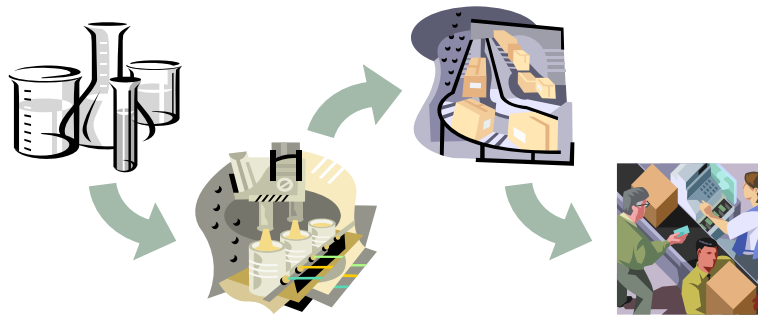




## 2020: A New World

- The ultimate goal has been met: to provide a standard means of describing product data through the life cycle of a product – a shared resource for all



55

## Smart Step Codification Phase 3

AC/135 have commissioned a Phase III of the SSC project

Phase I – Proved that STEP files could be used to generate codification records.

Phase II – Used SSC and ISO's 22745 & 8000 to create 100 Item of Supply Concepts for ROSOMAK.

Phase III – Will look to continue this work and develop true IT based automated data exchanges between Defence and Industry. A detailed Cost Benefits Analysis will also be produced.



MINISTRY OF DEFENCE



JOINT SUPPLY CHAIN

## The Task

To take a medium sized platform with mature enough data to be codified which is stored in an electronic Product Data Management (PDM System).

Using ISO 8000 exchange methods, create a fully codified platform direct from the PDM.

Return a copy of that data to the supplier in ISO 22745 format including the NSN as a completed field.

The successful completion of the project will result in demonstrable improvements in quality and time in the completion of a codification task and provide information on potential whole life cost savings



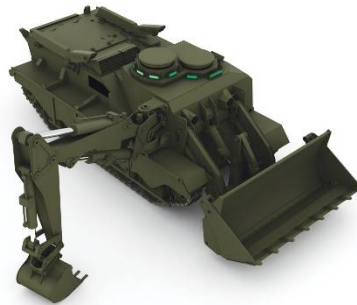
## TERRIER is a new generation Combat Engineering Vehicle (CEV)

- Used for Early entry
- Used for Combat support
- Used for Post conflict roles



## TERRIER® Capability

- **TERRIER Uses next generation Drive by Wire electronics**
- **Key points from TERRIER specification:**
  - 2 man crew
  - 31.5 tonnes
  - 700hp engine
  - 70kph top speed
  - 5 tonne clamshell bucket



## TERRIER® Capability

- **Key points from TERRIER specification:**
  - 2.5 tonne excavator arm
  - Thermal Imaging and low light cameras
  - Capable of being remotely controlled
  - 10 tonne integrated winch system
  - General Purpose Machine Gun
  - Scatterable Mine Clearance Device



## Counter Mobility / Survivability

High capacity bucket – 2.8 m<sup>3</sup>  
Excavator arm – 0.4 m<sup>3</sup>



Infantry and vehicle protective positions  
Deployed Force Infrastructure  
Host Nation Infrastructure



## The tale of the tape

**BETTER** - Current NATO Average for the creation of Type 1 records is approximately 16%.

Smart Step Codification Type 1 Creation = 60%

**FASTER** – UK NCB Average for the allocation of an NSN on receipt of the Source Data = approximately 50 minutes.

Smart Step Codification = 10 Minutes



So what does that mean in financial terms to the supplier?

389 Items for codification so far

129 Items screened out which is 33%

BAES will put forward approximately 2000 items for Terrier by project end.

That is a cost of approximately £44,000 in hard charging for codification

33% of £44,000 is **£14,520** which would be the estimated savings on codification costs.

BAES Don't have a classification system



So what does that mean in financial terms to the supplier?

**IF** a supplier was to place codification at the **design stage** and be able to accept the automated import of an R-XML File:

TERRIER had 129 Items Screened out as already existing in ISIS which UK NCB produced R-XML files which BAES GCS imported into the ISO 22745 Module they had access to.

It costs BAES GCS £3000 to introduce an item in to their catalogue

In accordance with the Shell UK commissioned survey 50% of those costs are for data.

129 x £1500 = **£193,500.00**



So what does that mean in financial terms to the supplier?

The potential to BAES GCS is far greater than that as UK NCB can provide data in r-XML format for 19,000 items that can be automatically loaded into any classification system they choose with XML capabilities. This data will be in ISO 22745 format and in accordance with ISO 8000 Pt 110.

If, we can get codification introduced at the design of a platform, before the engineers start to create properties and values:

The potential is there to save hundreds of thousands of pounds



## The Biggest Challenge

### **BAES GCS Has no classification system!**

This means that at present they have no supporting data electronically that can be used for codification.

For this project it means a work around by giving BAES GCS access to the suppliers modules available from both ESG and AURA.

For BAES it shows why they would be so interested in taking part in this project.



## The Cost of not codifying!

James Beer is the project manager at BAES GCS responsible for the introduction of a classification system, why?

He provided the following figures:

Cost to introduce an item into their Product Data Management Tool: **£3000**.

Average number of duplicates per item found in their PDM Tool: **10**

Each item has an un-necessary support cost of on average: **£27,000**

BAES GCS Newcastle has approximately **19,000** items registered against its NCAGE currently.



## Benefits & Barriers

### **Benefits already apparent**

The Data the supplier has access to is far greater than what is traditionally sent to NCBs.

The Supplier is in a better position to make judgement calls on the item.

### **Barriers still in place**

It was worrying that the supplier did not have a readily identifiable and accessible repository for their data.

The willingness of commercial companies like BAES to allow 'plug in software' into their systems is very limited.







## Implementation of ISO 22745/8000

- Many companies are now in the business of building ISO 22745/8000 compliant catalogs. Some examples:
  - PiLog – South Africa
  - Quadrem
  - ESG
  - AURA

69



## Implementation of ISO 22745/8000

- Many organizations are have implemented ISO 22745/8000 compliant catalogs, are testing them, or having committed to adopting them:
  - ArcelorMittal
  - PHP Billiton
  - Severstal
  - Aramco
  - Anglo-American Inc.

