**Consortium for:**

**OPEN SOURCE ENTITY RESOLUTION ARCHITECTURE**

**04/30/2014**

**Alan Johnson, Qbase**

**Gregg Hanold, Qbase**

**James Conklin, CCRi**

**Adam Norris, Blue Canopy**

**Cameron Hunt, TRA Holdings**

**Shane Lawson, Ponte Technologies**

**Sponsored by Qbase, LLC**

**12018 Sunrise Valley Drive, Suite 300, Reston, VA 20191**

**Toll Free: 888 458 0345**

**Table of Contents**

[Introduction 3](#_Toc386613036)

[Executive Summary 3](#_Toc386613037)

[Background 5](#_Toc386613038)

[The Challenges 5](#_Toc386613039)

[The Solution 6](#_Toc386613040)

[Collaboration Layer 7](#_Toc386613041)

[Visualization Layer 7](#_Toc386613042)

[Data Access Layer 7](#_Toc386613043)

[Protection Layer 7](#_Toc386613044)

[Common Access Layer 8](#_Toc386613045)

[Enrichment Engine 8](#_Toc386613046)

[Processing Engine 8](#_Toc386613047)

[Analysis Engine 8](#_Toc386613048)

[Data Stores (Relational, Document, Graph) 8](#_Toc386613049)

[Differentiators 8](#_Toc386613050)

[Overall Approach 9](#_Toc386613051)

**Table of Figures**

[Figure 1 -OSERA Operational View 4](#_Toc386613052)

[Figure 2 – Reference Architecture 6](#_Toc386613053)

[Figure 3 - OSERA Consortium Scrum Process and Roles (source: ec\_lds\_ERTProjectCharterFinal.docx) 9](#_Toc386613054)

# Introduction

This white paper presents and Open Source Entity Resolution Architecture and contains concise technical details on all processes, procedures, software, hardware, network connectivity, and expectations of the Open Source Entity Resolution Architecture (OSERA). Specifically the white paper illustrates and defines:

* The integration of multiple federated data sources, both structured and unstructured, with entity extraction and resolution, to provide a fused view into the data and its relationships based on a use case under analysis*.*
* The overarching principals and operational architecture of the OSERA.
* The overarching principals and operational architecture of the Entity Resolution Engine.

# Executive Summary

This white paper defines an Open Source Entity Resolution Architecture (OSERA) that provides for the fusion and analysis of multiple federated data sources that may be structured or unstructured across multiple network domains

A consortium of experts in the fields of Semantic Web, Natural Language Processing, System Security, Big Data Analytics, and Software and Hardware architectures as well as academia has been established that is uniquely positioned to apply methodologies, processes, and proven system security capabilities essential for solving the complex use cases surrounding data and information analysis vital to the commercial sector, our national security, and to the first response elements of our local, state and federal emergency responders. The OSERA defined in this white paperdemonstrates the capabilities of the consortium assembled to; conduct technology assessments to identify effective technologies; develop and integrate processes and technologies; test and evaluate systems; and transfer and integrate processes and technologies across multiple and diverse communities operating on multiple network domains.

The OSERA establishes the operational architecture, processes, and systems that can be reproduced and adapted or subscribed to by commercial, local, state, and federal organizations involved in information sharing and analysis, civil infrastructure protection, emergency management, and Homeland Defense. Further, the OSERA provides for establishing connectivity to such organizations as the private sector Information Sharing and Analysis Centers (ISACs), social networks, Joint Terrorism Task Forces (JTTFs), emergency management portals, law enforcement operations centers, state fusion centers, and military information operations centers. The operational view of the OSERA is depicted in Figure 1.



Figure 1 -OSERA Operational View

The OSERA establishes a reference architecture consisting of:

* Entity Extraction through the Data Access Layer
* Entity Resolution and Disambiguation in the Enrichment Engine
* Multi-domain Storage through configurable common database technologies (relational, document, and graph)
* Semantic and Geospatial Analysis in the Analysis Engine
* Data, corporate and personal information protection across network domains through the Protection Layer
* Enterprise Analysis across multiple network domains through the Common Access Layer
* Common Enterprise Processing of multi-network data sources through the Processing Engine
* Information sharing best practices, standards, and capabilities within a secure multi-domain environment through the Collaboration and Visualization Layers.

This OSERA is implemented through a consortium of industry and academia experts each representing their respective areas of expertise in the reference architecture and subsequent implementation.

