



Nationwide, states are tasked with using and sharing mass amounts of data for the greatest good: to improve the education experience for students and better prepare them for success in the future. However, states are also tasked with protecting the anonymity and privacy of the same students to whom the data pertain. The Arkansas Research Center (ARC) has taken unprecedented strides to this end through the development of two identity management systems: OYSTER and TrustEd.

The Knowledgebase Approach

Both OYSTER and TrustEd use what is known as a “knowledgebase approach.” All “known representations” of an entity (name, Social Security Number, etc.) are stored to facilitate data matching in the future and resolve past matching errors. Representations (e.g., list both “Bob” and “Robert”) are combined into clusters, each of which has a unique ID, and a knowledgebase is built.

Cluster	Representation
KB5765	Bob Smith, CHS
KB5765	Robert Smith, Acxiom
KB5765	Bob Smith, UCA

Figure 1: Example of Knowledgebase Approach

OYSTER

OYSTER (Open-System Entity Resolution) is a knowledgebase identity management system that maintains all representations of an entity (i.e., all IDs) and generates agency-specific identifiers—thereby protecting personally identifiable information (PII). When new information is inputted into OYSTER, OYSTER uses the R-Swoosh algorithm of entity resolution (attribute-based matching) to determine if that information matches a representation already included in the knowledgebase, or if the information represents an entirely new entity.

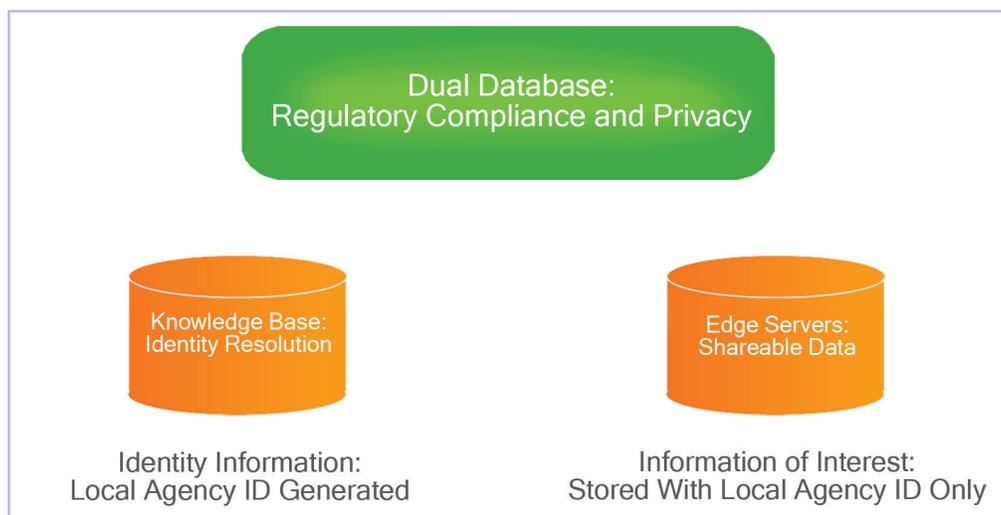


Figure 2. OYSTER Structure

This product of the Institute of Education Sciences (IES) was developed with the help of knowledgeable staff from state education agencies and partner organizations. The views expressed do not necessarily represent those of the IES SLDS Grant Program. We thank the following people for their valuable contributions:

Neal Gibson
Arkansas Research Center

Missy Cochenour
State Support Team, SLDS Grant Program

For more information on the IES SLDS Grant Program or for support with system development, please visit <http://nces.ed.gov/programs/SLDS>.

Data is de-identified and assigned an agency-specific ID (for example, PreK ID numbers begin with a 'P'). Thus, only the data of interest are stored in the agency's database. Because each agency has its own separate database and unique ID, it is impossible to automatically join two different agencies' sets of data.

When a cross-agency data request is received, OYSTER creates a temporary database of a crosswalk between local identifiers, which is used to join the databases. The results are then returned to the requesting agency without the use of personally identifiable information and the temporary database is destroyed.

OYSTER is an open source project that is available for free download in the Public Domain Clearinghouse in GRADS360° (<https://nces.grads360.org>). OYSTER is thoroughly documented, and has a graphical user interface and an XML structure. OYSTER provides direct support for assertion resolution. Users can also define multiple matching algorithm rule sets, such as exact-match, sub-string match, transpositions, and approximate match (Soundex, Edit Distance, QTR, etc.).

TrustEd

TrustEd combines Knowledgebase Identity Management (KIM) with TrustEd Identifier Management (TIM) to produce privacy-protected agency and partner databases for multi-agency data research. When source data are received, they receive a record ID, and the PII is sent to KIM. KIM then tries to match the information to an existing record; if no exact match is found, the Fellegi-Sunter method of probabilistic record linkage (record-based matching). The information is then either matched to an existing record in the database, or a new identity cluster is created.



Figure 3. TrustEd Structure

The knowledgebase ID (which points to PII) is encrypted as a KIMID (which represents the individual the PII pertains to) and sent to TIM. TIM turns the KIMID into a TIMID (an ID specific to the agency or researcher). The research data with the TIMID are then loaded into an agency-specific database (edge server). Thus, the needed data exists without PII, and there are no direct links between one agency's data and another's.

The only data TIM holds permanently are the source and record IDs it created when source data are submitted. If a research request requires integration of data from more than one agency, TIM generates a crosswalk of agency IDs that is later destroyed.

TrustEd does add some flexibility as compared to OYSTER. TIM can be configured to do multiple things, including creating a consistent individual ID that would follow the individual throughout an agency. TIM can also encrypt one-time or multi-use file(s) of that ID for research use. The user can set levels/thresholds for the consolidation rules in KIM, and the matching process can be modified to a particular user's needs or requirements for a particular data set.

Thus far, there is little documentation and no graphical user interface for TrustEd. More information can be found at <https://arc.arkansas.gov/trustEd>.

The q-Gram Tetrahedral Ratio Methodology

A working model of the q-Gram Tetrahedral Ratio matching methodology used by TrustEd can be found at <http://gregholland.com/qTR.asp>.

Greg Holland - Approximate String Compare Research q-Gram Tetrahedral Ratio - qTR - Methodology

Compare two strings...

String 1:

String 2:

- OR -

**** DETAILED COMPARISON - LYNN to LIN ****

Comparing L to LIN.

Found L in LIN. Points: 0+1=1

Comparing LY to LIN.

Not checking for additional strings containing LYNN.

Comparing Y to LIN.

Not checking for additional strings containing YNN.

Comparing N to LIN.

Found N in LIN. Points: 1+1=2

Comparing NN to LIN.

Not checking for additional strings containing NN.

Comparing N to LIN.

Found N in LIN. Points: 2+1=3

**** END OF DETAILED COMPARISON ****

LIN compared to LYNN...

L LI LIN

I IN

N

q-Gram points found 2.

Total points possible, $T_3 = 10$.

RESULT #2 - qRatio: $2 / 10 = 0.2$

CONCLUSION:

qTR = two results weighted by string length ($4 * 0.15 + 3 * 0.2$)

Figure 4. A comparison of the names "Lynn" and "Lin" using the qTR matching methodology.