OYSTER

An Open Source Tool for Master Data Management

Yinle Zhou

University of Arkansas at Little Rock

Oct 27, 2011

Agenda

- OYSTER Introduction
- OYSTER Operation Modes
- Example Runs

What is Entity Resolution (ER)?

- ER is the process of determining when two records in an information system refer to the same or to different real-world objects.

Sometimes called:
- Record de-duplication
- Record matching
- Record linking
- Co-reference problem

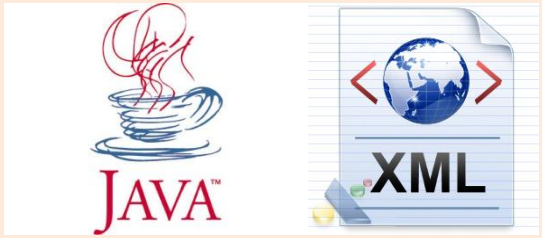
…

OYSTER

Open sYSTem for Entity Resolution

ERIQ Entity Resolution & Information Quality

http://ualr.edu/eriq



sourceforge **OYSTER Entity Resolution**

sourceforge.net/projects/oysterer

Entity Identity Information Management (EIIM)

However, Most Existing ER Systems Just Do One-time Resolution

Run1

RecID,FN,LN,DOB,SchCode
1,Edgar,Jones,20001104,G34
2,Eddie,Jones,20001104,H15
3,Super,Man,20011104,G19

→

RefID	LinkID
1	00000001
2	00000001
3	00000002

Run2

RecID,FN,LN,DOB,SchCode
1,Super,Man,20011104,G19
2,Edgar,Jones,20001104,G34
3,Eddie,Jones,20001104,H15

→

RefID	LinkID
1	00000001
2	00000002
3	00000002

…

What We Want Are Persistent Identifiers

LinkID: 00008888

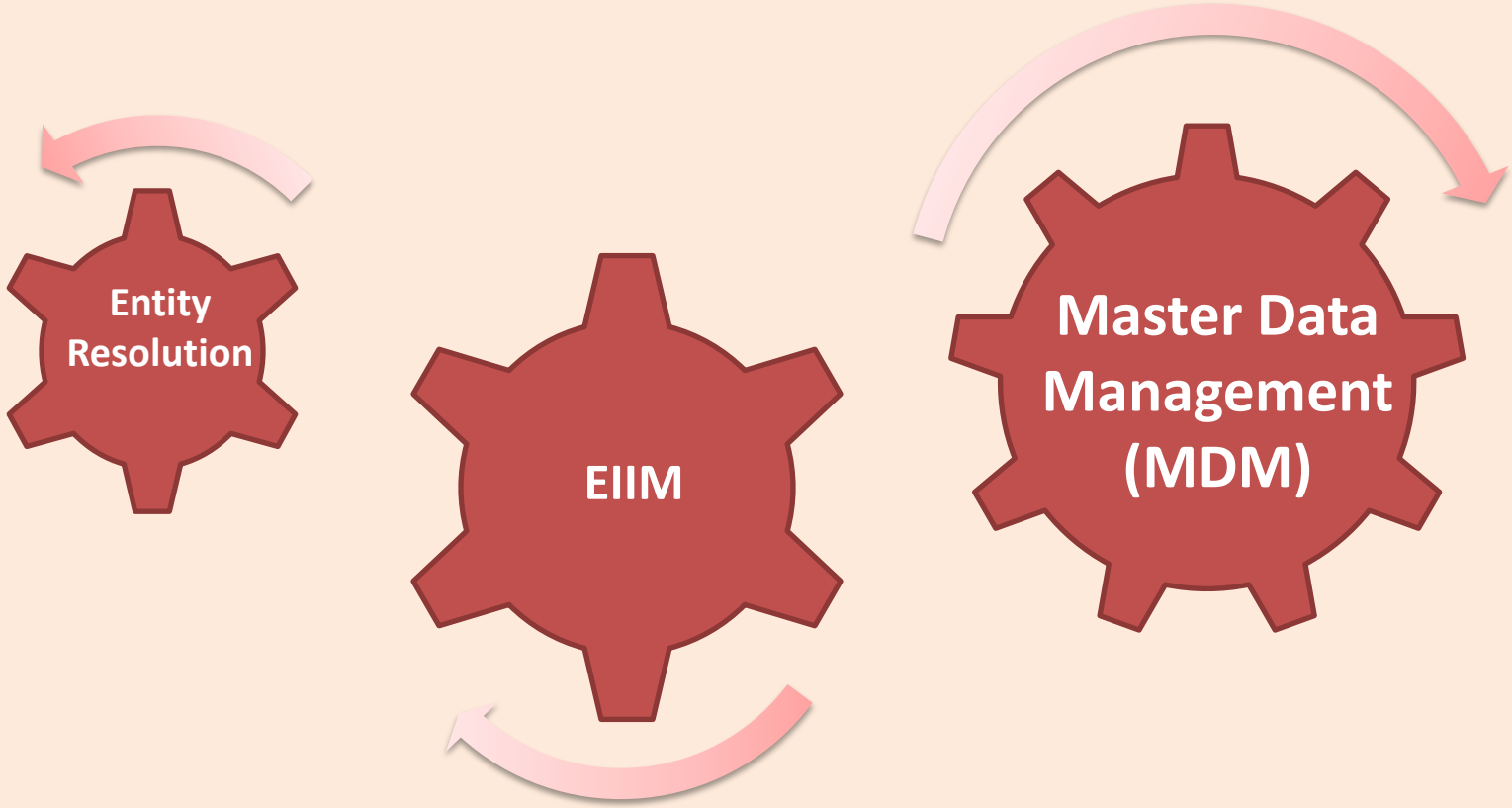
LinkID: 00008888

LinkID: 00008888

Entity Identity Information Management (EIIM)

Provide persistent entity identifiers (tokens) that do not change over time

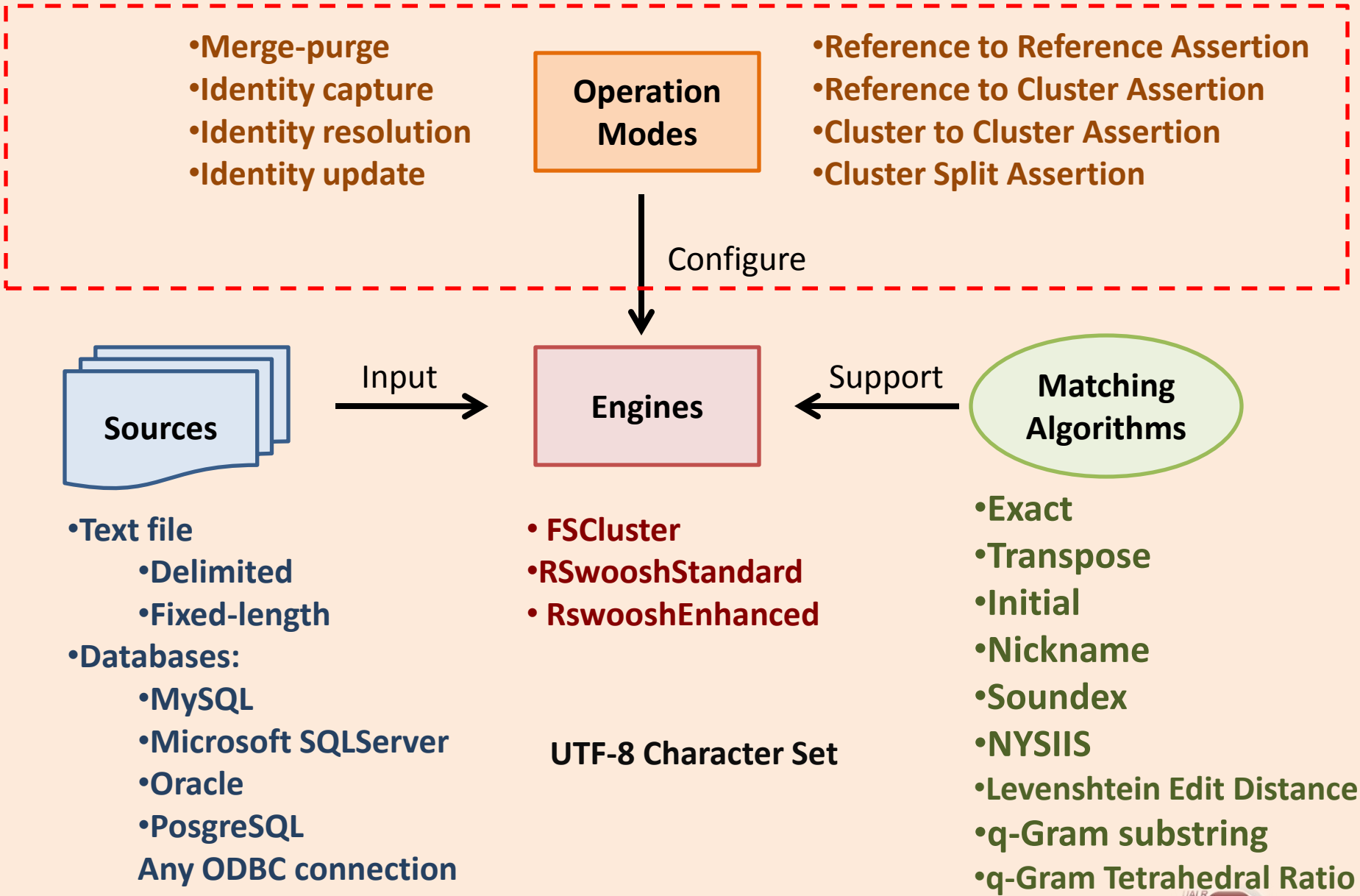
The Purpose of OYSTER Project Is to Create a Freely Available EIIM system supporting MDM