70



## Implementation of ISO 22745/8000

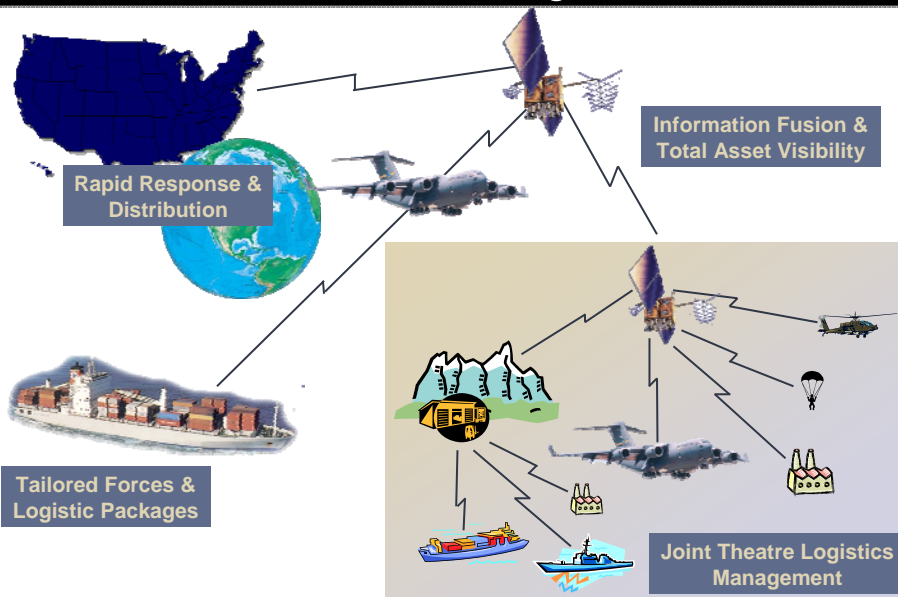
Many nations within the AC/135 community are running or planning to run pilot projects to test electronic data exchange between suppliers and government offices using 22745/8000, including Belgium, Czech Republic, Finland, New Zealand, Norway, Poland, Russia, Slovakia, United Kingdom, and the United States



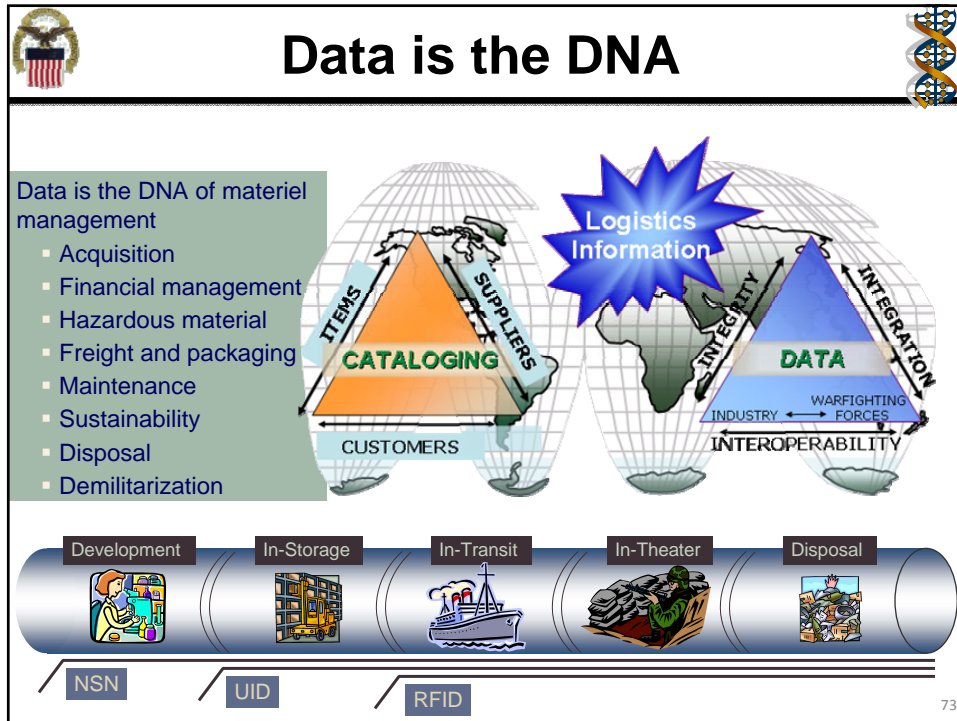
71



## Netcentric Logistics



72



## Summary

- The **NATO Codification System** is an international standard for exchange of catalog data in government
- **ISO 22745** is an e-catalog standard based on the NCS and **ISO 8000** ensures the quality of the data
- ISO 22745 and 8000 are working in practice and poised for wide implementation around the world

**Investment in ISO 22745 and 8000 =  
Strong Return on Investment**

74



## Useful International Web Site Addresses

- NATO CODIFICATION SYSTEM (NATO ALLIED COMMITTEE 135)
  - <http://www.nato.int/structur/AC/135/welcome.htm>
- NATO MAINTENANCE AND SUPPLY AGENCY (NAMSA)
  - [http://www.namsa.nato.int/home/www.namsa\\_e.htm](http://www.namsa.nato.int/home/www.namsa_e.htm)
- NATO MCRL
  - [http://www.nato.int/structur/AC/135/nmcrl/nmcrl\\_e/index.htm](http://www.nato.int/structur/AC/135/nmcrl/nmcrl_e/index.htm)
- NATO AMMUNITION DATA BASE (NADB)
  - [http://www.namsa.nato.int/ammo/nadb\\_e.htm](http://www.namsa.nato.int/ammo/nadb_e.htm)
- NATO HEADQUARTERS
  - <http://www.nato.int>
- PACIFIC AREA CATALOGING SYSTEM (PACS)
  - [http://www.defence.gov.au/dmo/\\_jlc/pacs](http://www.defence.gov.au/dmo/_jlc/pacs)

## Affordable Data Management

### ABSTRACT

---

\* ABSTRACT NOT AVAILABLE \*

### BIOGRAPHY

---

#### **Sebastiao Correia**

R&D Engineer  
Talend, France

Sebastiao is currently the team leader of the Data Quality products developed at Talend. He received a Ph.D in Theoretical Physics in 2000 and left the academic world in 2001 in order to tackle optimization problems at Chronopost, a parcel shipping company. There he used a mixture of genetic algorithms and renormalization ideas coming from his physics background in order to provide real solutions to the vehicle routing problem with time windows. After tackling UML and MDA (Model Driven Architecture), he participated in the development of the graphical software that provides routing information to the Chronopost drivers. Since then, he used models to develop several products with Eclipse RCP and EMF ranging from an optimization tool for a to an interactive task planner using a rule engine. His experience in modeling led him to work on the MDM (Master Data Management) project at Geopost and to be responsible of the existing master data. His interest in statistical analysis also led him to work on several Business Intelligence projects before joining Talend. His skills in algorithmic development, modelisation and statistical analysis helps him in the current development of Talend's data quality softwares.

#### **Steve Sarsfield**

Product Manager  
Talend



**talend\***  
open data solutions

## Affordable Data Management

July 14-16, 2010  
Sebastiao Correia  
Steve Sarsfield

MIT 2010  
Information Quality Industry Symposium



**talend\***  
open data solutions

## Speakers

- \* **Sebastiao Correia (Ph.D)**
  - Data Quality Team Leader
  - Blogger
    - [scoreiait.wordpress.com](http://scoreiait.wordpress.com)
- \* **Steve Sarsfield**
  - Product Manager
  - Blogger
    - [data-governance.blogspot.com](http://data-governance.blogspot.com)
  - Author of "The Data Governance Imperative"

© Talend 2010



## Agenda

- \* **Where are the Costs in Data Management?**
- \* **Different Methods of Managing Data**
- \* **Changing Landscape**
- \* **Low-Cost Resources**
  - **Standards**
  - **Tools**
  - **Reference Data**
  - **Regular Expressions**
- \* **Questions and Answers**

3



## Unifying Data Management

- \* **business processes**
  - Ensures that important data assets are formally managed throughout the enterprise
  - Instills trust in data can be trusted
  - Assigns accountability
- \* **an evolutionary process for a company**
  - Information-centric thinking
  - Empower people, setting up processes and getting help from technology


