

"... a comprehensive, detailed guide ..."

JOHN R. TALBURT, PhD, UNIVERSITY OF ARKANSAS AT LITTLE ROCK

Data Governance Tools

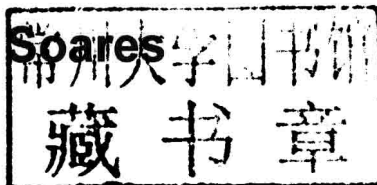
*Evaluation Criteria, Big Data Governance, and
Alignment with Enterprise Data Management*

SUNIL SOARES



Data Governance Tools

Sunil Soares



MC Press Online, LLC
Boise, ID 83703 USA

Data Governance Tools: Evaluation Criteria, Big Data Governance, and Alignment with Enterprise Data Management
Sunil Soares

First Edition

© Copyright 2014 Sunil Soares. All rights reserved.

Printed in Canada. *All rights reserved.* This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, contact mcbooks@mcpressonline.com.

Every attempt has been made to provide correct information. However, the publisher and the author do not guarantee the accuracy of the book and do not assume responsibility for information included in or omitted from it.

Ab Initio is a registered trademark of Ab Initio Software Corporation. Activiti is a registered trademark of Alfresco Software, Inc. ADABAS is a registered trademark of Software AG. Adaptive is a trademark or registered trademark of Adaptive Computing Enterprises, Inc. Adobe, Acrobat, and Reader are registered trademarks of Adobe Systems Incorporated in the United States and/or other countries. Amazon, DynamoDB, EC2, Elastic Compute Cloud, and Redshift are trademarks of Amazon.com, Inc., or its affiliates. Apache, Cassandra, CouchDB, Flume, Hadoop, HBase, Hive, Oozie, Pig, and Sqoop are trademarks of The Apache Software Foundation. ASG, ASG-becubic, ASG-metaGlossary, ASG-MyInfoAssist, and ASG-Rochade are trademarks or registered trademarks of ASG. Remedy is a registered trademark or trademark of BMC Software, Inc. ERwin is a registered trademark of CA, Inc. Clarabridge is a trademark of Clarabridge, Inc. Cloudera and Cloudera Impala are trademarks of Cloudera, Inc. Collibra is a registered trademark of Collibra Corporation. Concur is a registered trademark of Concur Technologies, Inc. Constant Contact is a registered trademark of Constant Contact in the United States and other countries. Couchbase is a registered trademark of Couchbase, Inc. ActiveLinux and MetaCenter are trademarks of Data Advantage Group, Inc. Denodo is a registered trademark of Denodo Technologies. Diaku and Diaku Axon are the trademarks of Diaku Ltd. Eclipse is a trademark of Eclipse Foundation, Inc. Eloqua is a trademark of Eloqua Corporation. Embarcadero and all other Embarcadero Technologies product or service names are trademarks, service marks, and/or registered trademarks of Embarcadero Technologies, Inc. EMC, Archer, Documentum, Greenplum, Pivotal, RSA, and SourceOne are trademarks or registered trademarks of EMC Corporation in the United States and/or other countries. Facebook and the Facebook logo are registered trademarks of Facebook, Inc. Financial Industry Business Ontology (FIBO) is a trademark of the EDM Council. Force.com, Salesforce, and Salesforce.com are registered trademarks of salesforce.com. Google, Maps, and Search Appliance are trademarks or registered trademarks of Google, Inc. EnCase and Guidance Software are registered trademarks or trademarks owned by Guidance Software in the United States and other jurisdictions. Hortonworks is a trademark of Hortonworks Inc. HP and HP Vertica are trademarks of Hewlett-Packard Development Company, L.P. IBM, AS/400, BigInsights, CICS, Cognos, DataStage, DB2, Domino, Guardium, IMS, InfoSphere, MQSeries, Notes, OpenPages, Optim, QualityStage, PureData, and SPSS are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Imperva is a registered trademark of Imperva. Informatica, AddressDoctor, Informatica Cloud, and PowerCenter are trademarks or registered trademarks of Informatica Corporation in the United States and in foreign countries. InfoTrellis is a trademark or registered trademark of InfoTrellis, Inc., in Canada and other countries. JIRA is a trademark of Atlassian. MapR is a registered trademark of MapR Technologies, Inc., in the United States and other countries. Marketo is a trademark of Marketo, Inc. Microsoft, Azure, Excel, Exchange, Outlook, SharePoint, SQL Server, Visual Basic, and Word are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. MongoDB is a registered trademark of MongoDB, Inc. Netezza is a registered trademark of IBM International Group B.V., an IBM Company. NetSuite is a registered trademark of NetSuite, Inc. All Nuix trademarks are the property of Nuix Pty Ltd. OpenText is a trademark or registered trademark of Open Text SA and/or Open Text ULC. Oracle, Endeca, Exalytics, Java and all Java-based trademarks and logos, and MySQL are trademarks or registered trademarks of Oracle and/or its affiliates. Orchestra Networks is a registered trademark of Orchestra Networks in France and in jurisdictions throughout the world. Pega is a registered trademark of Pegasystems, Inc. Pentaho is a registered trademark of Pentaho, Inc. Protegrity is a registered trademark of Protegrity Corporation. QlikView is a registered trademark of Qlik Technologies, Inc., or its subsidiaries in the United States, other countries, or both. Recommind and Accelerate are trademarks or registered trademarks of Recommind or its subsidiaries in the United States and other countries. Riak is a registered trademark of Basho Technologies, Inc. Sage is a registered trademark of Sage Software, Inc. SAP, BusinessObjects, HANA, NetWeaver, PowerDesigner, and Sybase are trademarks and registered trademarks of SAP SE in Germany and other countries. SAS is a registered trademark of the SAS Institute, Inc. Semarchy and Convergence are trademarks or registered trademarks of Semarchy. Symantec and Enterprise Vault are trademarks or registered trademarks of Symantec Corporation or its affiliates in the United States and other countries. Tableau is a registered trademark of Tableau Software. Talend and Talend ESB are trademarks of Talend, Inc. Teradata and Aster are registered trademarks of Teradata Corporation and/or its affiliates in the United States and worldwide. TIBCO and StreamBase are trademarks or registered trademarks of TIBCO Software, Inc., or its subsidiaries in the United States and/or other countries. Trillium Software, The Trillium Software System, and/or other Trillium Software, A Harte Hanks Company products referenced herein are either registered trademarks or trademarks of Trillium Software, A Harte Hanks Company Corporation in the United States and/or other countries. Twitter and the Twitter logo are registered trademarks of Twitter, Inc. Yahoo! is a registered trademark of Yahoo, Inc., in the United States, other countries, or both. ZyLAB is a registered trademark of ZyLAB North America. Other company, product, or service names may be trademarks or service marks of others.