OYSTER Users

- Arkansas Department of Education
- Computing Service, University of Arkansas at Little Rock
- University of Arkansas for Medical Sciences
- Students in ER and IQ class
- Others

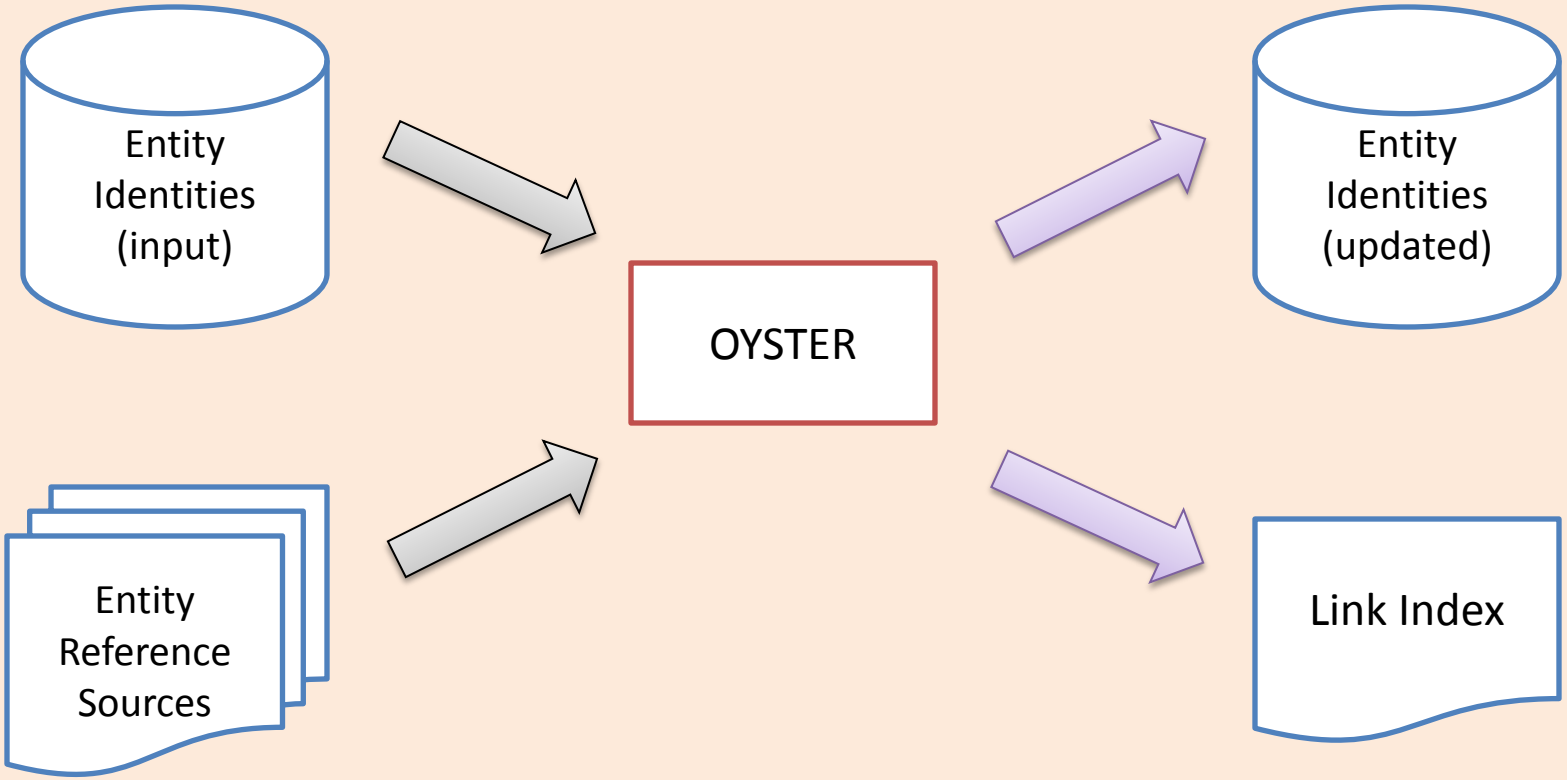
An Overview of OYSTER Components


•Merge-purge
•Identity capture
•Identity resolution
•Identity update
Operation Modes
•Reference to Reference Assertion
•Reference to Cluster Assertion
•Cluster to Cluster Assertion
•Cluster Split Assertion
Configure
Sources
Input
Engines
Support
Matching Algorithms
•Text file
•Delimited
•Fixed-length
•Databases:
•MySQL
•Microsoft SQLServer
•Oracle
•PosgreSQL
Any ODBC connection
• FSCluster
•RSwooshStandard
• RswooshEnhanced
UTF-8 Character Set
•Exact
•Transpose
•Initial
•Nickname
•Soundex
•NYSIIS
•Levenshtein Edit Distance
•q-Gram substring
•q-Gram Tetrahedral Ratio
…

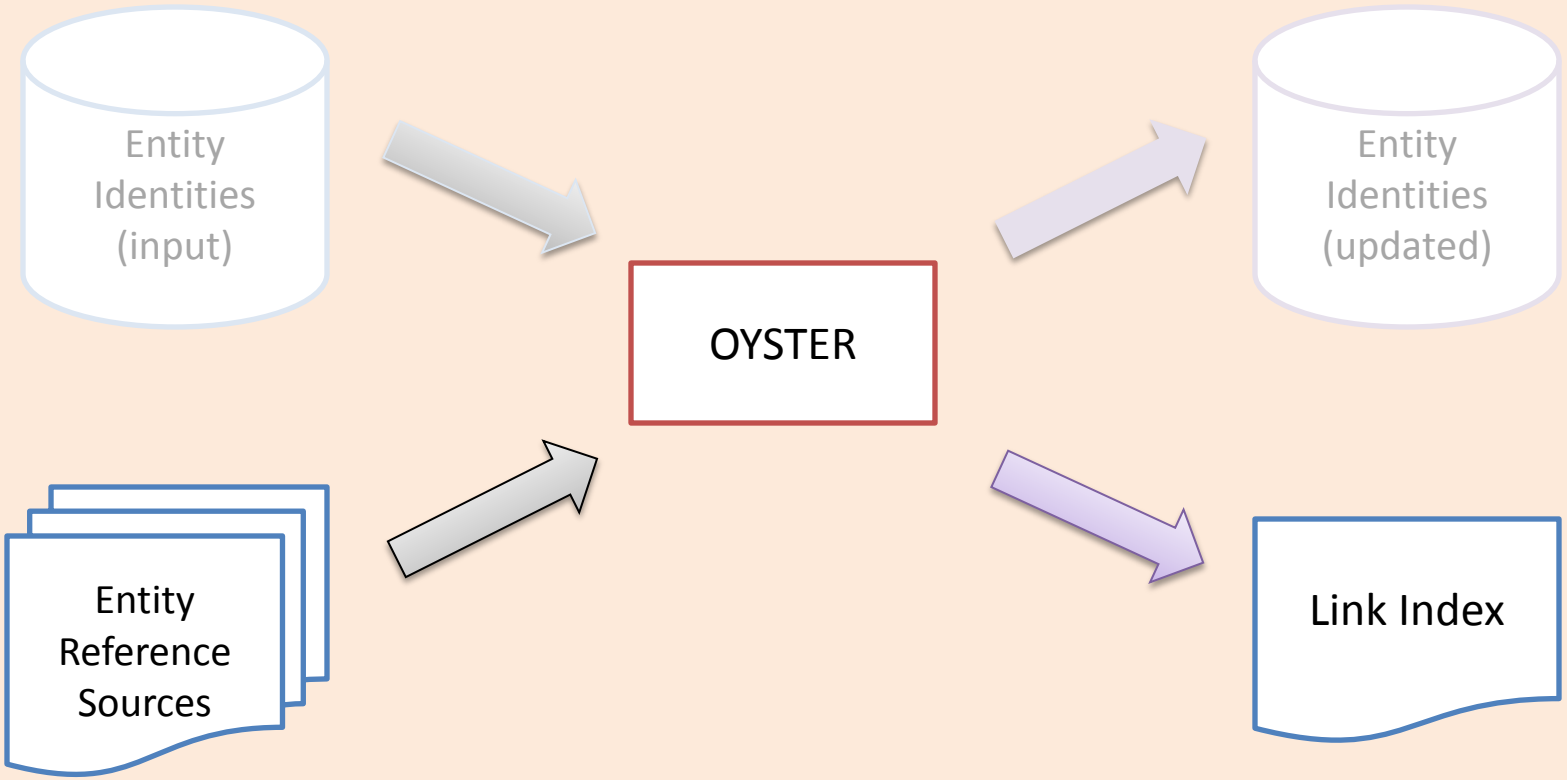
OYSTER 3.1 Can Be Configured to

- Merge-purge Mode
- Identity Capture Mode
- Identity Resolution Mode
- Identity Update Mode