4



## What's Costly About Data Management?

- \* **People**
  - Internal Resources
  - Consultants
- \* **Processes**
  - Time and resources to set up new ones
    - Getting buy-in
    - Process Change
    - Follow-up
- \* **Technology**
  - Data Management Technology
    - Profiling, Data Quality, Master Data Management

5



## What Does it Cost?

- \* **All-encompassing Approach (MDM)**
  - \$1.2 Million Software
  - \$4.0 Million Services
  - Total - \$5.2 Million

(source: CDI Institute Survey – in 2007 among Global 5000 companies)
- \* **Land and Expand Approach**
  - Data Management can take the shape of a series of coordinated projects
  - Projects >Business Units>Regions>Companies
- \* **Price Pressure from:**
  - Open Source/public domain/transparency
  - Internet

6





## Benefits of 'Land and Expand'

- \* **Lower up-front costs with smaller scope**
- \* **Evolution, not boiling the ocean**
- \* **Tack on data management to CRM, ERP, etc.**
- \* **Easier approval**
- \* **Ability to cherry-pick projects**
  - High ROI
  - Easy to do

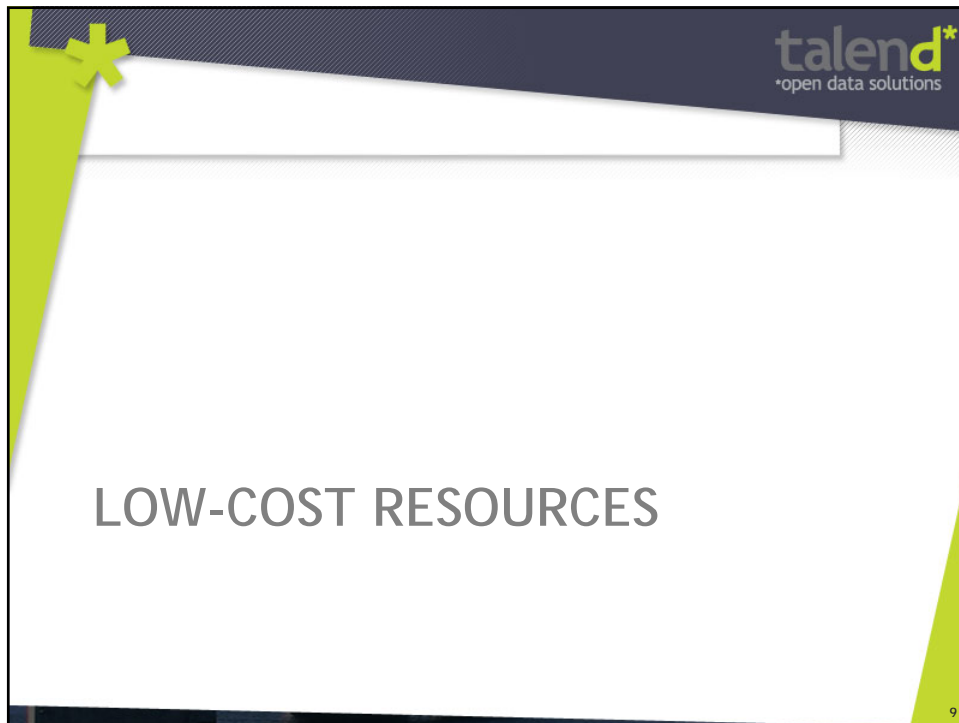
7



## Keys to Land and Expand Proposals

- \* **Find the ROI**
  - Revenue, Efficiencies, Compliance
- \* **Limit Scope and Hold**
- \* **Give clear metrics on success and failure**
- \* **Talk about project in terms of business benefits**
- \* **Market Your Team**
  - Be ready with an elevator pitch
  - Newsletters/Social Media/E-mails
- \* **Include a 'do-nothing' option**

13



**talend\***  
open data solutions

## Standards

- \* **Standards in data quality**
  - ISO codes for countries, currencies, languages...
  - Phone prefixes
  
- \* **Metadata standards**
  - Postal address format
  - ISO 8000 data quality standard (copyrighted)
  
- \* **Metadata and data quality metamodels**
  - What is a data?
  - How data quality is measured?

11

**talend\***  
open data solutions

## CWM (Common Warehouse Metamodel)

- \* **UML Model provided by the OMG (Object Management Group)**
- \* **Purpose of CWM: interchange metadata**

**The CWM Metamodel**

	Warehouse Process			Warehouse Operation		
Management						
Analysis	Transformation		OLAP	Data Mining	Information Visualization	Business Nomenclature
Resource	Object Model	Relational	Record	Multidimensional		XML
Foundation	Business Information	Data Types	Expression	Keys and Indexes	Type Mapping	Software Deployment
	Object Model					

© Talend 2010

**talend\***  
open data solutions

## Data Quality model

- \* **No formal specification of data quality by the OMG**
- \* **There exist a few tentatives to model data quality**
  - A Data Quality Metamodel Extension to CWM (P. Gomes *et al*)


© Talend 2010

**talend\***  
open data solutions

## Talend Open DQ model

- \* **Uses CWM**
- \* **Uses Gomes DQ metamodel**
- \* **Enhances both models**
  - Defines an analysis composed by indicators
  - Each indicator applies on CWM elements (columns...)
  - Each indicator defines a DQ domain (regular expression, thresholds...)
  - Each indicator stores a measure
  - ...


© Talend 2010



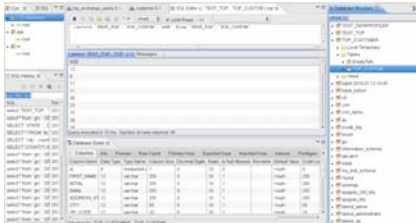

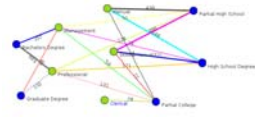
## Tools - You have Options

- \* **Profiling**
  - SQL Queries
  - Open Source
  - Vendors
- \* **Data Integration & Master Data Management**
  - Home-made code
  - Open Source
  - Vendors
- \* **Reference Data**
  - Pay for Use
  - Public Domain

15



## Examples of open source applications

- \* **SQL Explorer:**
  - database exploration tool
- \* **Jasper report: reporting in pdf, html, excel...**
- \* **Jung CERN network chart library**


© Talend 2010



## Available data matching libraries

- \* **Approximate string matching**
  - Second String <http://secondstring.sourceforge.net/>
    - W. W. Cohen from
  - Apache Jakarta Commons Codec (soundex)
  - Apache Commons Lang (Levenshtein, Metaphone)
- \* **Entity resolution**
  - SERF (Stanford Entity Resolution Framework)
  - Sun's Mural MDM solution <https://mural.dev.java.net/>
    - Probabilistic matching based on Fellegi-Sunter theory.
  - **FRIL**. Fine grained Records Integration and Linkage tool.


