

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
- * **Steve Sarsfield**
 - Product Manager
 - Blogger
 - 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

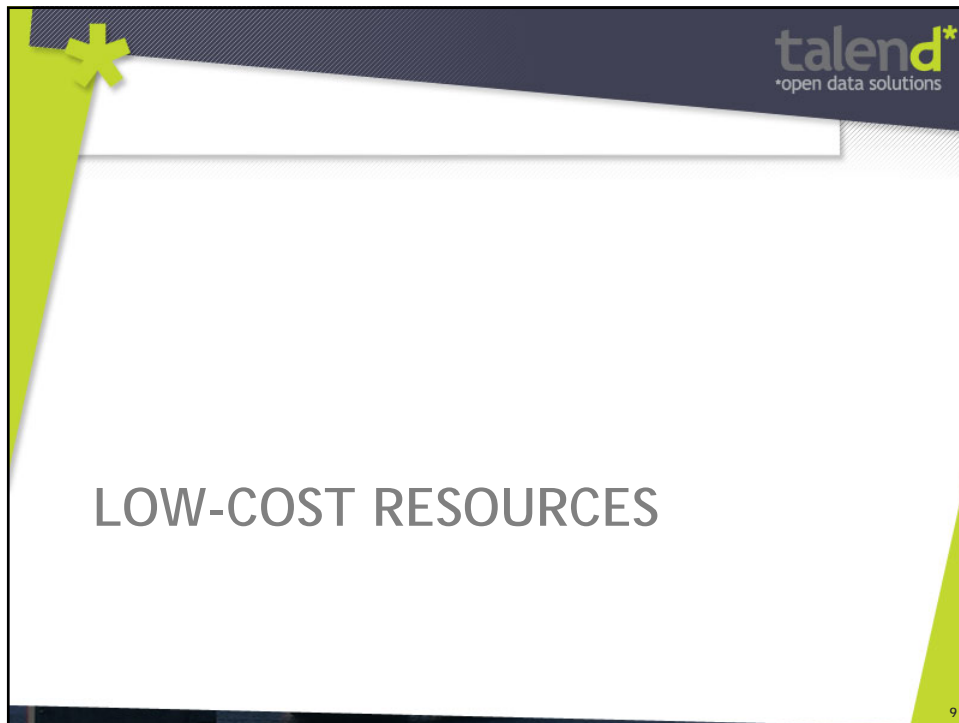
7




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






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



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



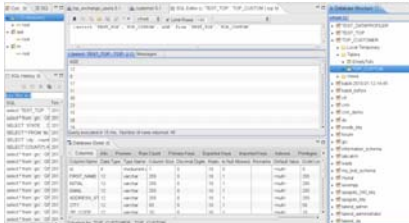

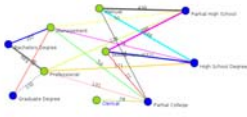
Tools - You have Options

- * **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

talend*
open data solutions

Reporting Data Quality Metrics

The screenshot shows a web interface for 'talend Data Quality'. The main content area is titled 'Birthdate Null tracking' and contains two visualizations. The first is a line chart titled 'Null Birthdate Evolution' showing the percentage of null birthdates over time from 16-Jan to 28-Apr. The second is a gauge chart titled 'COUNT OF NULL BIRTHDATE' showing the distribution of null birthdates, with a total of 314 rows. The gauge shows a large red segment (20,000) and a smaller yellow segment (100).