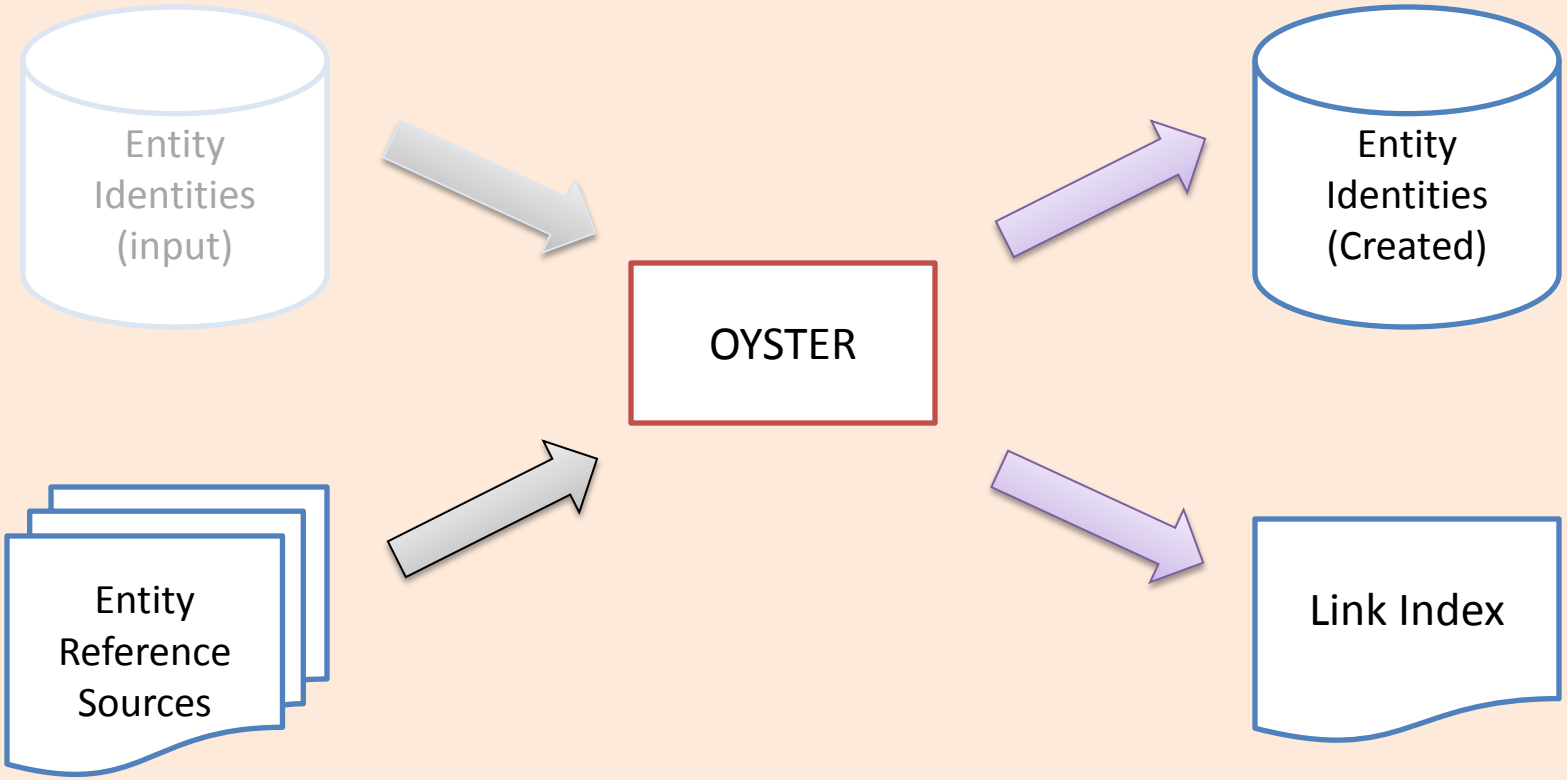
Dataflow of an OYSTER Run



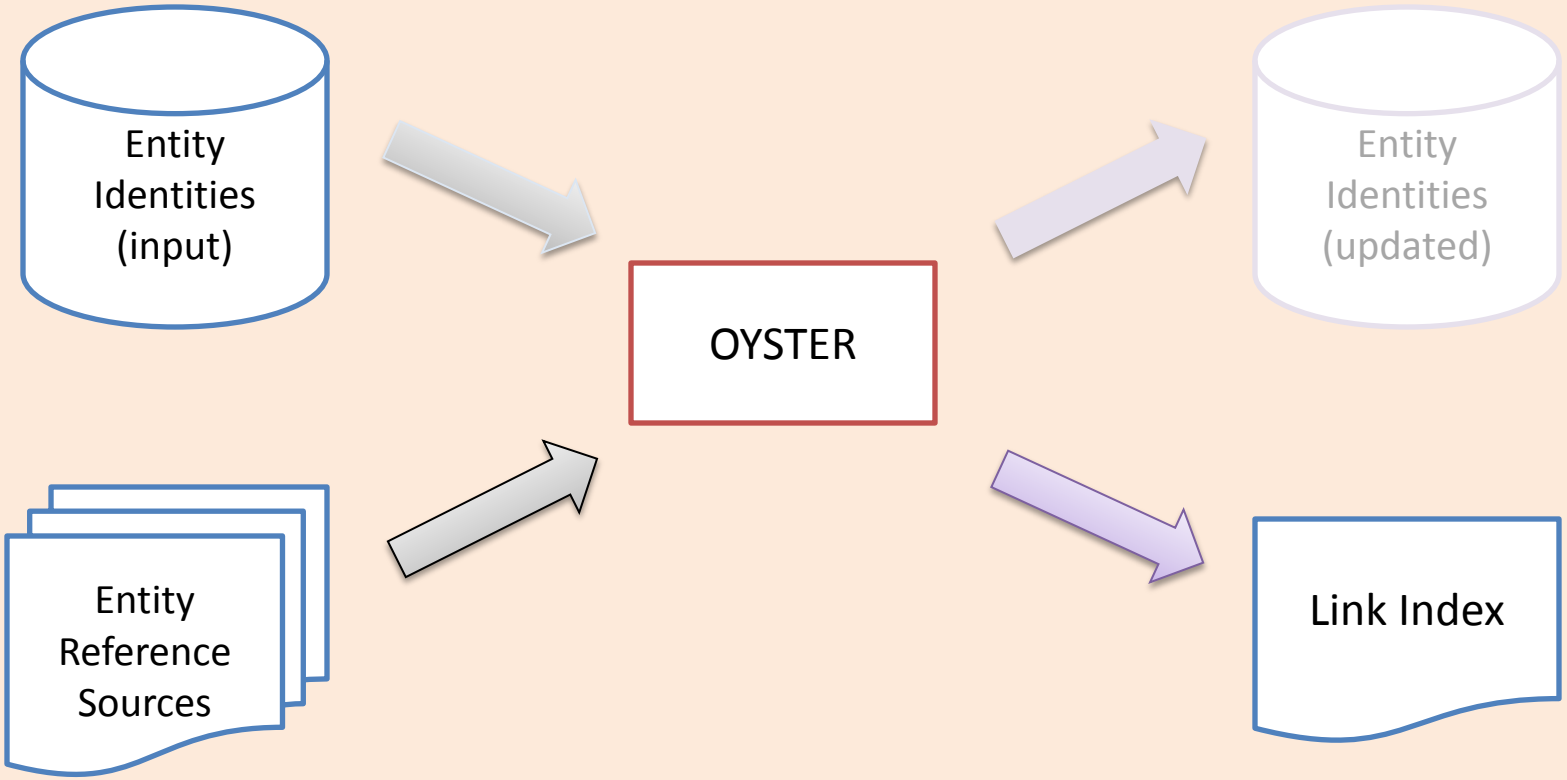
Merge-Purge Mode



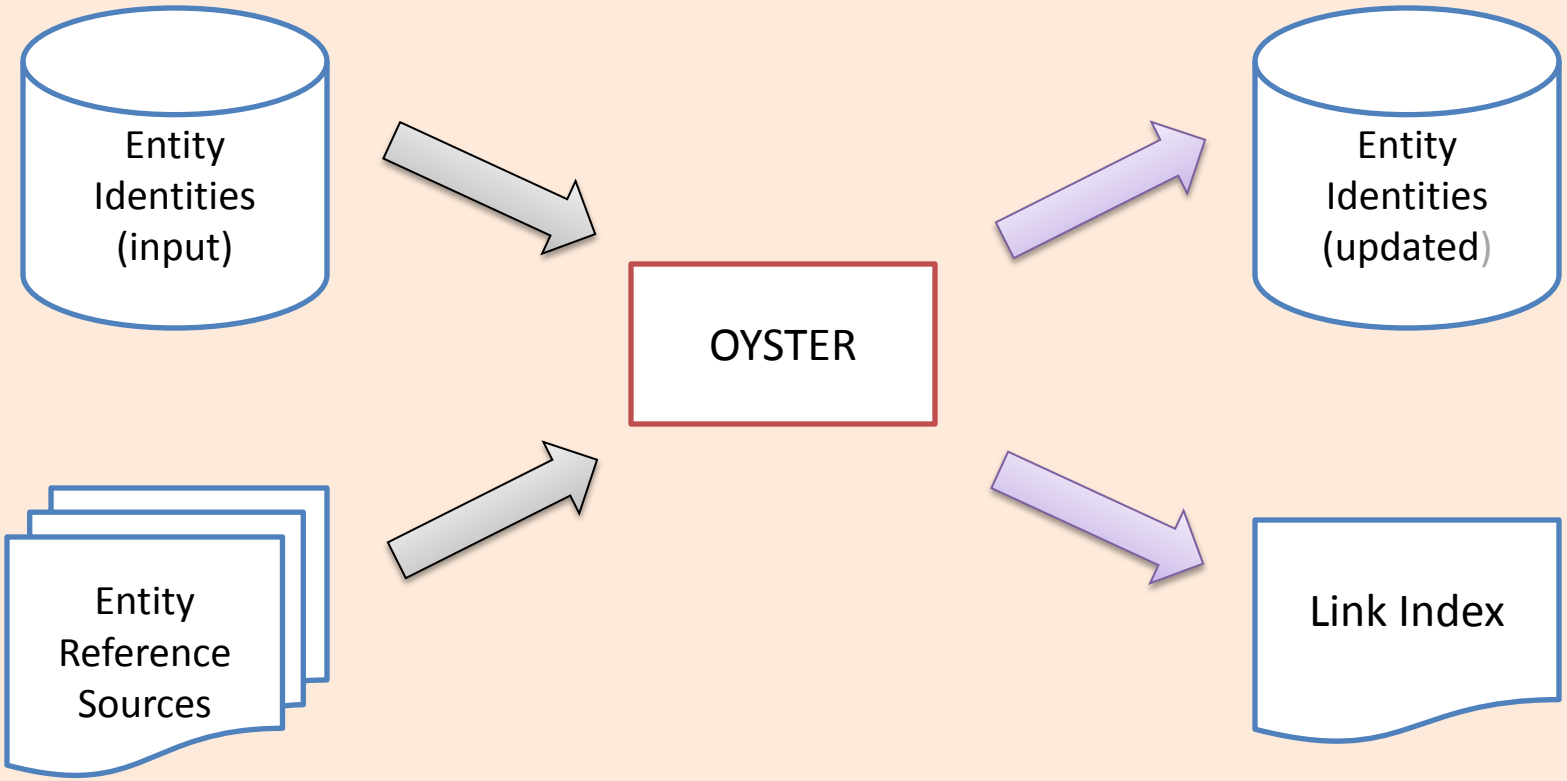
Identity Capture Mode



Identity Resolution Mode



Identity Update Mode

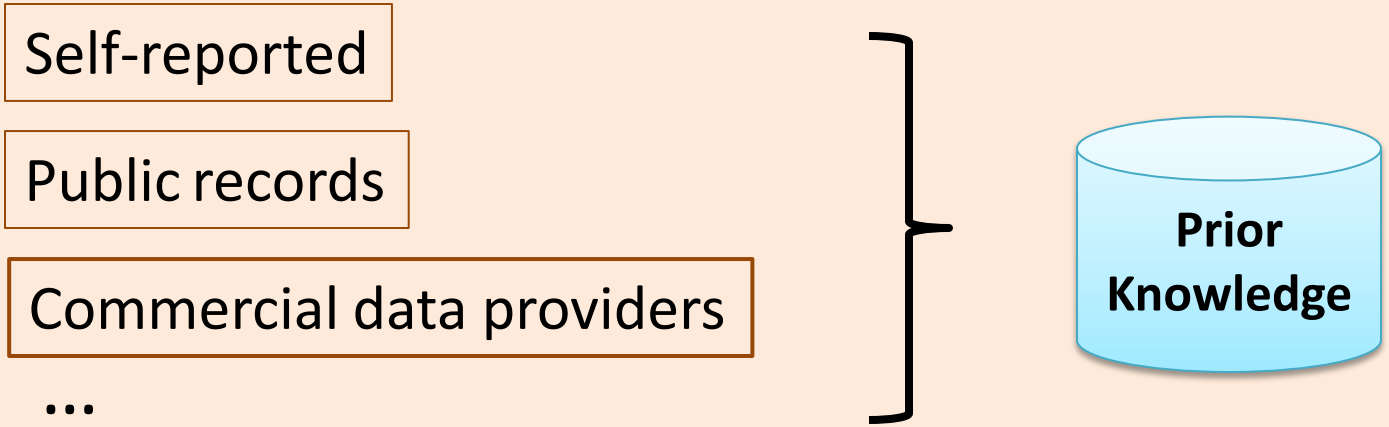


OYSTER 3.2 Adds Assertion Modes

- Reference to reference assertion mode
- Reference to cluster assertion mode
- Cluster to cluster assertion mode
- Cluster split assertion mode