MC Press offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include custom covers and content particular to your business, training goals, marketing focus, and branding interest.

MC Press Online, LLC 3695 W. Quail Heights Court, Boise, ID 83703-3861 USA • (208) 629-7275
service@mcpressonline.com • www.mcpressonline.com • www.mc-store.com

ISBN: 978-1-58347-844-8

WB201410

Dedicated to my beautiful daughters, Maya and Lizzie.

Many thanks to my wife Helena, who came up with the idea for this book.

A big thanks to my parents Cecilia and Hubert for their prayers and guidance.

I also want to acknowledge the Information Asset team, including Jatin Bhoir, Michelle D'Sa, Royson Mendonca, Yanxin Shi, and Dorothy Xavier. The Enterprise Data Management lab is a critical success factor in our client engagements and in the development of this book.

ABOUT THE AUTHOR

Sunil Soares is the founder and managing partner of Information Asset, a consulting firm that specializes in data governance. Prior to this role, Sunil was director of information governance at IBM, where he worked with clients across six continents and multiple industries. Before joining IBM, Sunil consulted with major financial institutions at the Financial Services Strategy Consulting Practice of Booz Allen & Hamilton in New York.

Sunil's first book, *The IBM Data Governance Unified Process* (MC Press, 2010), details the almost 100 steps to implement a data governance program. This book has been used by several organizations as the blueprint for their data governance programs and has been translated into Chinese. Sunil's second book, *Selling Information Governance to the Business: Best Practices by Industry and Job Function* (MC Press, 2011), reviews the best practices to approach information governance by industry and function. His third book, *Big Data Governance* (MC Press, 2012), addresses the specific issues associated with the governance of big data.

Sunil lives in New Jersey and holds an MBA in Finance and Marketing from the University of Chicago Booth School of Business.

FOREWORD

by Aditya Kongara

Enterprise Data Management (EDM) over the past few years has quickly become an important discipline as organizations look to establish governance over their information assets. Effective data management needs the three pillars of people, process, and technology to be mature and well-functioning.

I have spent the majority of my career in large financial services organizations and working with Big Four consulting firms setting up data management and governance programs. In my opinion, the technology pillar of EDM is as important as the other two pillars.