# Background

Entity resolution (ER, which is often also referred to as record linkage, or de-duplication), the problem of extracting, matching and resolving entity mentions in structured and unstructured data, is a long-standing challenge in artificial intelligence, statistics, information retrieval, and database management. It has been approached using a variety of techniques, including constraint-based methods, statistical methods, and methods which perform probabilistic inference. The “entities" (be they people, places or things) are referred to in semantically equivalent but linguistically different ways in multiple data systems. The goal of ER is to "resolve" these entities, through automated but guided means, based on the application of multiple knowledge models to the corpus of information being linked, de-duplicated and/or resolved. Accurate and fast entity resolution has huge practical implications in a wide variety of commercial, scientific and security domains. Some examples of the application of ER include:

* Commercial/E-Commerce
* Multiple presentations of an individual consumer; how to resolve, track and market?
* Master data management; products/product lines replace individuals Medical
* In a HIPPA compliant manner, how to bring together a person’s medical history such that it can be used in aggregate knowledge discovery (such as epidemiological assessments)
* Government, Law Enforcement and/or Intelligence
* Multiple presentations of an individual entity of interest; how to resolve, tag and track?
* Historical research, Census records
* Financial, Fraud Detection
* Multiple presentations of an individual entity of interest; how to resolve, tag and track?

This white paper describes an Open Source Entity Resolution Architecture (OSERA) the implements the best of the practical aspects and theoretical approaches to the problem of entity resolution. The OSERA leverages the ER research from AI, machine learning, database and information retrieval communities and the security measures available in SE Linux and Open Solaris. In addition to providing the necessary background information on existing ER models, algorithms and evaluation methods, the white paper also discusses the scalability of the OSERA, the security interfaces, data protection, and ER enhancements necessary for the processing and analysis of community specific use cases.

# The Challenges

Historically ER has faced a number of challenges:

* Name/Attribute ambiguity
* Errors due to data entry
* Missing Values
* Changing Attributes
* Data formatting
* Abbreviations / Data Truncation

The OSERA attempts to address these through the process of entity enrichment, however new challenges have also emerged:

* More Data
  + Need parallel techniques
* More Heterogeneity
  + Unstructured, Unclean and Incomplete data
  + Diverse data types
* More linked
  + Need to infer relationships in addition to “equality”
* Multi-Relational
  + Deal with structure of entities (Are Walmart and Walmart Pharmacy the same?)
* Multi-domain
  + Customizable methods that span across domains
* Multiple applications (web search versus comparison shopping)
  + Serve diverse application with different accuracy requirements

To address the historical challenges and the new emerging challenges, the OSERA initiative has established a consortium of experts from industry and academia.

# The Solution

Open Source Entity Resolution Architecture (OSERA) – A reference architecture that implements, through Open Source Software (OSS), the components necessary to support information dominance and address the current and future challenges identified with entity extraction, resolution, processing and analysis. The OSERA framework supports implementation in hosted (e.g. Amazon, Rackspace …) and private Amazon EC2 compliant data centers. The reference architecture (Figure 2) consists of the following components:

* Collaboration Layer
* Visualization Layer
* Data Access Layer
* Protection Layer
* Common Access Layer
* Enrichment Engine
* Processing Engine
* Analysis Engine
* Data Stores (Relational, Document, Graph)



Figure 2 – Reference Architecture

The OSERA provides the following key capabilities:

* High Performance, Multi-Modal Data Ingestion
* Structured
* Unstructured
* Modalities may include imagery, audio, video and/or biometric signatures
* Cognizant of Personally Identifiable Information (PII) and/or other Legal Data Access Requirements
* Able to obfuscate in an appropriate/approved manner, while retaining pedigree of information
* High Performance Data Processing, with a pluggable architecture
* Open API that supports a rich ecosystem of processing regimes
* Counterpart to the Data Analysis Component
* Data Storage that is
* High Performance/High Volume
* Multi-faceted, flexible, Extensible
* Reliable and Secure (ACID)
* Interoperable
* Open
* High Performance Data Analysis, also with a pluggable architecture
* Open API that supports a rich ecosystem of processing regimes
* Counterpart to the Data Processing Component
* Highly appealing, user friendly Data Visualization and Collaboration environment that is/has
* Extremely compelling user experience
* Open, interoperable and pluggable
* Appealing to a broad range of users (novice to expert)
* Accessible in a variety of manners (UI, Command Line, multiple APIs)

## Infrastructure

The OSERA is built in an Amazon VPC consisting of three Amazon EC2 micro server instances, each with 30G of EBS. The Domain Fusion Infrastructure micro-server is configured with the Collaboration Layer, Visualization Layer and Data Access Layer components. The trusted agents necessary for the multi-domain implementations, require SE Linux extensions applied in the AWS-EC2 Operating Systems. While straightforward, this is not on initial build out of the OSERA. The multi-domain Infrastructure micro-server is configured with the protection layer, common access layer, processing engine, enrichment engine and analysis engine components and services. The data Infrastructure micro-server is configured with the relational, document and graph database components and their respective service.