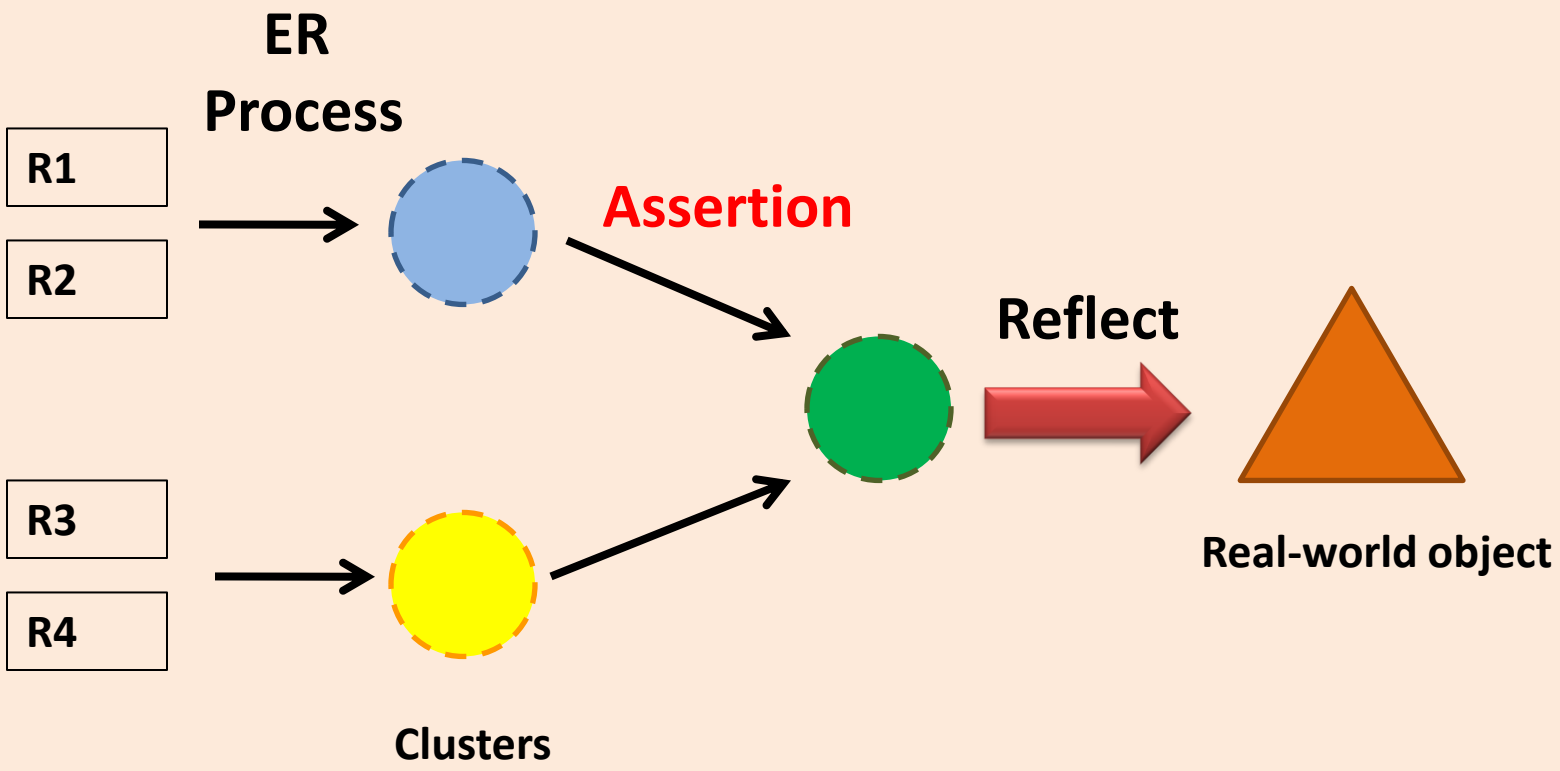
Asserted Resolution

- Employs a priori knowledge from an external source to determine whether two references are equivalent or not.