Assume you are the data governance lead at a large bank that has to pass a data audit from the regulators. The bank's systems consist of hundreds of thousands of data elements spread over hundreds of databases and schemas. How do you demonstrate data lineage to the regulators without a metadata tool? Are you able to convince the Chief Information Security Officer that all instances of sensitive data have been discovered? Can you do that without a data discovery tool? Are your SQL queries robust and automated enough to produce data quality scorecards on a regular basis? For these reasons and others listed in the book, I feel that companies will increasingly have to rely on data management tools to automate various manual tasks.

I have known Sunil Soares for many years in a variety of job roles. I am excited by his knowledge and passion for data governance and for his thought leadership around tools. This book is a great read for any practitioner who wants to be successful in the data management and governance field.

Aditya Kongara
Head of Enterprise Data Management
American Family Mutual Insurance Company

FOREWORD

by John R. Talburt

This book on data governance tools could not have come at a better time for the field of information quality. I say this having been in the most fortunate position to observe the explosive growth and evolution of information and data quality over the past three decades, from both a practitioner and academic perspective. Given this perspective, let me start by giving a bit of background that I think explains why this book is so timely.

Deeply rooted in practice, the emerging field of information quality had its genesis in the seemingly endless data cleaning efforts that were necessary to launch the data warehousing movement of the 1980s. From cleaning and correcting data, it started to mature, first embracing root cause analysis, then later fully adopting and incorporating the principles of TQM (Total Quality Management). Having embraced the concept of managing information as product, it continued to develop and mature. In its current incarnation, information quality goes far beyond just repairing things gone wrong, to having a seat at the table for information architecture planning and design, and now is an integral part of information policy and strategy in the role of data governance.

Like data warehousing, data governance is one of those new ideas that in retrospect seems so obvious. Why wouldn't any enterprise want to have a clear policy around and a shared understanding of its information assets? But like data warehousing, it has taken some time to "iron out the wrinkles" and make data governance really work. Now that we know that it does work, the competitive advantage imparted by a well-defined data governance program has elevated it to an essential part of corporate strategy.

Accepting data governance as essential is one thing, but making it work is another. In the early years of information quality, everyone had to develop their own tools to try and get the job done. It was not long before the demand for easier tools with more functionality created a market demand that was addressed by the many data quality tool

vendors we see today. Now we see a repeat of this¹ cycle with data governance. Many vendors now offer various tools and suites of tools to help organizations implement data governance programs. However, one difference is that data governance programs are more diverse because the reasons for adopting them and their goals are often quite different.

This comes to the point of why this book is so timely and important. In one source, the reader can have an overview of the various categories of data governance tools and their key components. This book also gives a clear description of how and where these tools integrate into the data management strategy of the enterprise. Moreover, it is written by someone with extensive experience in data governance implementation, someone who has been there and knows how it works. This experience is reflected in the large amount of detail and concrete examples given in the book.

One really invaluable section of this book is the survey of data governance tools offered by the leading vendors. The overview will be a tremendous help to those still on the sidelines and getting ready to start a data governance program, as well as those who have started on their own, but now see the potential value in adopting a third-party system.

Another very helpful section is on big data governance tools. It contains a great discussion on the use of Hadoop MapReduce and NoSQL tools to gain insights into data. There are also sections explaining approaches to streaming computing and text analytics.

All in all, *Data Governance Tools* is a comprehensive, detailed guide to the landscape of data governance tools that will be valuable to everyone involved with enterprise data management, both from business and IT. I hope that everyone will take advantage of the wealth of information that it provides.

*John R. Talburt, PhD, IQCP
Director of the Information Quality Graduate Program
University of Arkansas at Little Rock*

FOREWORD

by Aaron Zornes

While Sunil's prior books represented a Rosetta Stone for IT professionals to map their traditional IT experiences (MDM, RDM, data governance, etc.) to big data, at last we now have a "Domesday Book" to categorize and better understand the vast menagerie of solutions that comprise the data governance software market. There is quite a lot more beyond Microsoft Excel and SharePoint, and Sunil's "reference architecture" provides the foundational touchstone.

Given the synergy and codependence between MDM and data governance, Sunil's latest book is a must read for any MDM practitioner who is charged with establishing or upgrading the data governance processes inherently necessary for enterprise MDM or RDM programs. Among other benefits, it provides a much appreciated reference architecture and set of evaluation criteria, as well as examples illustrating the practical application of these tools.