© Talend 2010



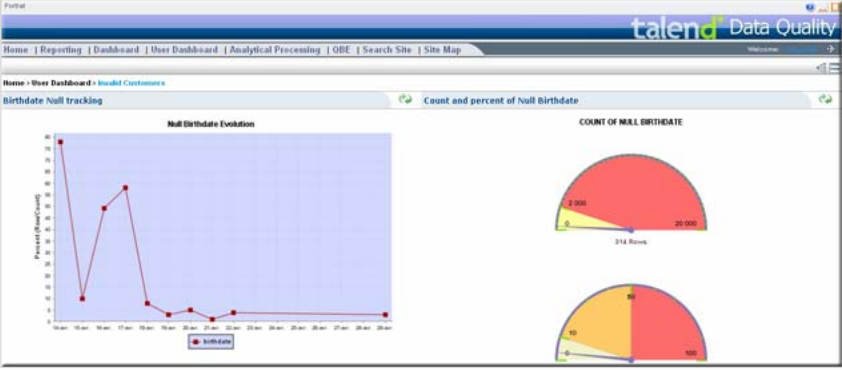
## Existing string matching implementations

Algorithm	SimMetrics	SecondString	Mural
Hamming distance	n	n	n
Levenshtein distance	Y	Y	n
Needleman-Wunch distance or Sellers Algorithm	Y	Y	n
Smith-Waterman distance	Y	Y	n
Gotoh Distance or Smith-Waterman-Gotoh distance	Y	n	n
Block distance or L1 distance or City block distance	Y	n	n
Monge Elkan distance	Y	Y	n
Jaro distance metric	Y	Y	Y
Jaro Winkler	Y	Y	Y
SoundEX	Y	n	n
Matching Coefficient	Y	n	n
Dice's Coefficient	Y	n	n
Jaccard Similarity or Jaccard Coefficient or Tanimoto coefficient	Y	Y	n
Overlap Coefficient	Y	n	n
Euclidean distance or L2 distance	Y	n	n
Cosine similarity	Y	n	n
Variational distance	n	n	n
Hellinger distance or Bhattacharyya distance	n	n	n
Information Radius (Jensen-Shannon divergence)	n	Y	n
Harmonic Mean	n	n	n
Skew divergence	n	n	n
Confusion Probability	n	n	n
Tau	n	n	n
Fellegi and Sunters (SFS) metric	n	Y	n
TIDF or TF/IDF	n	Y	n
FastA	n	n	n
BlastP	n	n	n
Maximal matches	n	n	n
q-gram	Y	n	Y
Ukkonen Algorithms	n	n	n
Metaphone	n	n	n
Double Metaphone	n	n	n

© Talend 2010




## Reporting Data Quality Metrics




The screenshot shows a dashboard for 'Invalid Customers' with a 'Birthdate Null tracking' section. It includes a line chart titled 'Null Birthdate Evolution' showing the percentage of null birthdates over time from 16-Jan to 28-Apr. The percentage starts at approximately 25% on 16-Jan, drops to about 5% by 17-Mar, and remains low through 28-Apr. To the right, there are two semi-circular gauges. The top gauge is labeled 'COUNT OF NULL BIRTHDATE' and shows 20,000 nulls and 314 rows. The bottom gauge shows 100 nulls and 100 rows.

What relevance does birth date have on business processes?  
How much has the team saved over time?



## Unified Data Management - Tools



The diagram illustrates the tools for Unified Data Management. It lists ten tools: Connectivity, Data Assessment, Metadata Management, Transformation, Cleansing, Standardization, Matching, Monitoring, Stewardship, and Active Data Model. These tools are organized into three main categories represented by arrows pointing right: Data Integration (covering Connectivity, Data Assessment, and Metadata Management), Data Quality (covering Transformation, Cleansing, Standardization, and Matching), and Master Data Management (covering Monitoring, Stewardship, and Active Data Model).

© Talend 2010



## The success of open source tools

- \* **A low price**
- \* **Transparency**
  - show how the tool works (Allow the user to see and modify the java/sql code)
- \* **Leverage on well-known standards**
  - Java, SQL, regular expressions...
- \* **Often permits customization**
  - add regex, indicators, components...

© Talend 2010



## Reference data

- \* **Open data resources is getting structured**
  - Open source like initiatives
  - Gouvernment initiatives
- \* **8 Principles of Open Data**
  - Complete: All public data is made available.
  - Primary: Data is as collected at the source
  - Timely: Data is made available as quickly as necessary
  - Accessible: Data is available to the widest range of users
  - Machine processable: Data is reasonably structured
  - Non-discriminatory: Data is available to anyone
  - Non-proprietary: open data format
  - License-free: Data is not subject to any copyright, patent, trademark or trade secret regulation.


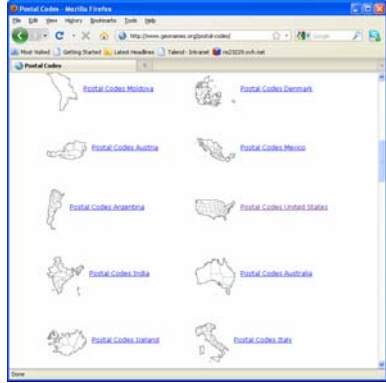
© Talend 2010



**talend\***  
open data solutions

## Example of open initiative

- \* **Geographical database: Geonames**
  - Countries, largest cities, highest mountains, capitals, postal codes
  - 8 million geographical names
- \* **Open Municipal Geodata Standard**
  - data at the city and municipal agency levels

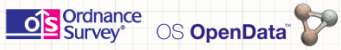



© Talend 2010

**talend\***  
open data solutions

## Government Sources

- \* **data.gov**
  - The purpose of Data.gov is to increase public access to high value, machine readable datasets
- \* **UK open public data**
  - Geographical data
  - Population data
  - Transport data
  - Education & skills data
  - ...
- \* **French initiative**



© Talend 2010



## Benefits for governments

- \* **Reduce time, effort and resources in fulfilling public information requests**
- \* **Increase data quality by providing correct data to public from the source**
- \* **Reduce duplication of effort**
- \* **Increase data access, availability, and speed of delivery**
- \* **Improve citizen satisfaction and create good public relations with your community**


25



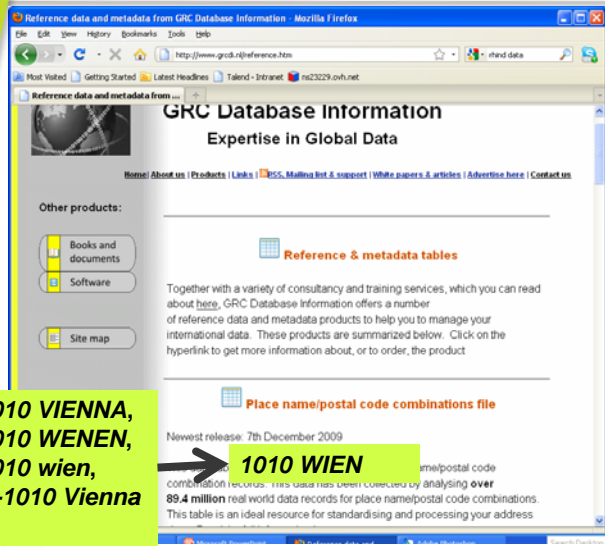
## Benefits for citizens

- \* **Open access to complete, formatted data rather than relying on third party interpretations or subsets**
- \* **Information accessibility leads to greater government accountability**
- \* **Fosters better community action on social issues, eg. crime, pollution, permits, accidents, and education**
- \* **Improves regional competitiveness by giving businesses quicker and fuller access to data**

