

Febri – A Freely Available Record Linkage System with a Graphical User Interface

Peter Christen

Department of Computer Science,
Faculty of Engineering and Information Technology,
ANU College of Engineering and Computer Science,
The Australian National University

Contact: peter.christen@anu.edu.au

Project Web site: <http://datamining.anu.edu.au/linkage.html>

Funded by the Australian National University, the NSW Department of Health,
and the Australian Research Council (ARC) under Linkage Project 0453463.

Outline

- What is record linkage?
- Record linkage techniques
- The record linkage process
- Overview of *Febrl*
- The *Febrl* graphical user interface (GUI)
- An example *Febrl* GUI screenshot
- Practical demonstration of *Febrl*
- Outlook and future work

What is record (or data) linkage?

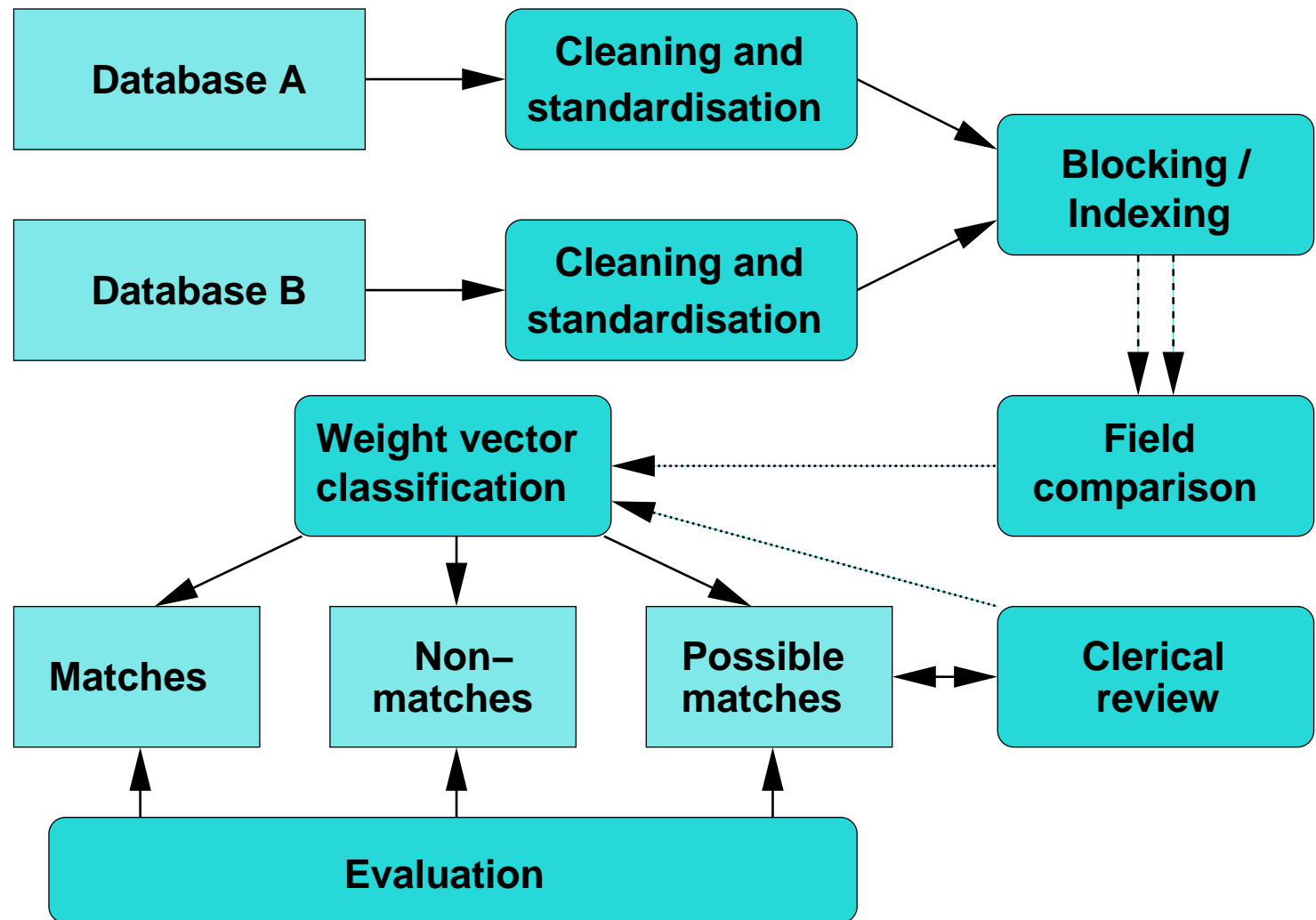
- The process of linking and aggregating records from one or more data sources representing the same entity (such as a patient, customer, or business)
- Also called *data matching*, *data integration*, *data scrubbing*, *entity resolution*, *object identification*, *merge-purge*, etc.
- Challenging if no unique entity identifiers available
For example, which of these three records refer to the same person?