In my consultancy practice and experience, MDM and RDM mandate the application of data governance (not just people and processes, but also software tools) to be effective and sustainable. Clearly, data governance for MDM is moving beyond simple stewardship to convergence of task management, workflow, policy management, and enforcement. Moreover, it is now time for MDM vendors to instantiate their data governance marketing claims and finally move from "passive-aggressive" mode to "proactive" data governance mode. The evaluation criteria provided in this book is proof that MDM vendors have recently begun to deliver (especially IBM, Informatica, Orchestra Networks, and SAP).

Data Governance Tools is the plenary source that can successfully tutor and guide you into becoming a “data governance professional.” Moreover, it is a key asset that I’ll be sharing with the 3,000+ annual attendees of my MDM & Data Governance Summit series.

Aaron Zornes
Chief Research Officer, The MDM Institute
Conference Chairman, The MDM & Data Governance Summit
(London, New York City, San Francisco, Shanghai, Singapore, Sydney, Tokyo, Toronto)

PREFACE

Data governance is the formulation of policy to optimize, secure, and leverage information as an enterprise asset by aligning the objectives of multiple functions. Data governance programs have traditionally focused on people and process. Cost has historically been a key consideration because data governance programs have often started from scratch, with little to no funding. As a result, Microsoft Excel and SharePoint have been the tools of choice to document and share data governance artifacts. While the marginal cost of these tools is zero, they are often missing critical functionality. Meanwhile, vendors have matured their data governance offerings to the extent that organizations need to consider tools as a critical component of their data governance programs.

It is not always clear, however, what “data governance tools” really mean. In this book, I review a reference architecture for data governance software tools. I seek to define the category called “data governance,” as well as lay out evaluation criteria for software tools, the vendor landscape, and the alignment with big data.

This book consists of the following sections:

1. *Introduction*

The chapters in this section provide an introduction to data governance and the Enterprise Data Management (EDM) reference architecture.

2. *Categories of Data Governance Tools*

These chapters discuss key data governance tasks that can be automated by tools for business glossaries, metadata management, data profiling, data quality management, master data management, reference data management, and information policy management.

3. *The Integration Between Enterprise Data Management and Data Governance Tools*

This section is an overview of the integration points between EDM tools and data governance. EDM tools relate to data modeling, data integration, analytics and reporting, business process management, data security and privacy, and information lifecycle management.

4. *Big Data Governance Tools*

The chapters in this section provide an overview of how data governance tools interact with big data technologies, including Hadoop, NoSQL, stream computing, and text analytics.

5. *Evaluation Criteria and the Vendor Landscape*

This section is a review of the overall evaluation criteria for data governance tools. This section also provides an overview of key vendor platforms, including ASG, Collibra, Global IDs, IBM, Informatica, Orchestra Networks, SAP, and Talend.

This book is geared toward business users and is relatively nontechnical. Sample roles who might be interested in this book include the following:

- Chief Information Officer
- Chief Data Officer
- Data Governance Lead
- Business Intelligence Lead
- Data Warehousing Lead
- Enterprise Data Management Lead
- Chief Information Security Officer
- Chief Privacy Officer
- Chief Medical Information Officer

All the best, and happy reading.

CONTENTS

About the Author	iv
Forewords	xv
<i>by Aditya Kongara</i>	xv
<i>by John R. Talburt</i>	xvi
<i>by Aaron Zornes</i>	xviii
Preface	xxi
PART I—INTRODUCTION	1
1: An Introduction to Data Governance	3
Definition	3
Case Study	5
The Pillars of Data Governance	5
Summary	6
2: Enterprise Data Management Reference Architecture	7
EDM Categories	8
Big Data	13
Data Governance Tools	14
Summary	14
PART II—CATEGORIES OF DATA GOVERNANCE TOOLS	15
3: The Business Glossary	17
Bulk-Load Business Terms in Excel, CSV, or XML Format	17
Create Categories of Business Terms	20

