The OYSTER Open Source Project for Introducing Entity Resolution and MDM into Information Systems Curricula

TREO Talk Paper

John R. Talburt  
University of Arkansas at Little Rock  
jrtalburt@ualr.edu

Yumeng Ye  
University of Arkansas at Little Rock  
yye1@ualr.edu

Bingyi Zhong  
University of Arkansas at Little Rock  
bxzhong@ualr.edu

Xinning Li  
University of Arkansas at Little Rock  
xxli3@ualr.edu

James True  
University of Arkansas at Little Rock  
jltrue@ualr.edu

Sapna Srimal  
University of Arkansas at Little Rock  
sxsrimal@ualr.edu

Abstract

Entity resolution (ER), or record linking, is foundational for identity management, master data management (MDM), and data integration. Even though these topics may be covered in various parts of an IS curriculum, these modules often lack a hands-on component because appropriate software is easily available. Even when an IT vendor provides an educational license, students are exposed to only one design and implementation of ER and MDM. OYSTER is an open source Java platform written specifically to support teaching and research in entity resolution, identity management, and MDM. OYSTER has been used to teach a graduate-level course in ER each year since 2010 and has been downloaded from SourceForge.net thousands of times.

Some of the features supported in the current 3.5 version of the software (now resident on BitBucket.net) are deterministic (Boolean) matching, probabilistic (scoring) matching at both the attribute and value level, list matching, property-value pair list matching, and multivalued attributes matrix matching. It also supports inverted index match key blocking on multiple match keys. In addition to basic deduplication (merge-purge linking) operation, the software also supports the full life cycle of identity management from identity capture through identity update and assertions for knowledgebase (graph) maintenance. All of the oyster features and functions are configurable through three XML scripts executed at run time.

New features, bug fixes, and documentation updates are continually being added by the OYSTER development team. New features in development include a logistic regression match rule and a probabilistic weight calculator. The system comes with documentation and a set of pre-configured scripts and test data to help students learn the basic operation of the system and fundamentals of ER. Instructors can also request a set of 4 files (272K records) comprising synthetic customer occupancy data. These files can be used as the basis for a team “challenge” project to link all of the same customer records within and across files. Because the challenge data was synthetically generated, teams can easily check the precision and recall of their linking results at any time through a web service that scores the results against the correct linking.

The talk will include lessons learned from using OYSTER in classes, textbooks available for covering ER and MDM topics, ideas for improving student experience with the system, and an invitation to contribute to OYSTER development and lesson development.