26




## GRC Database Information



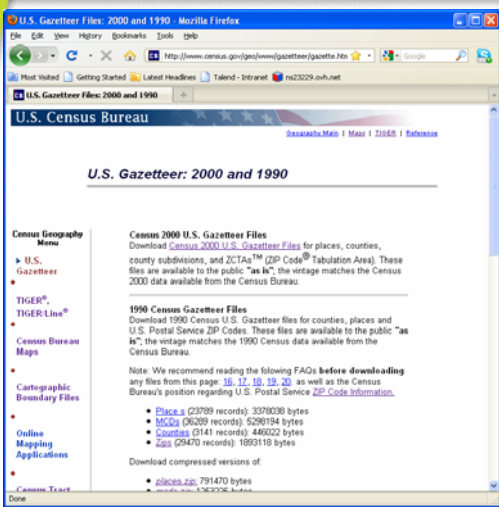
**1010 VIENNA,  
1010 WENEN,  
1010 wien,  
A-1010 Vienna** → **1010 WIEN**

- \* **Common misspellings of:**
  - Name
  - Address
  - Job Title
  - More...
- \* **Gathered by examining real-life data and developing tables.**

27



## Enrichment



- \* Field 1 - State Fips Code
- \* Field 2 - 5-digit Zipcode
- \* Field 3 - State Abbreviation
- \* Field 4 - Zipcode Name
- \* Field 5 - Longitude in Decimal Degrees
- \* Field 6 - Latitude in Decimal Degrees
- \* Field 7 - 2000 Population
- \* Field 8 - Allocation Factor (decimal portion of state within zipcode)

**Dedham, MA**  
 "25", "02026", "MA", "DEDHAM", 71.163741, 42.243685, 23782, 0.003953

**talend\***  
open data solutions

## Mapping Data

29

**talend\***  
open data solutions

## Regular Expressions

- ✳ **Great for verifying shape and structure of data**
  - Example: emails
- ✳ **Used in most databases**
- ✳ **Used in most programming languages**
- ✳ **May help to find invalid data**
  - Example of random text

Data Type	Matching Percentage
Email Address	99.21%
Name from text	96.27%
Random sequence of text	99.71%

Use of regex

Random emails

email

mkslfq@qdmfj.sdmkif

afvfmgtc

klmjfdgs@tmlskjg.fdsj

Data Type	Matching Percentage
Email Address	99.04%

© Talend 2010

**talend\***  
open data solutions

## Where to find Regular Expressions?

**Regular-Expressions.info**

Tutorial Tools & Languages Examples Books & References

### More Detailed Examples

- [Numeric Ranges](#). Since regular expressions work with...
- [Matching a Floating Point Number](#). Also illustrates th...
- [Matching an Email Address](#). There's a lot of contro...
- [Matching Valid Dates](#). A regular expression that mat...
- [Finding or Verifying Credit Card Numbers](#). Validate c...
- [Matching Complete Lines](#). Shows how to match com...
- [Removing Duplicate Lines or Items](#). Illustrates simp...
- [Regex Examples for Processing Source Code](#). How to...
- [Two Words Near Each Other](#). Shows how to use a re...

**RegExLib.com**  
Regular Expression Library

Home Search Regex Tester Browse Expressions Add Regex Login

Subscribe  
Recent Expressions

Site Links  
Regex Cheat Sheet  
Search  
Regex Tester  
Browse Expressions  
Add Regex  
Manage My Expressions  
Contributors  
Regex Resources  
Web Services  
Advertise  
Contact Us  
Register  
Recent Expressions  
Recent Comments

Browse Expressions by Category  
Email Url Numbers Strings Date and Time Misc Address/Phone Markup/Code  
28 regular expressions found in this category!

Expressions in Category: Email  
Change page 1 of 2 pages: Items 1 to 20

Title	Expression
email address (RFC 2822 mailbox)	<code>*{([a-z0-9!#\$%&amp;'*+,-./:;=?@^_`{ }~]+@) ([a-z0-9!#\$%&amp;'*+,-./:;=?@^_`{ }~]+@[a-z0-9]+) ([a-z0-9!#\$%&amp;'*+,-./:;=?@^_`{ }~]+@[a-z0-9]+@[a-z0-9]+) ([a-z0-9!#\$%&amp;'*+,-./:;=?@^_`{ }~]+@[a-z0-9]+@[a-z0-9]+@[a-z0-9]+)}</code>

Description: This accepts RFC 2822 email addresses in the form: <br> blah@blah.com OR <br> Blah BL@blah.com!g@<br> RFC 2822 email = mailbox <br> mailbox = name+addr | add+spec <br> name+addr = [display-name] "+" add+spec <br> "+" <br> add+spec = local-part "@" domain <br> domain = rfc2822domain | rfc2822domain-literal <br> <br> local-part

31

**talend\***  
open data solutions

## The World is Changing...

- \* Greater access to tools
- \* Shared specifications
- \* More data available in the public domain
- \* More transparency and sharing of data
- \* More reusable and extensible tools

\* Benefits

- Price pressures from Open Source
- New low-cost ways to implement data governance

© Talend 2010



## Questions and Answers



Sebastiao Correia	<a href="mailto:scorreia@talend.com">scorreia@talend.com</a>
Steve Sarsfield	<a href="mailto:ssarsfield@talend.com">ssarsfield@talend.com</a>

26

## Entity and Identity Resolution

### ABSTRACT

---

Entity resolution (ER) is concerned with determining whether two entity references (records) point to the same or to different real-world entities. ER, sometime called Identity Resolution, is tightly coupled with Information Quality, and is a key component of Customer Data Integration (CDI), fraud detection, law enforcement, and national intelligence. The tutorial will cover

- The five major ER activities
  - Entity Reference Extraction
  - Entity Reference Preparation
  - Entity Reference Resolution
  - Identity Management
  - Entity Relationship Exploration
- How record linking is different from record matching
- Entity Resolution versus Identity Resolution
- The role of “identity” in ER
- Current architectures and techniques used in ER processing
- Metrics for evaluating ER outcomes
- Demonstration of an ER System

The objective of the tutorial is to give participants a better understanding of the principles and terminology of entity resolution, and how it is being implemented in by leading companies and government agencies.

### BIOGRAPHY

---

#### John R. Talburt

Director, ERIQ Laboratory  
University of Arkansas at Little Rock

Dr. John R. Talburt is Professor of Information Science and Axiom Chair of Information Quality at the University of Arkansas at Little Rock (UALR) where he serves as the Coordinator for the Information Quality Graduate Program. He also holds appointments as Executive Director of the UALR Laboratory for Advanced Research in Entity Resolution and Information Quality (ERIQ), Associate Director of the Axiom Laboratory for Applied Research (ALAR), and Co-Director of the MIT Information Quality Program’s Working Group on Customer-Centric Information Quality Management. He also serves a Technical Advisor to the Board of Directors of the International Association for Information and Data Quality (IAIDQ), the only international professional organization devoted entirely to the field of information and data quality.





Prior to his appointment at UALR he was a leader for research and development and product innovation at Acxiom Corporation, a global leader in information management and customer data integration. Professor Talburt is an inventor for several patents related to customer data integration and the author of numerous articles on information quality and entity resolution. He is a co-editor of the textbook *Data Engineering: Mining, Information and Intelligence* (Springer, 2009), and is the author of the forthcoming textbook *Entity Resolution and Information Quality* scheduled for publication by Morgan Kaufmann in November, 2010.

His current research interests are at the intersection of information quality and information integration, particularly the areas of entity and identity resolution. Dr. Talburt is the winner of the 2008 DAMA International Academic Award and has earned the designation of Certified Data Management Professional (CDMP) from the Institute for Certification of Computing Professionals at the mastery-level with specialties in data and information quality and information technology.



