The dGrail Toolkit for Iterative Deterministic Record Linkage
Jeremy Ey, Andrew Walker, Douglas A. Talbert, PhD
Tennessee Technological University, Cookeville, TN

Abstract
The distributed Graph rule automata for iterative linkage (dGrail) toolkit is a software package for deterministic record linkage. This toolkit allows for iterative development of linked record sets. While intended for the generation of gold standard record sets, the toolkit is applicable to any offline deterministic record linkage task. The dGrail toolkit embodies a flexible rule engine allowing the user to implement a wide variety of record matching rules, including those found in the literature.

Introduction
Gold standard record sets are vital for the proper measurement of record linkage algorithm performance. However the creation of gold standard record sets is a time consuming and difficult process. In many cases gold standard sets are either absent or undersized. The distributed Graph rule automata for iterative linkage (dGrail) toolkit is designed to provide a configurable system for deterministic record linkage. Through a set of user definable conditions and a user definable rule graph, record pairs are determined to be of match, non-match, or unknown status. This system is designed to allow users to develop and test novel linkage rules as well as implement a variety of deterministic rules found in the record linkage literature. In order to boost performance of the system, dGrail supports the distribution of record pair comparison across systems.

Condition and Rule Graph Definition
The classification of record pairs into the match, non-match, or unknown status is performed by a graph based rule engine. This rule engine is supplied with a list of defined conditions and a acyclic, non-deterministic graph in the dot format. Graph nodes represent conditions; each condition is associated with a label. Figure 1 shows the graph structure for Grannis and Overhage’s NYSIIS rule. In this rule a record is said to match if the social security number (A), NYSIIS encoding of the first name (B) and gender (C) are exact match and at least one of the day (D), month (E), or year (F) of birth match.

When evaluated by the processing unit, each condition is evaluated to either true or false until all possible paths have reached a result node. Conditions are defined based on exact match of a field, match of an encoding of a field (e.g. Soundex), or meeting the minimum score on edit distance, longest common substring, or longest common subsequence.

Conclusion
The dGrail toolkit provides a system for iterative deterministic record linkage for the development of gold standard record set creation. The included rule engine can be used to either implement existing rule definitions or for the creation and evaluation of user-created rules.

Acknowledgments
This work is supported by Tennessee Tech University, Vanderbilt University, the State of Tennessee, and the Agency for Healthcare Research and Quality (AHRQ), Contract no. 290-05-0006. This publication has not been approved by AHRQ.

References