<i>Dr Smith, Peter</i>	<i>42 Miller Street 2602 O'Connor</i>
<i>Pete Smith</i>	<i>42 Miller St, 2600 Canberra A.C.T.</i>
<i>P. Smithers</i>	<i>24 Mill Street; Canberra ACT 2600</i>

Record linkage techniques

- Deterministic linkage
 - Exact linkage (if a *unique* entity identifier of high quality is available: has to be precise, robust, stable over time)
Examples: *Medicare*, *ABN* or *Tax file number* (?)
 - Rules based linkage (complex to build and maintain)
- Probabilistic linkage (*Fellegi and Sunter, 1969*)
Use available (personal) information for linkage (which can be missing, wrong, coded differently, and/or out-of-date)
Examples: *names*, *addresses*, *dates of birth*, etc.
- Modern approaches
Based on machine learning, data mining, artificial intelligence, and information retrieval techniques

The record linkage process



Overview of Febrl

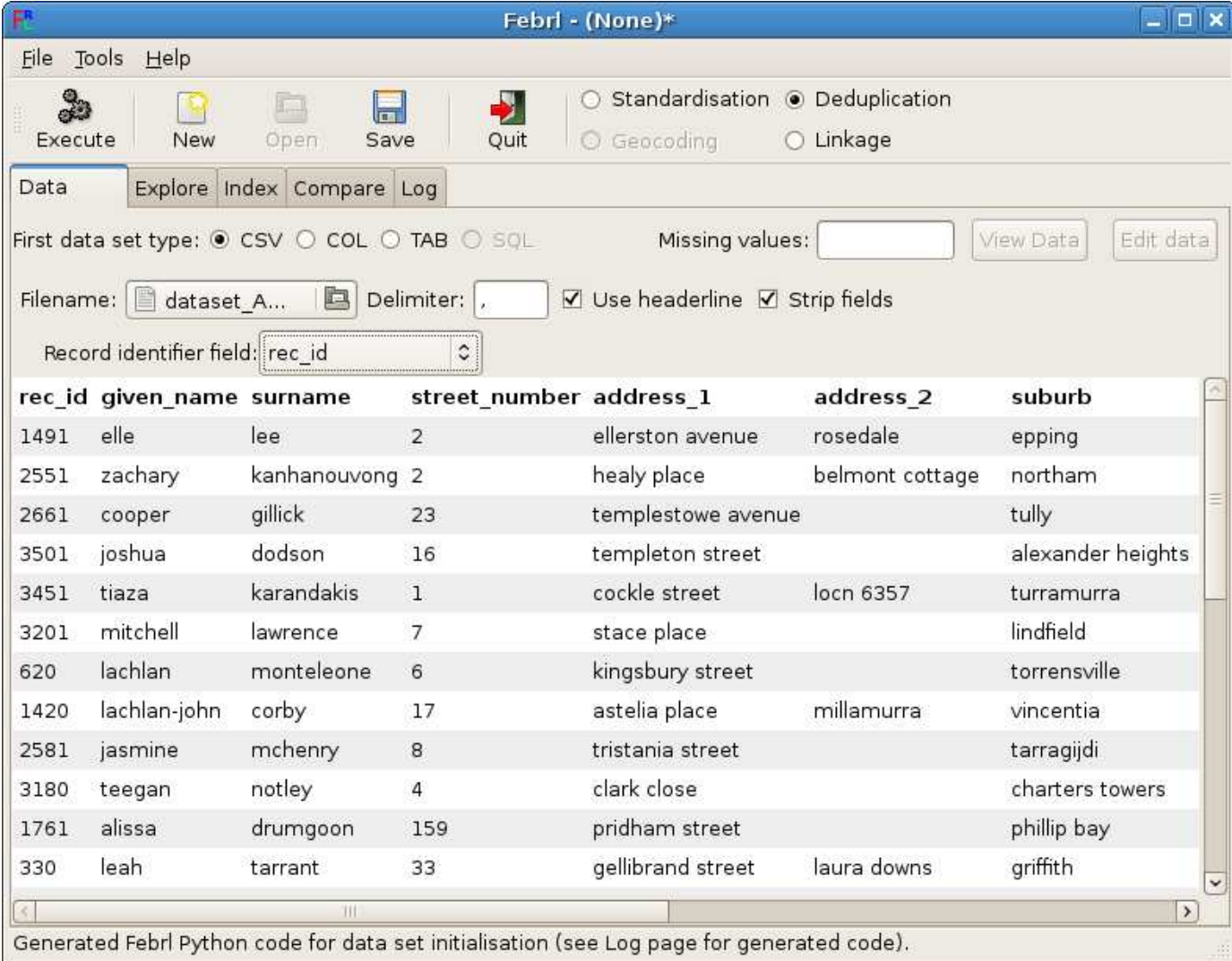
- Has been developed since 2002
(as part of an ARC Linkage Project between the ANU and the NSW Department of Health)
- Is implemented in Python (a freely available object-oriented programming language)
- Its source code is available
- Includes many recently developed algorithms and techniques for indexing (blocking), field comparisons and record pair classification
- An ideal tool to learn about record linkage
- Is freely available at *Sourceforge.net*