# Entity and Identity Resolution

MIT IQ Industry Symposium  
July 14, 2010

John Talburt, PhD, CDMP  
Department of Information Science



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Topics

- Principles of Entity Resolution
- Entity Resolution Models



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## PRINCIPLES OF ER



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

### Pair-wise Definition

- ER is the process of determining whether two references to real-world objects are referring to the same, or to different, objects.
- **Entity** – because of the real-world object
- **Resolution** – because it poses a question



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Dataset Definition

- The process of identifying and merging records judged to represent the same real-world entity (Stanford InfoLab)
- Systematic and successive application of pair-wise resolution to a larger set of references



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

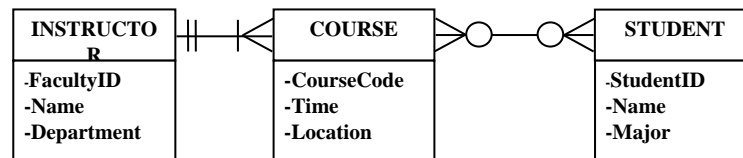
## Entity-Relation Model (ERM)

- Foundation of modern data models
- **Entity Types** define objects that have
  - **Attributes**
  - Attributes have **values** that describe a particular **instance** of an entity type
- **Relations** define connections between entity types
- **Identity attributes** – attributes whose values distinguish one instance from another



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Example



## Primary Key Problem

- Every table should have one
- Simplifies bringing together information about the same entity
  - **Table Join Operation**
- Problems
  - Different tables/databases often use different keys for same entity instance
  - Some records may not have keys
  - **Heterogeneous database join**



## ER Principle #1

- IS store and manipulate references to entities, not the entities.
- Entities are real-world objects --  
References are rows in a database table
  - In ER, instance of STUDENT entity type is a reference to a student -- the student is a person walking around campus
  - Data modelers call an instance an “entity”, but in ER it’s a reference



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## CDI

- When entity type is a customer, ER is called Customer Data Integration (CDI)
- Essential to support Customer Relationship Management (CRM)



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Big ER – Five Activities

1. Entity Reference Extraction
2. Entity Reference Preparation
3. Entity Reference Resolution
4. Entity Identity Management
5. Entity Relationship Analysis



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Entity Reference Extraction

- Identifying and extract entity reference from unstructured information
  - Free Text
  - Audio
  - Video
- Easy for people, hard for computers
- 80% of an organizations information is in unstructured text – reports, email, etc. – (Inmon, Nesavitch)



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Entity Reference Preparation

- Where IQ meets ER
- References are often
  - Incomplete
  - Inaccurate
  - Inconsistently represented, etc.
- Degrade ER processes and outcomes
- Reference clean-up often consumes large portion of ER effort



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Entity Reference Resolution

- Terminology : Linking vs. Matching
- Two references to the same entity are **equivalent** and should be **linked**
- **Matching** reference have the same (or mostly the same) identity attribute values
  - Matching records may not be equivalent
  - Equivalent records may not match
  - Mary Doe, Elm St – Mary Smith, Oak St
  - John Doe, Elm St – John Doe, Elm St



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## ER Principle #2

- ER is about linking equivalent references – matching is a means to an end
- Fundamental Law of ER  
Two entity references should be linked if and only if they reference the same entity (i.e. are equivalent).



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## False Negatives/Positives

- Two equivalent references that are not linked makes a **False Negative**
- Two non-equivalent references that are linked makes a **False Positive**
- Matching attribute values between two references is the most common (an intuitive) basis for making linkage decision, **but not the only one**



UNIVERSITY OF ARKANSAS AT LITTLE ROCK



### ER Principle #3

- False negative links are a more difficult problem to detect and solve in ER than false positive links
- Because ratio of true positives to true negatives is usually small – more non-links to checks for false, than links to check for false
- By definition, system doesn't give you something to look at



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

### ER Principle #4

- ER processes are generally designed to favor false negatives over false positives
- In business applications - Impact of a false positive decision is considered higher than impact of false negative decision – In other applications may be different
- False negative decisions are easier to defend than false positive decisions



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Identity Resolution

- Identity resolution is resolving an entity reference against a collection of known identities
- When known identities are for customers it is called **Customer Recognition**
- Identity resolution implies ER, but ER does not imply identity resolution



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## ER Principle #5

- Entity resolution is not the same as identity resolution
- Like fingerprints at a crimes scene
  - Can determine if two sets are for same or different suspects without knowing identity
  - Must get a “hit” against fingerprint database of known identities to identify
- Determining that references are to different entities without identifying them is called **disambiguation**



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Entity Identity Management

- All ER systems use identity, but not all systems manage (store and update) identity information
- ER system that manage identity can append **persistent links** -- consistently assign references to the same entity the same link identifier over time
- Allows transactional ER processing
- Allows linking by association and assertion



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## ER Principle #6

- ER systems that provide persistent link values must also implement some form of identity management
- Identity resolution systems
- Identity capture systems
  - “smart” merge-purge



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Four Methods for Linking

By

- Direct Matching
- Transitive Linking
- Linking by Association
- Asserted Linking



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## By Direct Matching

- Comparing the attributes between two references
- **Deterministic matching** – link if and only if all attributes agree
- **Probabilistic matching** – link if and only if certain combinations of attributes agree
- **Fuzzy matching** – “similar” attribute values can be counted as “agreeing”



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## By Transitive Linking

- Linking references through a chain of intermediate links
- If A links to B, B links to C, then A links to C
- Also called transitive closure
- Example: Probabilistic match on 2 out of 3 attributes
  - “Joe, GX, 56” matches “Joe, GX, 75”
  - “Joe, GX, 75” matches “Joe, TW, 75”
  - Link “Joe, GX, 56” and “Joe, TW, 75”



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## By Association

- Linking entity references based on relationships to other entities
- Example
  - Have an established link between “John Doe, Elm St” and “John Doe, Oak St”
  - Household association between “John Doe, Elm St” and “Sue Doe, Elm St”
  - Household association between “John Doe, Oak St” and “Sue Doe, Oak St”
  - Decision to link by association “Sue Doe, Elm St” and Sue Doe, Oak St”



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## By Assertion

- Linking references based on information from a reliable, external source – **knowledge-based linking**
- Example  
Magazine publisher reports that “Mary Doe, Oak St” is the same subscriber as “Mary Smith, Elm St”



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Approximate (Fuzzy) Matching

- Approximate String Matching (ASM) is based on the similarity of two strings in terms of shared characters and character sequences (Syntax)
  - “KELLEY” and “KELLY” differ by 1 char
- Alias Matching is based on the similarity of two strings in terms of their meaning (Semantics)
  - “ED” and “EDWARD” differ by 4 chars, but one is a “nickname” for the other



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## ASM – Edit Distance

- Levenshtein Edit Distance
  - Minimum number of transformations needed to change one string into another (delete, insert, replace)
  - “SALLIE” to “SALLY” distance = 2
  - Usually normalized by length of longest string, e.g.  $(6-2)/6 = 4/6 = 0.667$
  - Does not consider phonetic similarity
  - Does not consider position of difference
    - “THOMPSON” to “THOMAS” = 3
    - “THOMPSON” to “COMPTON” = 3



