

**Point of Contact:** Pearl McElfish [pamcelfish@uams.edu](mailto:pamcelfish@uams.edu) 479-264-8690

**Title:** Managing Duplicate Addresses Utilizing an Open-Source Entity Resolution Tool

**Authors:** Umit Topaloglu, Ph.D.; Bo Liu; William R. Hogan, M.D., M.S

**Affiliation:** Arkansas Study Center; University of Arkansas for Medical Sciences

### **Abstract**

The National Children's Study Arkansas Study Center (ASC) uses an open-source software application called Open sYSTem Entity Resolution (OYSTER), developed at the University of Arkansas at Little Rock (available at <http://ualr.edu/eriq/downloads/>), to resolve multiple records of a participant's address. The duplicate records arise because addresses are collected from multiple sources, including instruments, the participant's healthcare provider, and other data-collection forms. The ASC conducts study instruments using the open-source application LimeSurvey. Most address information is obtained via the pregnancy screener, but other instruments also require an address if the subject has moved or plans to move. Participants' demographic information, including address, is entered and managed in caBIG Central Clinical Participant Registry (C3PR). To properly submit participant address and instrument information to Vanguard Data Repository (VDR), we must ensure that a participant's addresses recorded in these applications are resolved if duplicated. Furthermore, given that entry of address data in both applications is error prone and subject to variability (e.g., entering St. vs. Street), resolving duplicates is not straightforward and simple string matching will frequently fail to detect duplicates. OYSTER is an entity resolution system that supports probabilistic direct matching, transitive linking, and asserted linking. To facilitate prospecting for match candidates (blocking), the system builds and maintains an in-memory index of attribute values to identities. Once OYSTER identifies the duplicates, we manually resolve them in LimeSurvey and C3PR, and we are moving to an automated process.