White Paper

The Foundations of Successful Reference Data Management

INTRODUCTION

Reference data is a special type of data.

It is essentially codes whose basic job is to turn other data into meaningful business information and to provide an informational context for the wider world in which the enterprise functions.

Data management is becoming more and more central to the business model of enterprises. The time when data was looked at as little more than the by-product of automation is long gone, and today we see enterprises vigorously engaged in trying to unlock maximum value from their data, even to the extent of directly monetizing it. Yet, many of these efforts are hampered by immature data governance and management practices stemming from a legacy that did not pay much attention to data. Part of this problem is a failure to understand that there are different types of data, and each type of data has its own special characteristics, challenges and concerns.

Reference Data Management Overview

Reference data is a special type of data. It is essentially codes whose basic job is to turn other data into meaningful business information and to provide an informational context for the wider world in which the enterprise functions.

Reference data is also the most widely shared class of data in an enterprise; applications as different as Human Resources and Trade Settlement will need the same state table, postal code table, and currency table. Yet, while reference data is very important for modern enterprises, it is rarely managed well — which has significant associated costs (see *The Costs of Poor Reference Data Management*).

A major reason is lack of clarity about the specific governance and management needs involved. A further reason is that until recently there have been no dedicated tools to help enterprises deal with the large number of specialized tasks and wide scope involved in reference data management. Enterprises have been left to themselves to cope as best they can, using generalized products such as Excel.

As we shall see in this paper, the challenges that need to be addressed to enable effective reference data management cannot be solved with such limited technologies. Equally important are the organizational response to the reference data challenge and the need for an effective methodology.

We will explore all of these themes, focusing on the most important areas of reference data management that an enterprise must address. We shall also identify the problems that can arise if reference data is not managed well. The overall objective is to outline the capabilities that are required to achieve modern reference data management and to provide the enterprise with a foundation to mature its practices.

The Costs of Poor Reference Data Management

The ultimate cost of reference data management problems for a business varies widely. In financial services, and elsewhere, a great deal of back office staff is dedicated to correcting problems that have their origin in mismanagement of reference data. In almost every enterprise, analytics are hampered by misunderstandings about reference data that lead to unsafe results that cannot be used for decision support.

Enterprises that rely on large-scale data entry, such as the healthcare industry, find an abundance of data quality problems due to "miscodings" of reference data. So, while each individual reference data issue may seem insignificant, in the aggregate they are very costly.

But it is not just errors in reference data that can be costly. Immature management practices can be incredibly inefficient. Consider a country code table in a medium-sized enterprise: such an enterprise may have 100 different applications, each with a country code table. Suppose it takes one hour every three months for each team that uses each application to check if the country code table is up to date. For the entire enterprise, this adds up to 400 hours per annum.

Now suppose there is an average of 20 reference data tables per application, and the average time to check each one is the same as for the country code table. The enterprise will spend some 8,000 hours per year checking whether its reference data tables are fully updated. This is the equivalent of roughly four full-time staff. Additionally, checking that a table is up to date is only one task out of many in reference data management.

Given that there are not enough resources in any enterprise to do this, something has to give, and what usually happens is that many of the necessary tasks of reference data management are simply not done.

What is Reference Data?

Many define reference data as "codes", "lookup tables", "domains", or "static data", but it can be formally defined as follows:

Reference data is any kind of data that is used solely to categorize other data found in a database, or solely for relating data in a database to information beyond the boundaries of the enterprise.

Enterprise applications typically implement reference data as database tables that have just a couple of columns — a code and a description — and which contain a few hundred rows at most and change slowly over time.

Figure 1 shows an example of the beginning of a typical code table, the UN Country List. Applications in multiple domains use such standard country codes to categorize other data — for instance, to indicate the location of a business office or customer address.

Because of its perceived structural simplicity, relatively low volume, and slow rate of change, reference data is often overlooked. On the other side of the ledger, however, are these facts:

- Anywhere from 20% to 50% of the tables in a database are reference data tables.
- Any data quality issue in reference data can have widespread results, such as errors in reporting and data integration.
- Tables covering the same or similar reference data get widely duplicated across many applications.