## Collaboration Layer

There are many open source products that can be leveraged in the implementation of the collaboration layer on the domain micro-server. These include Report and Query Tools, Collaboration Tools, and Analyst Tools. OSERA can support commercial products such as SharePoint through its well defined API which governs the interface requirements for access to the protection layer and the ER capabilities it protects and for integration of the visualization layer components and user/cloud access points and devices.

The OSERA baseline Collaboration layer integrates the following OSS:

* Liferay portal
* Add more

## Visualization Layer

The Visualization Layer integrates the OSS components necessary for the visualization and display through the Collaboration Layer of the Entities resolved through the Analysis engine and reporting tools. These include Visual Link Views, Report Views, and Geospatial Views.

The OSERA baseline Visualization Layer integrates the following OSS:

* Add components here

## Data Access Layer

The Data Access Layer provides the connectors necessary for the Collaboration and Visualization Layers to access the Data Stores and is tightly coupled with the Protection Layer through single domain trusted agents and the OSERA Identity, Access and Authorization interfaces. In addition, the Data Access Layer provides the enrichment, processing and analysis engines access to the domain cloud’s data sources through secure single domain trusted agents. The Data Access Layer ensure the pedigree of the data is preserved and access identity is maintained.

The OSERA baseline Data Access Layer integrates the following OSS:

* OrientDB JDBC connector
* Add additional components here

## Protection Layer

The Protection Layer provides the Single domain access for the Domain Fusion Infrastructure including the Domain Trusted Agents to the Mult-Domain Infrastructure components (Enrichment Engine, Processing Engine, and Analysis Engine) through the defined Controlled Interface API and Common Access Layer. The Protection Layer identifies the domain of access, and identity, access, and authorization of the user and preserves them across the Common Access Layer.

The OSERA baseline Protection Layer integrates the following OSS:

* SE Linux extensions
* Add additional components here

## Common Access Layer

The Common Access Layer is tightly coupled to the protection layer to give authenticated and authorized access to the global multi-domain components provide enrichment, processing and analysis of entities stored within the data layer. The Common Access Layer also provides the interfaces to these components and persists the identities as they move from a single domain to a multi-domain infrastructure.

The OSERA baseline Protection Layer integrates the following OSS:

* SE Linux extensions
* Add additional components here

## Enrichment Engine

Need input on a good discussion of this Layer and supporting OSS

## Processing Engine

Need input on a good discussion of this Layer and supporting OSS

## Analysis Engine

Need input on a good discussion of this Layer and supporting OSS

## Data Store (Relational, Document, Graph) Layer

The Data Store Layer hosts the database components. The OSERA baseline will include OSS relational, document and graph databases.

The OSERA Data Store Layer integrate the following OSS:

* MySQL (please comment on the best OSS choice for a relational DB)
* monDB (please comment on the best OSS choice for a document DB)
* orientDB (please comment on the best OSS choice for a graph DB)

# Differentiators

Why OSERA?

OSERA provides and open source, scalable, extensible and pluggable secure multi-domain reference architecture. The OSERA consortium has introduced the following concepts with this implementation that serve as differentiators from commercial alternatives:

* Open Source
* Data Loading

Most current Extraction Transformation and Load (ETL) tools focus on Relational Targets/SQL based storage. The OSERA consortium recognized that ETL for “NoSQL” is an emerging market with great potential and has introduced this capability in its data access and processing layers.

* Data Storage

Most NoSQL/Graph DB have been constructed with little/no regard to information protection concerns. The OSERA adds multi-domain security and data and information protection capabilities in its protection common access and data access layers.

* Data Processing/Analysis

The OSERA implements in its processing and enrichment layers architectures such as UIMA and/or GATE that operate at a lower level than is appropriate for ER. By including these architectures, OSERA provides data processing and analysis capabilities that are integrated with more traditional ER solutions.

* Model Driven Mechanism

Most semantic analysis is “over the (collective) heads” of typical domain analysts. The OSERA open source framework and API allows for the introduction of tools (open source, internal or commercial) that implement sematic analysis in a “user friendly manner”.

