN
2	Typical	Complete Person Name (fuzzy)
		DOB
		College and Student ID
3*	Typical	Gender
		DOB
		First Name
		Complete Person Name (fuzzy)

* 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:

Rule Number	Search Level	Matching Columns
1	Typical	Complete Person Name (fuzzy)
		College and Student ID
2	Typical	Complete Person Name (fuzzy)
		SSN
3	Typical	Complete Person Name (fuzzy)
		DOB

Manual Review in Excel Example

- Pairwise evaluation of matches:

Match	P20ID	Class	LastName	FirstName	MiddleName	DOB	College ID	SSN	GENDER	COUNTY
1	1392	1	OVAK	STEPHEN	J	9/21/1987	111	55555555	M	Grays Harbor
	92827	1	OVAK	STEPHEN	J	9/21/1987	111	55555555	M	Jefferson
0	27284	2	ILI	LISA		9/14/1989	234		F	Douglas
	43323	2	ILI	LISA	M	9/14/1989	233		F	Douglas
1	767	3	COOK	ALICE	M	12/13/1990	123	53333333	F	Jefferson
	28579	3	COOK	ALICE		12/13/1990	123		F	Jefferson
0	767	4	COOK	ALICE	M	12/13/1990	123	53333333	F	Jefferson
	46342	4	COOK	ALICE	LUDMILA	12/13/1990	124		F	Jefferson

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:

Jane Smith/Jan Smith

P20ID	Source	TokenID Definition*	Data Warehouse
354	OSPI	Jane-Smith|1/5/1995|17111|0014353	Grade 8 data
354	OSPI	Jane-Smith|1/5/1995|17892|0014353	Grade 9 to 12 data
354	SBCTC	Jan-Smith|1/5/1995|111|W4543935	SBCTC data

* 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:

Jane Smith

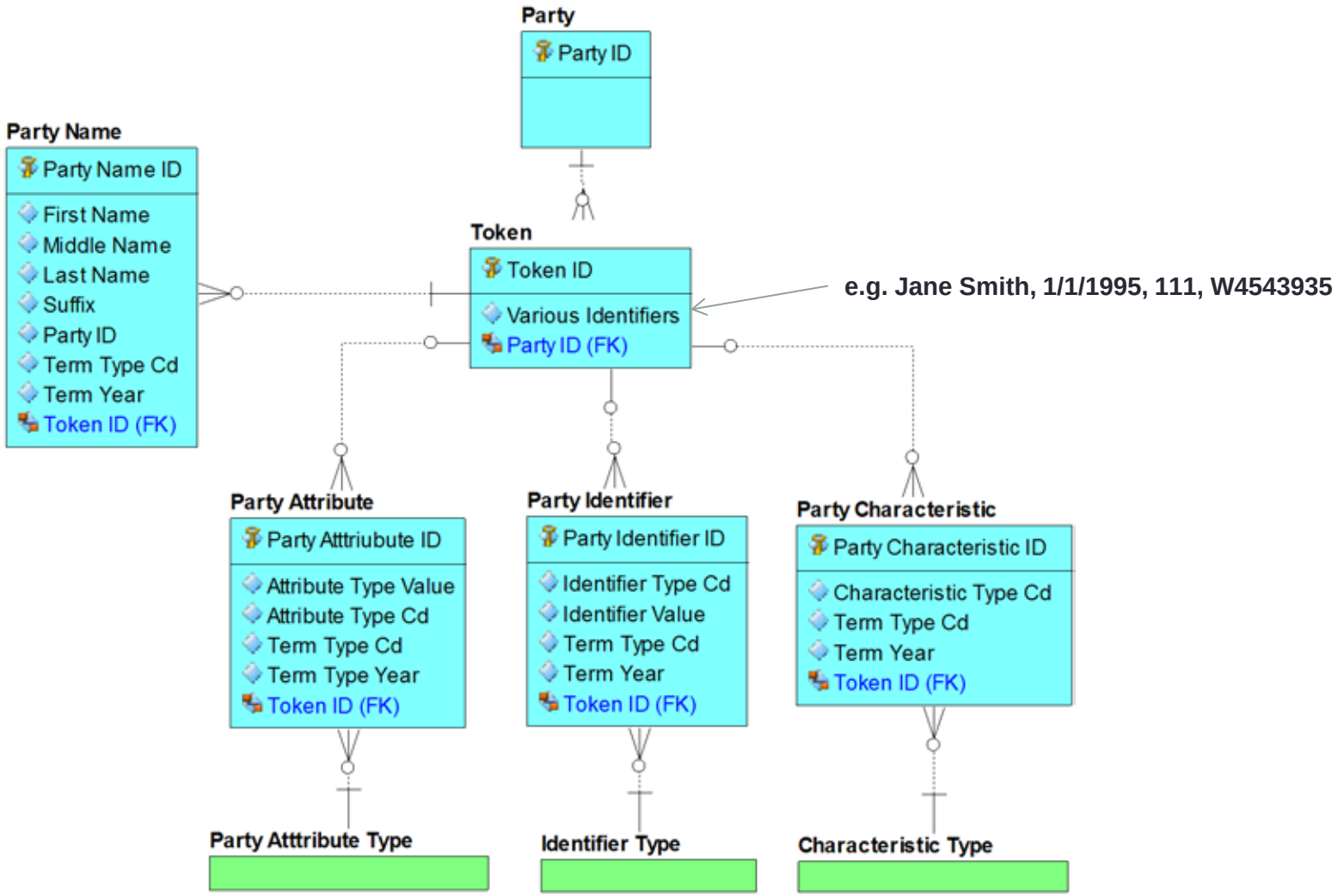
P20ID	Source	TokenID Definition	Data Warehouse
354	OSPI	Jane-Smith|1/5/1995|17111|0014353	Grade 8 data
354	OSPI	Jane-Smith|1/5/1995|17892|0014353	Grade 9 to 12 data

Jan Smith

P20ID	Source	TokenID Definition	Data Warehouse
57289	SBCTC	Jan-Smith|1/5/1995|111|W4543935	SBCTC data

- Jane Smith and Jan Smith are now unmerged.

Tokens in the MDM Hub Database



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