While in the context of a single application, the implementation, maintenance and use of reference data are fairly simple, in the broader context of an enterprise they are complex. *Figure* 2 illustrates some of these facts.

Numerical Code	Country or Area Name	ISO ALPHA-3 CODE
004	Afghanistan	AFG
248	Åland Islands	ALA
008	Albania	ALB
012	Algeria	DZA
016	American Samoa	ASM
020	Andorra	AND
024	Angola	AGO
660	Anguilla	AIA
028	Antigua and Barbuda	ATG
032	Argentina	ARG
051	Armenia	ARM
533	Aruba	ABW
036	Australia	AUS

Figure 1: Example of Reference Data — Fragment of UN Country List *(http://unstats.un.org/unsd/methods/m49/m49alpha.htm)*

What is Reference Data? *(continued from page 3)*

In *Figure 2*, the fragment of the Customer record shown has six reference data fields. These require six tables in each system that has a Customer table. Typically, due to constraints on speed of delivery or development resources, not every system implements all the tables, but instead create custom shortcuts in coding.

Tables implemented by different systems may also contain different code values for the same business concepts. In *Figure 2*, the Gender table may be implemented differently in each system, where ideally each table should code for "Male", "Female", and "Unknown" in the same way.

Thus, even in this simple example we see several key challenges:

- Many tables are needed to represent reference data.
- These tables must be implemented in many different systems.
- Discrepancies can easily arise in reference data across systems.

Figure 3 shows how reference data compares to other kinds of data found in databases. Each kind of data has its own particular characteristics and governance and management needs.

We will now look at what these characteristics and needs are for reference data, starting with what makes reference data important.

Why is Reference Data Management Important?

Enterprises that manage reference data well gain significant benefits.

The benefits include:

- ***Agile responsiveness to new data requirements:*** The ability to implement a new business concept in an application without database restructuring — for example, adding a new code for a new Customer Type, as opposed to changing the Customer table to add a new column.
- ***Description of the external world:*** The ability to easily capture data about the world outside the enterprise, even expected future changes. In one example, many enterprises populated their Currency tables with a code for the Euro years before they started to use it in transactions.
- ***Aid to analytics:*** Reference data provides a fast way to categorize data, even for very short term needs, such as when a six-week marketing campaign for a charity needs to classify donors according to their likely attitudes towards a particular cause.

Conversely, organizations that fail to effectively manage reference data face critical operational risks.

The risks and costs of poor reference data management include: (see also on page 2: *The Costs of Poor Reference Data Management*)

- ***Coding errors:*** If reference data is misunderstood, data entry operators can make "coding errors." For example, if a data entry operator onboarding an institutional customer does not understand the difference between "HF-Hedge Fund" and "AM-Asset Manager" they can invoke the wrong compliance checks which the prospective customer cannot possibly comply with.
- ***Miscommunications across enterprise systems:*** If different systems do not share the same reference data, they cannot communicate effectively. In Figure 2, System 1 may send System 2 records with Gender having values of "M", "F", or "9", where System 2 expects either "1" or "2". Such problems often lead to transaction rejection.
- ***High costs and errors in data integration:*** Differences in implementation of reference data create the need for "mappings" in systems that integrate data such as data warehouses. If System 1 and System 2 in Figure 2 both feed into a data warehouse, the warehouse must "map" the Gender Codes between the two sources. Perhaps "1" is equivalent to "M" and "2" to "F", but the problem still remains of what to do with the "9" from System 1.