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## ASM - Soundex

- Capitalize all letters, drop punctuation
- Remove 'A', 'E', 'I', 'O', 'U', 'H', 'W', 'Y' after the first letter
- Change letters to digits as
  - 1 = 'B', 'F', 'P', 'V'
  - 2 = 'C', 'G', 'J', 'K', 'Q', 'S', 'X', 'Z'
  - 3 = 'D', 'T'
  - 4 = 'L'
  - 5 = 'M', 'N'
  - 6 = 'R'



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Soundex (Continued)

- Replace consecutive sequences of same digit with a single digit if original letters were adjacent or separated by "H" or "W"
- Truncate or pad with zeros to make a total of 4 characters
- Example:
  - PHILLIP – PLLP – P441 – P41 – P410
  - PHILIP – PLP – P41 – P410
  - PETERSON – PTRSN – P3625 – P362



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Soundex Examples

- LEE -> L000 (both "E"s are dropped)
- SHAW -> S000 ("H", "A", "W" in drop list)
- GAUSS->GSS->G22->G2->G200
- CHERRY->CRR->C66->C6->C600
- CHECKER->CCKR->C226->C26->C260
- COUSSACSK->C 22 222 ->C22->C220



UNIVERSITY OF ARKANSAS AT LITTLE ROCK



## Soundex Anomalies

- Group 1
  - LEE -> L000
  - LEIGH -> L200
  - LIU -> L000
- Group 2
  - GAUSS & GHOSH -> G200
  - WACHS & WAUGH -> W200
- Other issues
  - Lloyd, van Buren, von Munching



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## ASM - Jaro String Comparator

- Accounts for
  - Difference in length
  - Transposition of characters  
“JHON” vs “JOHN”
  - Number of characters in common
- Let  $s_1$  and  $s_2$  be strings
  - If index of char  $x$  is  $n_1$  in  $s_1$
  - If index of char  $x$  is  $n_2$  in  $s_2$
  - If  $|n_1 - n_2| \leq \min\{|s_1|, |s_2|\} / 2$
  - Then  $x$  is counted as a common char



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Jaro Formula

If  $c > 0$  then

$$\Phi(s_1, s_2) = W_1 \cdot \frac{c}{L_1} + W_2 \cdot \frac{c}{L_2} + W_3 \cdot \frac{(c - \tau)}{c}$$

$W_1$  = Weight assigned to first string

$W_2$  = Weight assigned to second string

$W_3$  = Weight assigned to transpositions

$W_1 + W_2 + W_3 = 1$

$c$  = common character count

$L_1$  = Length of first string

$L_2$  = Length of second string

$\tau$  = Number of chars transposed

If  $c = 0$  then  $\Phi(s_1, s_2) = 0$



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Example 1

- Higbee – Higvee
- $L_1 = L_2 = 6$ ,  $c = 5$ ,  $\tau = 0$ ,  $W_1 = W_2 = W_3 = 1/3$

$$\Phi_J(s_1, s_2) = W_1 \cdot \frac{c}{L_1} + W_2 \cdot \frac{c}{L_2} + W_3 \cdot \frac{(c - \tau)}{c}$$

$$= \left(\frac{1}{3}\right) \cdot \left(\frac{5}{6}\right) + \left(\frac{1}{3}\right) \cdot \left(\frac{5}{6}\right) + \left(\frac{1}{3}\right) \cdot \left(\frac{5 - 0}{5}\right)$$

$$= 8/9 = 0.889$$



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Example 2

- Shackleford– Shackelford
- $L_1 = L_2 = 11, c = 11, \tau = 2, W_1=W_2=W_3=1/3$

$$\begin{aligned}\Phi_J(s_1, s_2) &= W_1 \cdot \frac{c}{L_1} + W_2 \cdot \frac{c}{L_2} + W_3 \cdot \frac{(c-\tau)}{c} \\ &= \left(\frac{1}{3}\right) \cdot \left(\frac{11}{11}\right) + \left(\frac{1}{3}\right) \cdot \left(\frac{11}{11}\right) + \left(\frac{1}{3}\right) \cdot \left(\frac{11-2}{11}\right) \\ &= 31/33 = 0.939\end{aligned}$$

## ASM- Winkler String Comparator

- Modification of the Jaro Comparator
- Gives higher weight to agreement of initial characters of strings

$$\Phi_W(s_1, s_2) = \Phi_J(s_1, s_2) + i \cdot 0.1 \cdot (1 - \Phi_J(s_1, s_2))$$

- Where
  - $i = \min\{j, 4\}$
  - $j = \text{number of initial chars in common}$
- Example Shackleford – Shackelford
- $= 0.939 + 4 \cdot 0.1 \cdot (0.061) = 0.963$

## Other ASM

- n-grams (q-grams) based on number of shared substrings of length n
- LCS - longest common substring
- Variations of Soundex
  - NYSIIS - New York State Identification and Intelligence System – avoids first letter problem
  - Phonex – preprocess names before using Soundex
  - Phonix – an improved version of Phonex



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## ER MODELS



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Fellegi-Sunter Model

- Standard for probabilistic matching
- Context
  - Two unduplicated lists of references A, B
  - Both lists have N corresponding identity attributes
- Given a false positive rate P and false negative rate N, the model defines a linking strategy that will
  - Not exceed P and N,
  - Minimize cases requiring intervention



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Fellegi-Sunter Conditions

- A and B two lists of references
- Consider  $A \times B$  (all pairs)
- $M$  = True positives, i.e.  $(a, b) \in M$  if and only if "a" should be linked to "b"
- $U$  = True negatives, i.e.  $(a, b) \in U$  if and only if "a" should NOT be linked to "b"
- $\Gamma$  = all attribute match/no-match combinations of the N attributes. There will be  $2^N$  of these.



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Fellegi-Sunter Weight Ratios

- For an agreement pattern  $\gamma \in \Gamma$  define

$$R_\gamma = \frac{P(\gamma \in \Gamma | r \in M)}{P(\gamma \in \Gamma | r \in U)}$$

- Represents the ratio of the “probability of Good Links” to “probability of Bad Links” for a given match pattern
- Very large value means good link rule
- Very small value means bad link rule

## Fellegi-Sunter (cont)

- Establish two values U (upper) and L (lower) in the series of decreasing values of R

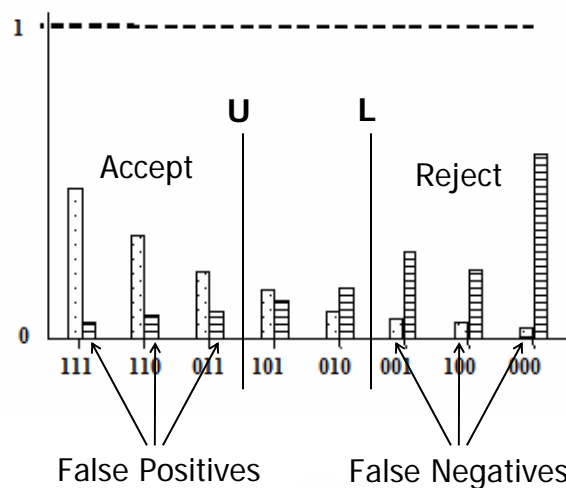
$$\underbrace{R_1 > R_2 > \dots > U}_{\text{Accept}} > \dots > \underbrace{R_n > R_{n+1} > \dots > L}_{\text{Review}} > \dots > \underbrace{R_m > R_{m+1}}_{\text{Reject}}$$