# Overall Approach

The Consortium’s overall technical is divided into two broad categories. The first is the installation, integration and configuration of Open Source products in an Amazon EC2 compliant environment that form the reference ERA upon which community specific requirements can be added. The second is the development of both hosted and deployable instances of the ERA with a defined API to support integration of community specific components defined by their unique requirements. The Consortium has adopted a variant of the agile methodology as depicted in Figure 3 used in the Wisconsin Department of Public Instruction’s Project P377: “Entity Resolution Tool (ERT) Selection and Implementation For the Early Childhood Longitudinal Data System (EC LDS)”(June 5, 2013).

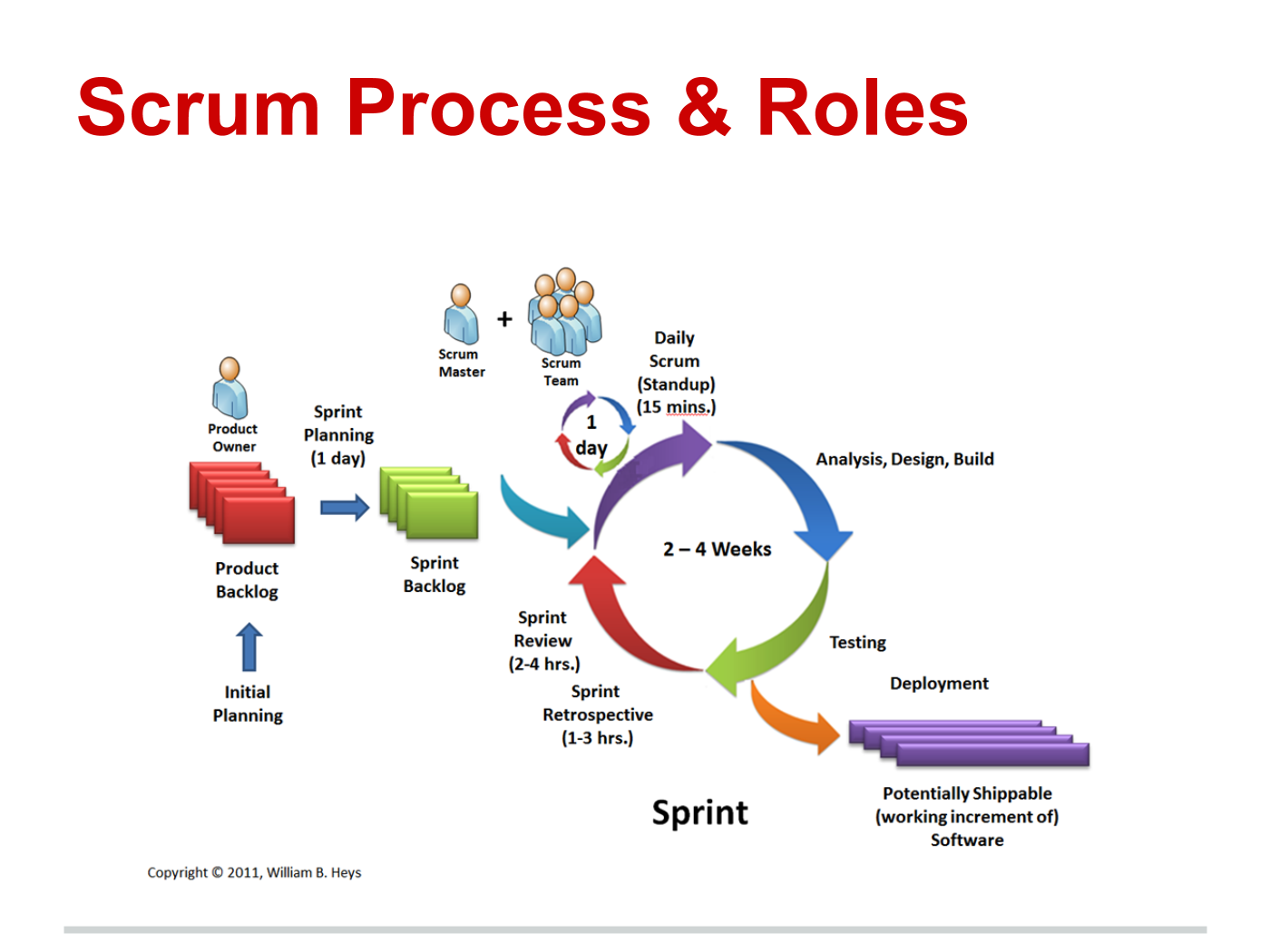


Figure 3 - OSERA Consortium Scrum Process and Roles (source: ec\_lds\_ERTProjectCharterFinal.docx)