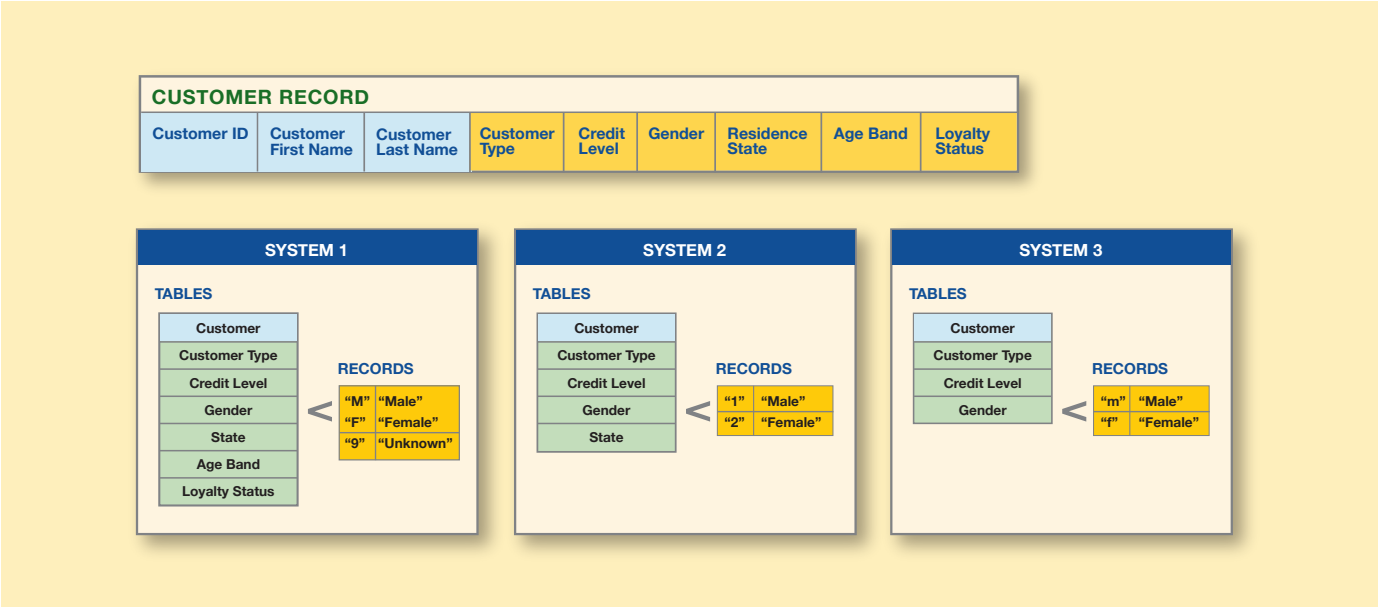

Figure 2: A Typical Reference Data Scenario

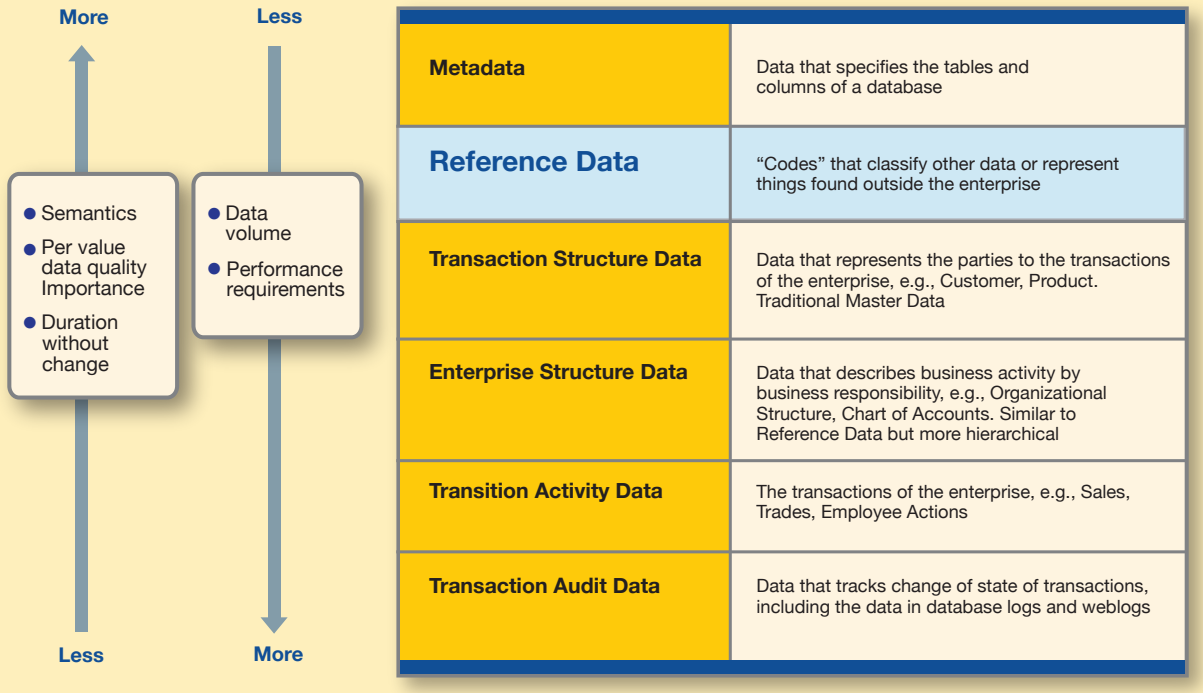

Figure 3: Putting Reference Data in Context

Meeting the Challenges of Reference Data Management

Once an enterprise understands the purpose of reference data and decides that it wants to effectively govern and manage their reference data, what does it do next? Here we meet the problem of immature practices head on. There are no pre-existing templates into which reference data management can be fitted.

It might seem that either Information technology (IT) or Operations could do this work, but it is a poor fit for both. IT tends to think only in terms of time-bound projects that deliver something to business users, after which IT moves on to the next project. Operations, typically the business focus of data management, is inherently tactical and suffers from its goals being oriented to quantity and timeliness, rather than quality.

Given that neither IT nor Operations are natural homes for reference data management, executive management needs to provide leadership. However, doing this requires a vision of how an enterprise can achieve effective reference data governance and management. Such a vision must account for organizational, process and infrastructure needs in order to provide an integrated solution that meets the challenges of reference data management.

Today, many enterprises have a Chief Data Officer (CDO) who can provide the leadership to implement this vision, but if not, this leadership can also come from a Chief Information Officer (CIO) or Chief Operations Officer (COO).

What are the best practices for RDM, and what capabilities must be present in the RDM solutions to effectively support these best practices?

BEST PRACTICES

Best Practices for Managing Reference Data (see page 8) summarizes the best practices that are essential components of any vision for the governance and management of reference data. We will now consider each of these in further detail.

Establishing a Central Reference Data Unit (RDU)

Today, the best practice is to establish a central reference data unit (RDU) that governs both internal and external reference data and plays a strong role in the management of external reference data. (We will discuss the differences between external and internal reference data later.) This unit is staffed with personnel who have expertise in reference data management and governance.