- U and L are selected so that N and P (respectively) are not exceeded

## Example: Student Records

- Two enrollments list from consecutive years
- Match first name, last name, DOB
- Expect large overlap, but
  - Some first year students leave
  - Some new students second year
- Not all records have DOB
- Use 3-bit binary numbers to represent agreement patterns

## True and False Positives



## Stanford SERF Model

- Developed at Stanford InfoLab
- **S**tanford **E**ntity **R**esolution **F**ramework
- Intended to be a “generic” ER Model
- Fellegi-Sunter gives a way to evaluate matching, SERF does not
- SERF does describe
  - Conditions that must hold for ER outcome to be unique
  - How pair-wise matching can resolve a set (merge-purge algorithm)



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Match and Merge Functions

- R is set of references
- Two functions defined
  - Match Function M
    - $M: D \times D \rightarrow \{\text{true}, \text{false}\}$
    - $R \subset D$  the domain of M
  - Merge Function  $\mu$ 
    - If  $a, b \in D, M(a, b) = \text{true}$ , then  $\mu(a, b) \in D$
- Definition
  - If  $\mu(a, b) = a$ , then “a dominates b”



UNIVERSITY OF ARKANSAS AT LITTLE ROCK



## SERF definition of ER

$ER(R) \subseteq D$  such that

- Any record that can be derived from  $R$  is either in  $ER(R)$  or is dominated by a record in  $ER(R)$
- No two records in  $ER(R)$  match and no record in  $ER(R)$  is dominated by any other
- Think of merged records in  $ER(R)$  as clusters of equivalent records



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Consistent ER

- Consistent ER means that  $ER(R)$  exists, is finite, and is unique
- Will be consistent if the following condition hold
  - $M(a, b) = M(b, a) \ \& \ \mu(a, b) = \mu(a, b)$
  - $M(a, a) = \text{true} \ \& \ \mu(a, a) = a$
  - $M(a, \mu(a, b)) = M(b, \mu(a, b)) = \text{true}$
  - $\mu(a, \mu(b, c)) = \mu(\mu(a, b), c)$



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## R-Swoosh Algorithm

- Systematic way to find  $ER(R)$  if match & merge functions are consistent
1. Start:  $D = R$ , and  $ER(R) = \emptyset$
  2. Start comparing first record  $\mathbf{x}$  in  $D$  to each record  $\mathbf{y}$  in  $ER(R)$
  3. If  $M(\mathbf{x}, \mathbf{y}) = \text{true}$ 
    - Stop comparing
    - Replace  $\mathbf{x}$  in  $D$  with  $\mu(\mathbf{x}, \mathbf{y})$
    - Remove  $\mathbf{y}$  from  $ER(R)$



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## R-Swoosh Algorithm (cont)

4. If  $M(\mathbf{x}, \mathbf{y})$  not true for any  $\mathbf{y}$  in  $ER(R)$ 
  - Put  $\mathbf{x}$  in  $ER(R)$
  - Remove  $\mathbf{x}$  from  $D$
5. If more items in  $D$  to process, go back to Step 3, otherwise algorithm is finished



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Example: D at Start of Process

	First	Last	DOB	SCode
r1	Edgar	Jones	20001104	G34
r2	Mary	Smith	19990921	G55
r3	Eddie	Jones	20001104	G34
r4	Mary	Smith	19990921	H17
r5	Eddie	Jones	20001104	H15

• Match if references agree on

– First, Last, DOB, or Last, DOB, SCode

• Merge combines attributes



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Example: ER(R) at End

	First	Last	DOB	SCode
r7	Mary	Jones	20001104	{H17,G55}
r8	{Eddie, Edgar}	Jones	20001104	{G34, H15}

• r7 represents original r1, r3, r5

• r8 represents original r2 and r4



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Algebraic Model (Background)

### Definitions

- Given a set  $S$  and a subset  $T \subseteq S \times S$ , then  $T$  is said to be a relation on  $S$
- $T$  is said to be an equivalence relation on  $S$  if and only if
  - For every  $a \in S$ , then  $(a, a) \in T$
  - If  $(a, b) \in T$ , then  $(b, a) \in T$
  - If  $(a, b) \in T$  and  $(b, c) \in T$ , then  $(a, c) \in T$



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Background Continued

- If  $T$  is an equivalence relation on  $S$  then  $[a] = \{b \in S \mid (b, a) \in T\}$  is the equivalence class of  $a$
- A partition  $P$  of a set  $S$  is a collection of subsets  $P_1, P_2, \dots, P_n$  such that
  - $P_j \neq \emptyset$  for all  $j=1 \dots n$
  - $P_j \cap P_k = \emptyset$  whenever  $j \neq k$
  - $S = P_1 \cup P_2 \cup \dots \cup P_n$
- If  $T$  is an equivalence relation on  $S$  then  $P = \{[a] \mid a \in S\}$  is a partition of  $S$ .



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Algebraic Model Defined

- Defines ER only in terms of outcome
  - Let  $R$  be a set of references where every  $a \in R$  references one and only one real-world object
  - Define  $E \subseteq R \times R$  by  $(a, b) \in E$  if and only if  $a$  and  $b$  reference the same real-world object.
- Then
  - $E$  is an equivalence relation on  $R$
  - The equivalence classes of  $E$  define a unique partition of  $R$



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## From Previous Example

- $R = \{r1, r2, r3, r4, r5\}$ , then
- $E = \{(r1, r1), (r2, r2), (r3, r3), (r4, r4), (r5, r5), (r1, r3), (r3, r1), (r1, r5), (r5, r1), (r3, r5), (r5, r3), (r2, r4), (r4, r2)\}$
- Partition defined by  $E$  is  
 $P(E) = \{\{r1, r3, r5\}, \{r2, r4\}\}$



UNIVERSITY OF ARKANSAS AT LITTLE ROCK

## Comparing ER Outcomes

- Comparing ER outcomes is same as comparing partitions
- Let P and Q be two partitions of S
- Define  $V = \{P_j \cap Q_k \mid P_j \cap Q_k \neq \emptyset\}$
- The Talburt-Wang Similarity Index (TWI) is defined by

$$\text{TWI} = \frac{\sqrt{|P| \cdot |Q|}}{|V|}$$

- TWI is a number from 0 to 1
- TWI = 1 iff P = Q



## Example

- $S = \{a, b, c, d, e, f, g, h\}$
- $P = \{\{a, d, e\}, \{b\}, \{c, f, g\}, \{h\}\}$
- $Q = \{\{a, b, d\}, \{e\}, \{c, f\}, \{g\}, \{h\}\}$
- $V = \{\{a, d\}, \{e\}, \{b\}, \{c, f\}, \{g\}, \{h\}\}$
- $|P| = 4, |Q| = 5, |V| = 6$
- $\text{TWI} = \text{SQRT}(4 \times 5)/6 = \text{SQRT}(20)/6 = 0.745$



## Questions and Discussion

John R. Talburt

[jrtalbert@ualr.edu](mailto:jrtalbert@ualr.edu)

Coming in November

***Entity Resolution and  
Information Quality***, Morgan  
Kaufmann Publishers



UNIVERSITY OF ARKANSAS AT LITTLE ROCK