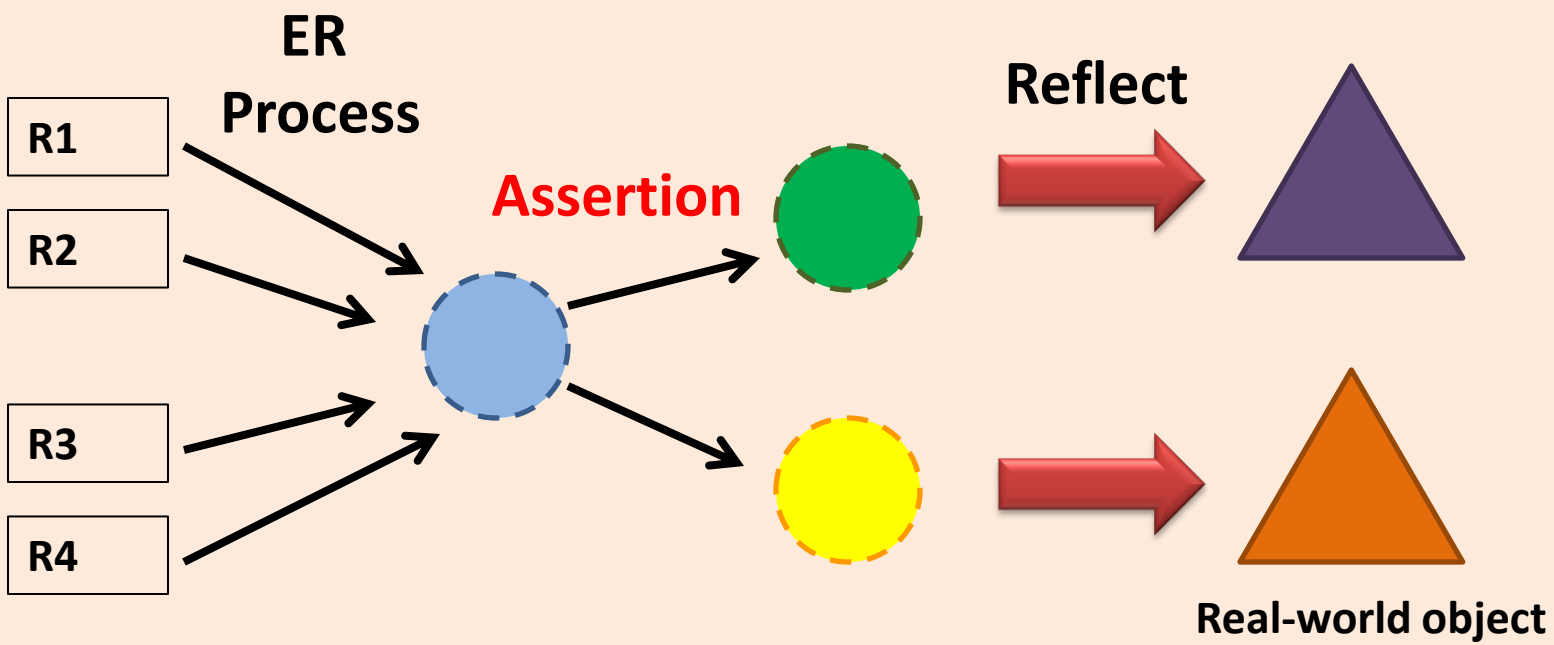
- Knowledge-based resolution

Repair False Negative



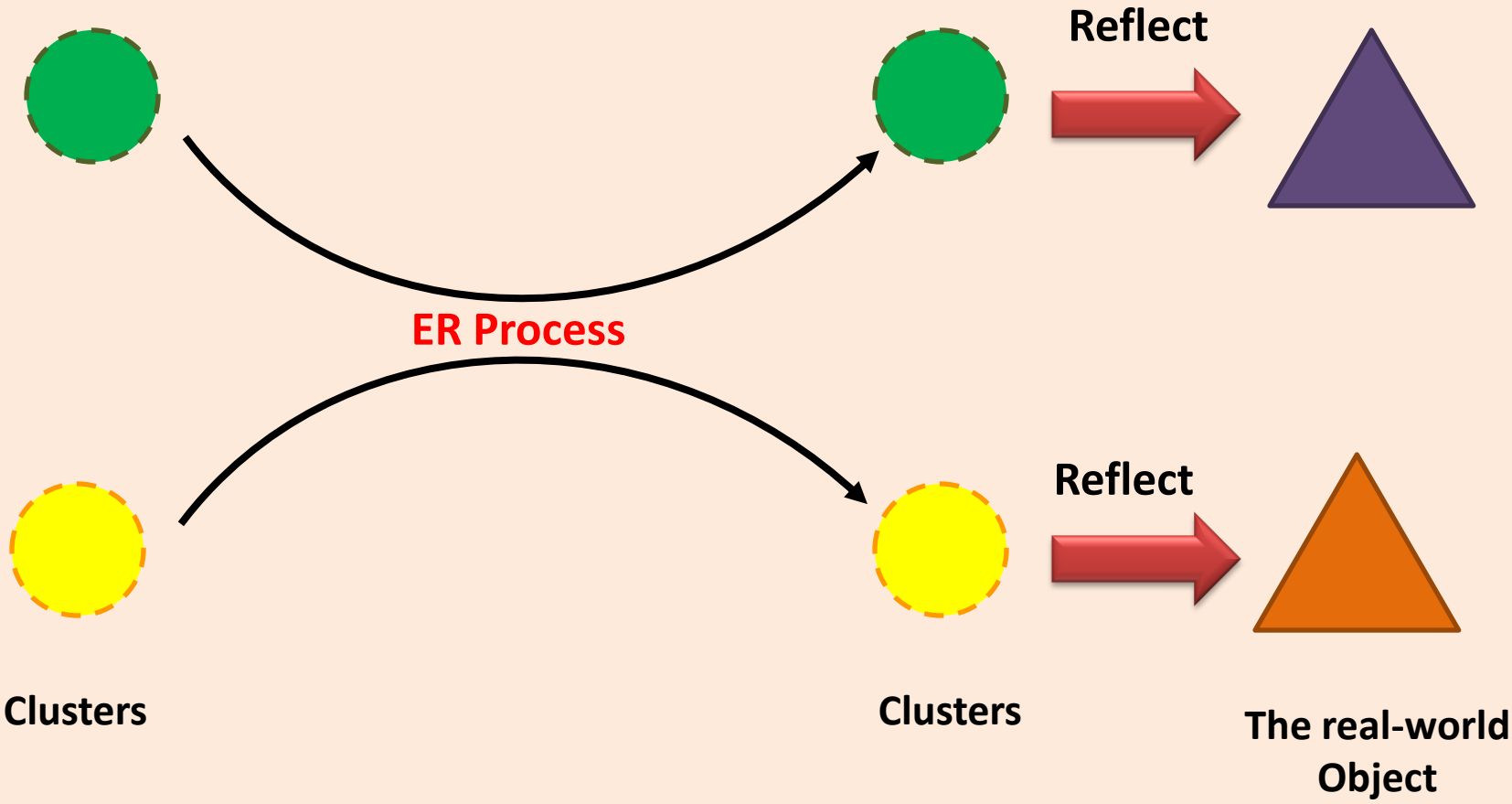
Cluster to Cluster Assertion

Repair False Positive

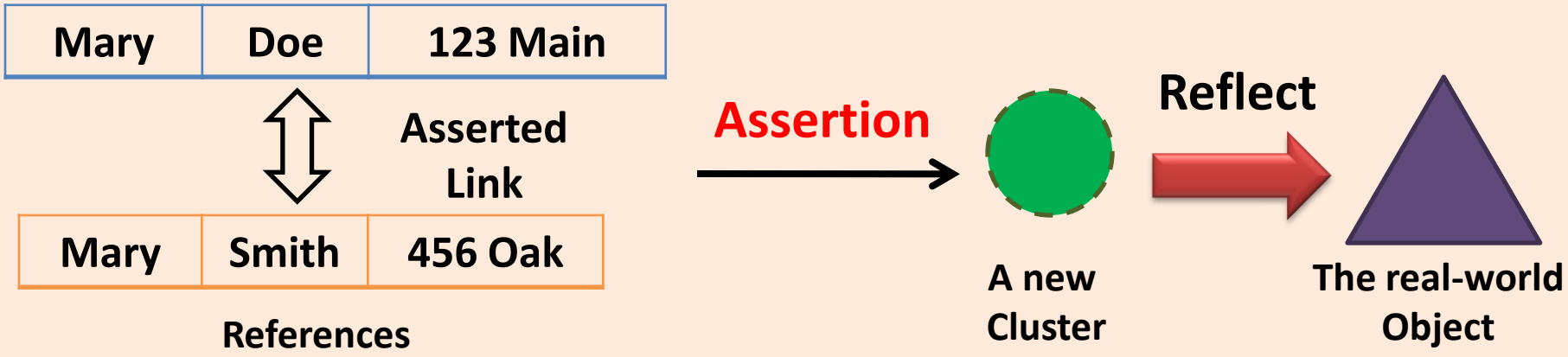


Cluster Split Assertion

Clusters after Split are Negatively Marked

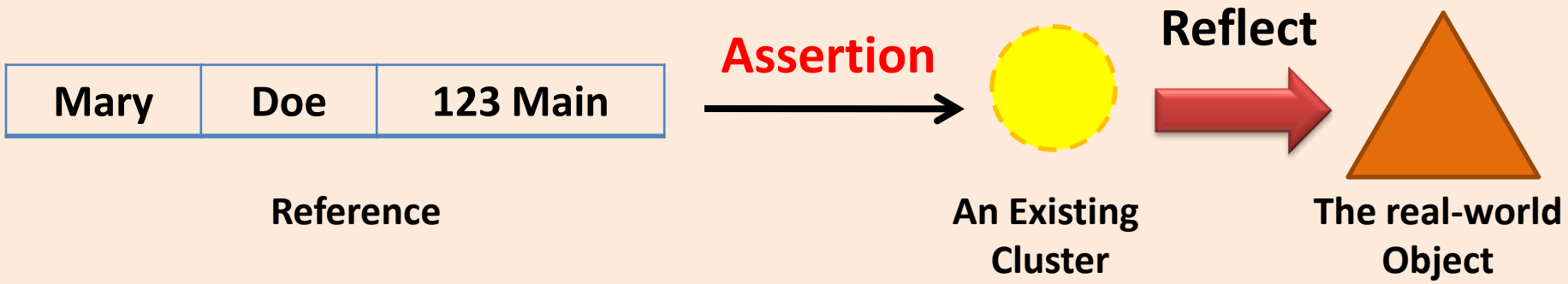


Reference-to-Reference Assertion

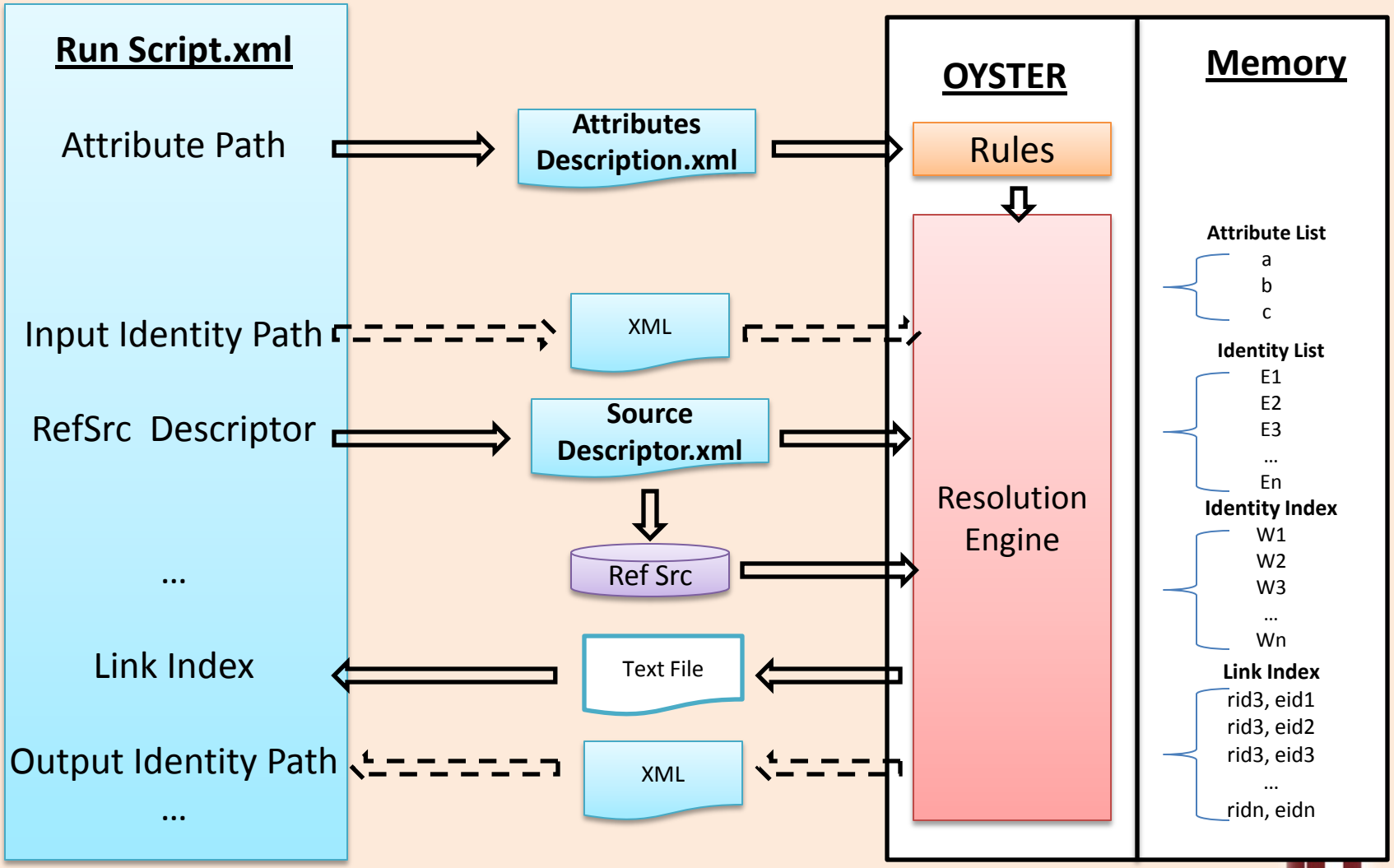


Known to be equivalent because Mary reported her change of name (marriage) and new address to the publisher of her favorite magazine

Reference-to-Cluster Assertion



OYSTER Run Control and Data Flow



Example RunScript.xml

```
<?xml version="1.0" encoding="UTF-8"?>

<!--
   Document   : RunScript.xml
   Created on : 10/21/2011
   Author     :
   Description: Example Run Script
-->

<OysterRunScript>
    <Settings Explanation="On/Off" Debug="On/Off" />

    <LogFile Num="Integer(e.g. 5)" Size="Integer(e.g.100000000)">
        Path
    </LogFile>

    <EREngine Type="FSCluster/RSwooshStandard/RSwooshEnhanced"/>

    <!-- Attributes read from file only -->
   <AttributePath>
        Path
   </AttributePath>
    <!-- Merge-purge does not start with any managed identities -->
    <IdentityInput Type="None/TextFile/Database">
        Path
    </IdentityInput>

    <!-- Merge-purge does not produce any managed identities -->
    <IdentityOutput Type=" None/TextFile/Database">
        Path
    </IdentityOutput>

    <!-- Merge-purge only output is the Link Index  -->
    <LinkOutput Type="TextFile">
        Path
    </LinkOutput>

    <!-- Sources to Run -->
    <ReferenceSources>
      <Source Capture="Yes/No"> Path </Source>
        …
    </ReferenceSources>
</OysterRunScript>
```