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

talend*
open data solutions

Unified Data Management - Tools

The diagram illustrates a workflow of data management tools. The tools are listed in a sequence from left to right: Connectivity, Data Assessment, Metadata Management, Transformation, Cleansing, Standardization, Matching, Monitoring, Stewardship, and Active Data Model. These tools are grouped into three main categories represented by arrows at the bottom: Data Integration (covering Connectivity, Data Assessment, and Metadata Management), Data Quality (covering Transformation, Cleansing, and Standardization), and Master Data Management (covering Matching, 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


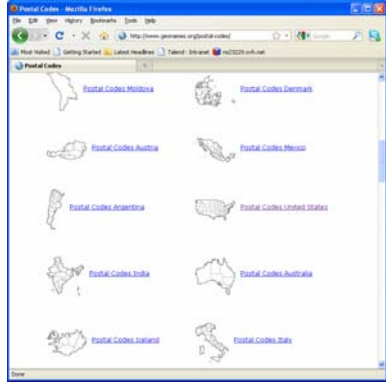
- * **Open data resources is getting structured**
 - Open source like initiatives
 - Government initiatives
- * **8 Principles of Open Data**
 - Complete: All public data is made available.
 - Primary: Data is as collected at the source
 - Timely: Data is made available as quickly as necessary
 - Accessible: 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



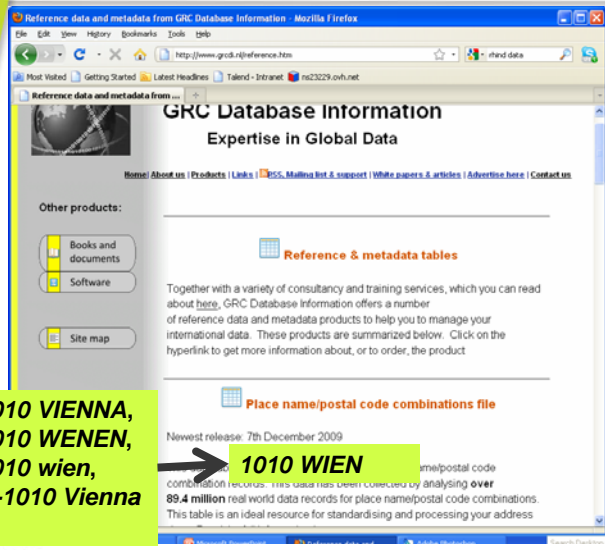
Benefits for citizens

- * **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

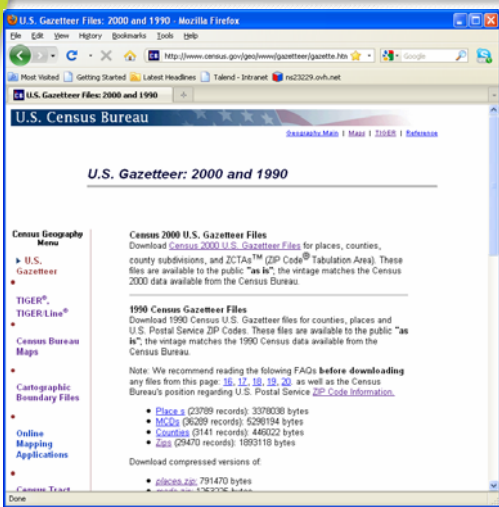


- * Common misspellings of:
 - Name
 - Address
 - Job Title
 - More...
- * Gathered by examining real-life data and developing tables.

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?



The screenshot shows the RegExLib.com website. The main content area displays search results for the category 'Email'. The first result is titled 'email address (RFC 2822 mailbox)' and includes the following regular expression: `^[iI](?!a-z\d+)(?![_-])[a-z\d+](?!@)(?![^\s@]+@)[a-z\d+](?!@)@(?!@)[a-z\d+](?!@)(?![^\s@]+@)[a-z\d+](?!@)$`. The description explains that this accepts RFC 2822 email addresses in the form: `
blah@blah.com
 Blah BL@blah.com!g

 RFC 2822 email=mailbox
 mailbox = name+addr | add+spoc
 name+addr = [display-name] "+" add+spoc
 "+"
 add+spoc = local-part "@" domain
 domain = rfc2822domain | rfc2822domain-literal

 local-part`

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 simple
- [Regex Examples for Processing Source Code](#). How to
- [Two Words Near Each Other](#). Shows how to use a re

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	scorreia@talend.com
Steve Sarsfield	ssarsfield@talend.com

26