Facilitate Social Collaboration	20
Automatically Hyperlink Embedded Business Terms	21
Add Custom Attributes to Business Terms and Other Data Artifacts	22
Add Custom Relationships to Business Terms and Other Data Artifacts	23
Add Custom Roles to Business Terms and Other Data Artifacts	23
Link Business Terms and Column Names to the Associated Reference Data	24
Link Business Terms to Technical Metadata	25
Support the Creation of Custom Asset Types	26
Flag Critical Data Elements	28
Provide OOTB and Custom Workflows to Manage Business Terms and Other Data Artifacts	28
Review the History of Changes to Business Terms and Other Data Artifacts	32
Allow Business Users to Link to the Glossary Directly from Reporting Tools	33
Search for Business Terms	34
Integrate Business Terms with Associated Unstructured Data	35
Summary	36
4: Metadata Management	37
Pull Logical Models from Data Modeling Tools	37
Pull Physical Models from Data Modeling Tools	38
Ingest Metadata from Relational Databases	40
Pull in Metadata from Data Warehouse Appliances	41
Integrate Metadata from Legacy Data Sources	42
Pull Metadata from ETL Tools	43
Pull Metadata from Reporting Tools	44
Reflect Custom Code in the Metadata Tool	45
Pull Metadata from Analytics Tools	47
Link Business Terms with Column Names	48
Pull Metadata from Data Quality Tools	48
Pull Metadata from Big Data Sources	50
Provide Detailed Views on Data Lineage	51
Customize Data Lineage Reporting	52
Manage Permissions in the Metadata Repository	55
Support the Search for Assets in the Metadata Repository	57
Summary	58
5: Data Profiling	59
Conduct Column Analysis	59

Discover the Values Distribution of a Column	61
Discover the Patterns Distribution of a Column	62
Discover the Length Frequencies of a Column	63
Discover Hidden Sensitive Data	64
Discover Values with Similar Sounds in a Column	65
Agree on the Data Quality Dimensions for the Data Governance Program	66
Develop Business Rules Relating to the Data Quality Dimensions	67
Profile Data Relating to the Completeness Dimension of Data Quality	69
Profile Data Relating to the Conformity Dimension of Data Quality	69
Profile Data Relating to the Consistency Dimension of Data Quality	71
Profile Data Relating to the Synchronization Dimension of Data Quality	71
Profile Data Relating to the Uniqueness Dimension of Data Quality	73
Profile Data Relating to the Timeliness Dimension of Data Quality	74
Profile Data Relating to the Accuracy Dimension of Data Quality	75
Discover Data Overlaps Across Columns	76
Discover Hidden Relationships Between Columns	80
Discover Dependencies	81
Discover Data Transformations	84
Create Virtual Joins or Logical Data Objects That Can Be Profiled	86
Summary	88
6: Data Quality Management	89
Transform Data into a Standardized Format	89
Improve the Quality of Address Data	93
Match and Merge Duplicate Records	95
Create a Data Quality Scorecard	98
<i>Select the Data Domain or Entity</i>	98
<i>Define the Acceptable Thresholds of Data Quality</i>	98
<i>Select the Data Quality Dimensions to Be Measured for the Specific Data Domain or Entity</i>	99
<i>Select the Weights for Each Data Quality Dimension</i>	99
<i>Select the Business Rules for Each Data Quality Dimension</i>	100
<i>Assign Weights to Each Business Rule in a Given Data Quality Dimension</i>	101
<i>Bind the Business Rules to the Relevant Columns</i>	102
View the Data Quality Scorecard	103
Highlight the Financial Impact Associated with Poor Data Quality	104
Conduct Time Series Analysis	104

Manage Data Quality Exceptions	106
Summary	108
7: Master Data Management	109
Define Business Terms Consumed by the MDM Hub	109
Manage Entity Relationships	111
Manage Master Data Enrichment Rules	112
Manage Master Data Validation Rules	113
Manage Record Matching Rules	114
Manage Record Consolidation Rules	116
View a List of Outstanding Data Stewardship Tasks	117
Manage Duplicates	119
View the Data Stewardship Dashboard	121
Manage Hierarchies	122
Improve the Quality of Master Data	122
Integrate Social Media with MDM	125
Manage Master Data Workflows	126
Compare Snapshots of Master Data	127
Provide a History of Changes to Master Data	128
Offload MDM Tasks to Hadoop for Faster Processing	129
Summary	131
8: Reference Data Management	133
Build an Inventory of Code Tables	134
Agree on the Master List of Values for Each Code Table	135
Build Simple Mappings Between Master Values and Related Code Tables	137
Build Complex Mappings Between Code Values	137
Manage Hierarchies of Code Values	139
Build and Compare Snapshots of Reference Data	140
Visualize Inter-Temporal Crosswalks Between Reference Data Snapshots	141
Summary	143
9: Information Policy Management	145
Manage Information Policies, Standards, and Processes Within the Business Glossary	147
Manage Business Rules	147
Leverage Data Governance Tools to Monitor and Report on Compliance	149
Manage Data Issues	149
Summary	157