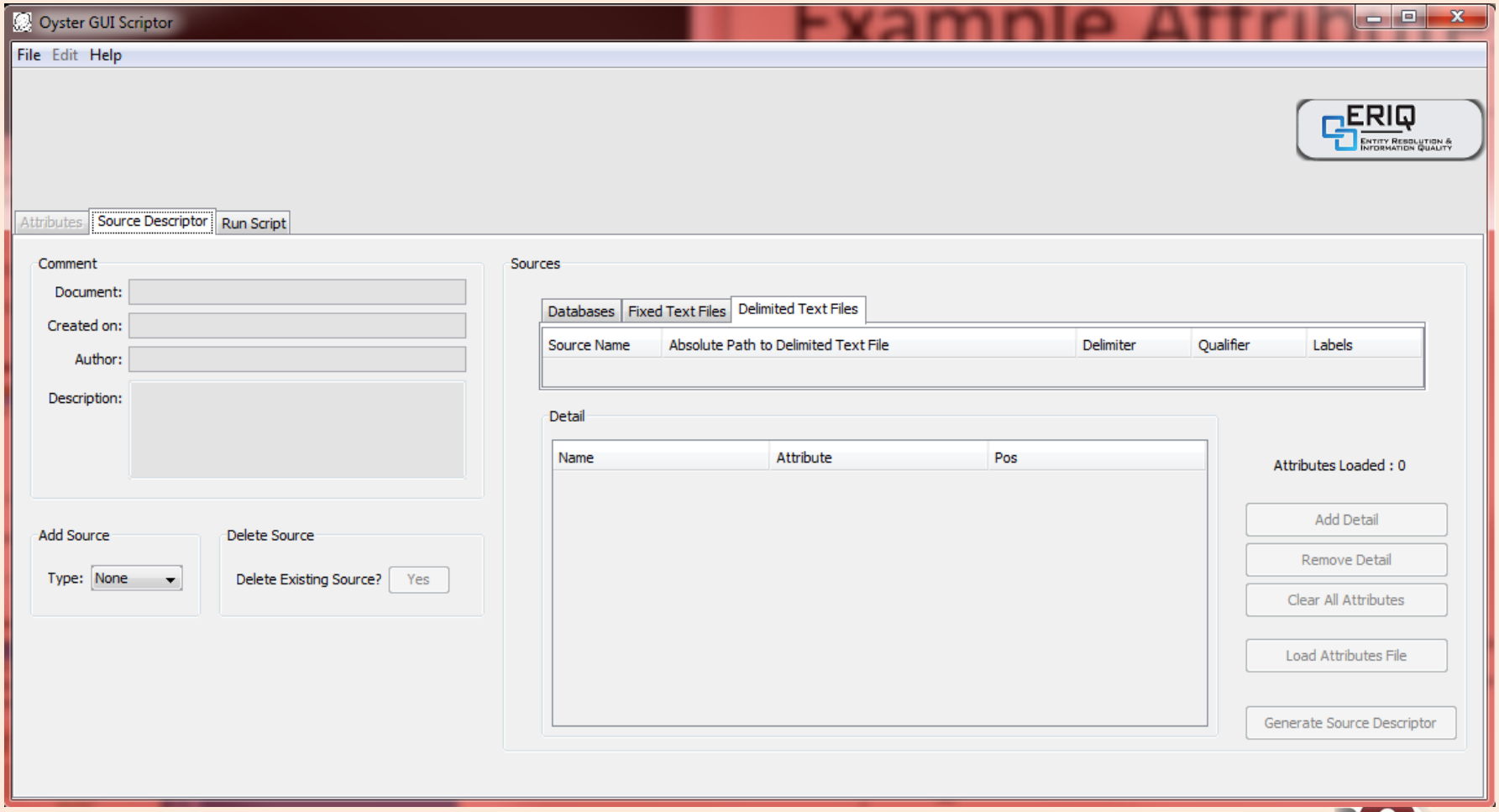
Example SourceDescriptor.xml

```
<?xml version="1.0" encoding="UTF-8"?>

<!--
   Document   : SourceDescriptorq.xml
   Created on : 10/21/2011
   Author     :
   Description: Example Source Descriptor
-->

<OysterSourceDescriptor Name="sourceName">
   <!-- Delimited -->
   <Source Type="FileDelim/FileFixed/Database" … > Path </Source>

   <!-- Items in Source (One for each item in the source including reference identifier --
        >
   <ReferenceItems>
      <!-- For Delimited -->
      <Item Name="FiscalName" Attribute="@RefID" …/>
      <Item Name="FiscalName" Attribute="LogicalName" …/>
</ReferenceItems>
</OysterSourceDescriptor>
```

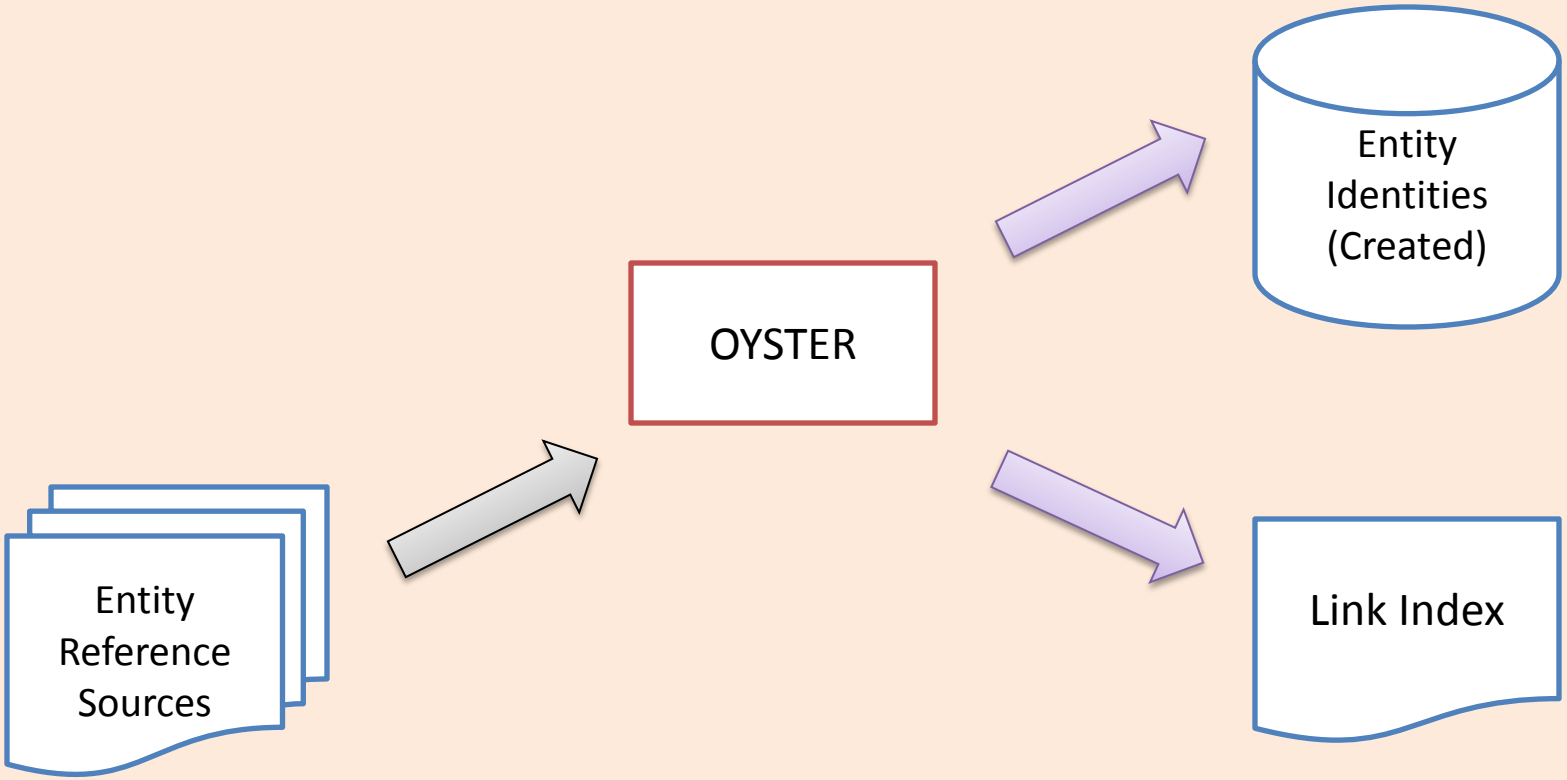
Example Attributes.xml

```
<?xml version="1.0" encoding="UTF-8"?>

<!--
   Document   : MergePurgeAttributes.xml
   Created on : 10/21/2011
   Author     :
   Description:  Example Attribtues
-->

<OysterAttributes System="SystemName">
   <Attribute Item="LogicalName"  Algo= "algoName" />
        …
<IdentityRules>
      <Rule Ident="ruleID">
        <Term Item="LogicalName" MatchResult="code"/>
        <Term Item="LogicalName" MatchResult="code"/>
        …
      </Rule>
        …
</IdentityRules>
</OysterAttributes>
```