<https://sourceforge.net/projects/febrl/>

The Febrl graphical user interface

- A page (tab) based approach with one page per major step of the record linkage project
- Three different project types
 - Clean and standardise one data set
 - Deduplicate one data set
 - Link two data sets
- Clicks on 'Execute' will validate and confirm the settings on a GUI page
- *Febrl* Python code will be generated (and can be run outside the GUI)
- GUI structure similar to the *Rattle* open source data mining tool (<http://rattle.togaware.com>)

An example Febrl GUI screenshot



The screenshot shows the Febrl GUI window titled "Febrl - (None)*". The interface includes a menu bar (File, Tools, Help) and a toolbar with icons for Execute, New, Open, Save, and Quit. Below the toolbar, there are radio buttons for "Standardisation" (unselected) and "Deduplication" (selected), and "Geocoding" (unselected) and "Linkage" (unselected). The "Data" tab is active, with sub-tabs for "Explore", "Index", "Compare", and "Log".

Configuration options include:

- First data set type: CSV, COL, TAB, SQL
- Missing values:
- View Data and Edit data buttons
- Filename: Delimiter: Use headerline Strip fields
- Record identifier field:

The main data table is displayed with the following columns and rows:

rec_id	given_name	surname	street_number	address_1	address_2	suburb
1491	elle	lee	2	ellerston avenue	rosedale	epping
2551	zachary	kanhanouvong	2	healy place	belmont cottage	northam
2661	cooper	gillick	23	templestowe avenue		tully
3501	joshua	dodson	16	templeton street		alexander heights
3451	tiaza	karandakis	1	cockle street	locn 6357	turrumurra
3201	mitchell	lawrence	7	stace place		lindfield
620	lachlan	monteleone	6	kingsbury street		torrensville
1420	lachlan-john	corby	17	astelia place	millamurra	vincentia
2581	jasmine	mchenry	8	tristania street		tarragijdi
3180	teegan	notley	4	clark close		charters towers
1761	alissa	drumgoon	159	pridham street		phillip bay
330	leah	tarrant	33	gellibrand street	laura downs	griffith

Generated Febrl Python code for data set initialisation (see Log page for generated code).

Outlook and future work

- *Febri* is a tool suitable for both practitioners and new record linkage users
- Contains many different linkage techniques
- Allows small to medium sized experimental standardisations, deduplications and linkages
- Can be used alongside commercial linkage systems for comparative linkage studies
- Future work on *Febri*
 - Include *Febri* geocoding module into the GUI
 - Include privacy-preserving record linkage techniques
 - Add new indexing, comparisons and classification techniques as they become available