OYSTER GUI 1.0



Example Run01 --Identity Capture



Run01 Input

ReferenceID,FirstName,LastName,DateOfBirth,SSN

1,Edgar,Jones,20011010,123-45-9999

2,Eddie,Jones,20011010,123-45-9999

3,Arianna,Johnson,20000505,456-78-6666

Source1.txt

Pre-defined Identity Equivalence Rules(PIERs)

	FirstName	LastName	DOB	SSN
Rule 1	Exact	Exact	Exact	
Rule 2				Exact

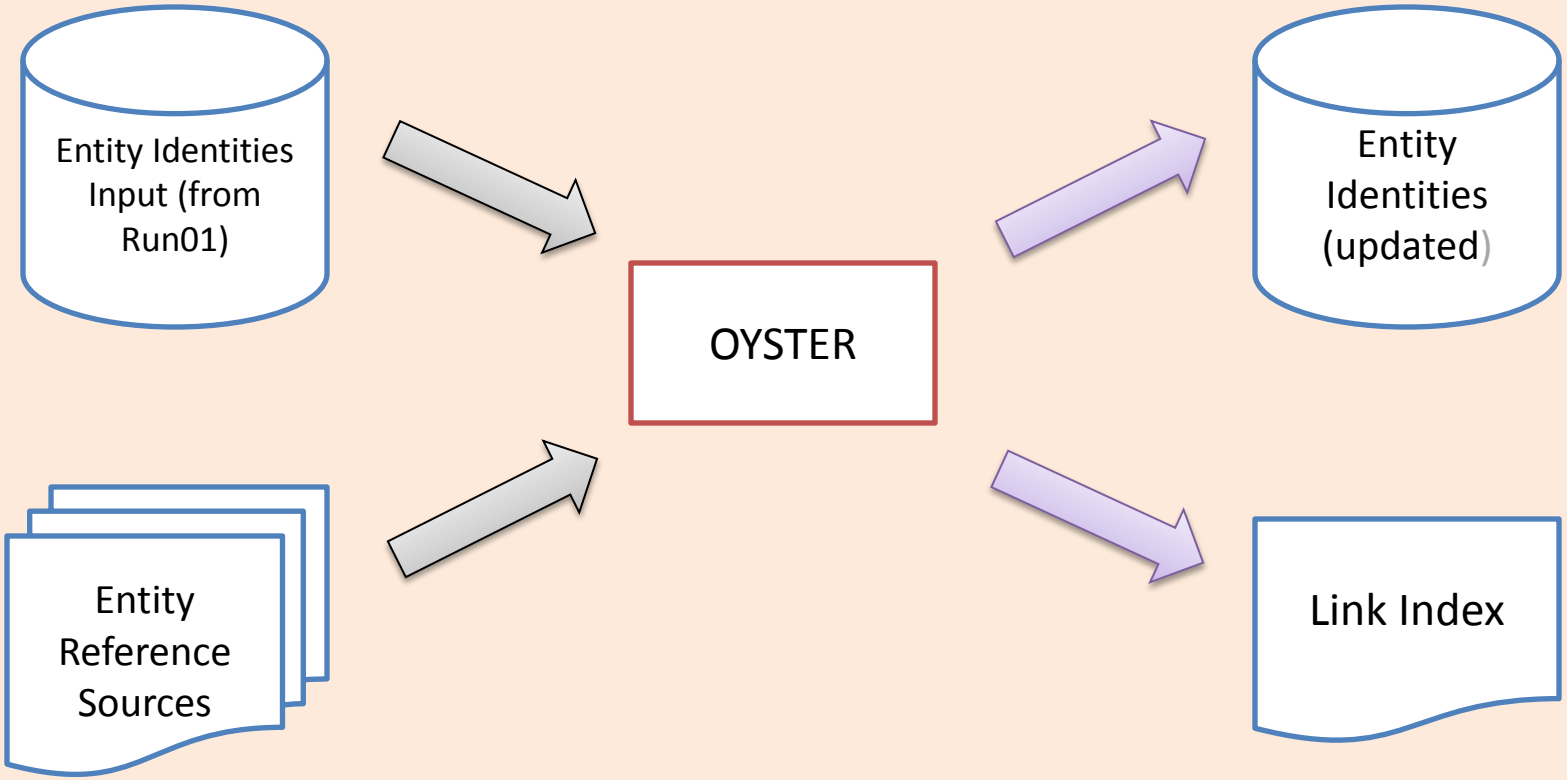
Run01 LinkIndex.link

RefID	OysterID	Rule
Source1.1	9O7DLRY4JIWWEK2W	[2]
Source1.2	9O7DLRY4JIWWEK2W	[2]
Source1.3	ADS5YVNKX198GI21	null

Run01 OutputIdetities.idty

```
...
<Identity Identifier="9O7DLRY4JIWWEK2W" CDate="2011-10-17">
        <References>
                <Reference Value="A^Source1.1|B^123-45-9999|C^Jones|D^20011010|E^Edgar"/>
                <Reference Value="A^Source1.2|B^123-45-9999|C^Jones|D^20011010|E^Eddie"/>
        </References>
    </Identity>
    <Identity Identifier="ADS5YVNKX198GI21" CDate="2011-10-17">
        <References>
                <Reference Value="A^Source1.3|B^456-78-6666|C^Johnson|D^20000505|E^Arianna"/>
        </References>
...
```

Run 02 -- Identity Update



Run02 Input

ReferenceID,FirstName,LastName,DateOfBirth,SSN

1,Edgar,Jones,20011010,132-45-9999

2,Brianna,Johnson,20000505,456-78-6666

Source2.txt

Run02 OutputIdetities.idty

...

<Identity Identifier="9O7DLRY4JIWWEK2W" CDate="2011-10-17">

<References>

<Reference Value="A^Source1.1|B^123-45-9999|C^Jones|D^20011010|E^Edgar"/>

<Reference Value="A^Source1.2|B^123-45-9999|C^Jones|D^20011010|E^Eddie"/>

<Reference Value="A^Source2.1|B^132-45-9999|C^Jones|D^20011010|E^Edgar"/>

</References>

</Identity>

<Identity Identifier="ADS5YVNKX198GI21" CDate="2011-10-17">

<References>

<Reference Value="A^Source1.3|B^456-78-6666|C^Johnson|D^20000505|E^Arianna"/>

<Reference Value="A^Source2.2|B^456-78-6666|C^Johnson|D^20000505|E^Brianna"/>

</References>

...

Run 3—Cluster Split Assertion

RefID,	@RID,	@OID,	@AssertSplitStr
R1,	Source1.3,	ADS5YVNKX198GI21,	1
R2,	Source2.2,	ADS5YVNKX198GI21,	2

AssertionSource.txt

Run 03 OutputIdentities.idty

```
…
<Identity Identifier="ADS5YVNKX198GI21" CDate="2011-10-17">
    <References>
        <NegStrStr>
            <OID> YVQUB43BVDFFY3H2 </OID>
        </NegStrStr>
        <Reference Value="A^Source2.2|B^456-78-6666|C^Johnson|D^20000505|E^Brianna|"/>
    </References>
</Identity>

<Identity Identifier="YVQUB43BVDFFY3H2" CDate="2011-10-17">
    <References>
        <NegStrStr>
            <OID>ADS5YVNKX198GI21</OID>
        </NegStrStr>
        <Reference Value="A^Source1.3|B^456-78-6666|C^Johnson|D^20000505|E^Arianna|"/>
    </References>
</Identity>
…
```

OYSTER Identity Change Report

Date : Oct 17, 2011
RunScript Path: Run03RunScript.xml
RunScript Name: Run03RunScript

Identity Change Summary Section
Count of Output Identities: 3
Count of Input Identities: 2
Count of Input Identities Updated and Written to Output: 1
Count of Input Identities Not Updated and Written to Output: 1
Count of Input Identities Merged: 0
Count of New Identities Created: 1

Identity Change Detail Section
New Identities Created
Identifier References
YVQUB43BVDFFY3H2
…

Thank you!

yxzhou@ualr.edu