The central RDU should be responsible for creating all the policies needed for reference data management and ensuring compliance with them. It also ensures that appropriate reference data standards are chosen for the enterprise, and it deals with regulatory issues where these exist such as in healthcare and finance. Sometimes, the central RDU may also undertake a certain amount of management of external reference data.

Locating the Central Reference Data Unit

Important stakeholders should agree upon a charter for the central RDU that reflects the overall vision. The charter should state the benefits that the central RDU will provide to the enterprise, which will help with judging its success in the long term. It should also define the scope of the unit, which will help prevent arguments about where the central RDU has authority and where it does not.

The next question is where the RDU should be located. A stand-alone RDU is most advisable, and is indispensable in large global enterprises. Ideally, such an RDU will be located close to similar functions. These could be Data Governance (usually located in the Office of the CDO) or Master Data Management (often in Operations). It is less advisable to have a central

RDU in an analytics environment, since this is too far downstream in terms of data flow to be effective. Likewise, housing a central RDU in IT is not a good idea, because IT is often shunned by the rest of the enterprise, particularly Operations.

If a suitable organizational location for the RDU cannot be found, then an alternative is a highly federated model. As we shall see below, some degree of federation will always be needed. However, in a highly federated model, a group such as Data Governance takes responsibility only for creating policies and standards and ensuring compliance with these. The actual reference data management tasks get divided up amongst the groups most willing and best able to take them on. If Data Governance is chosen for the governance aspects of reference data, then Data Governance must not undertake operational reference data management tasks, particularly for external reference data, as this would create a conflict of interest with their role-setting policies.

Each enterprise must work out how it can implement a central RDU or a highly federated model. *Figure 4* illustrates how one such arrangement might work.

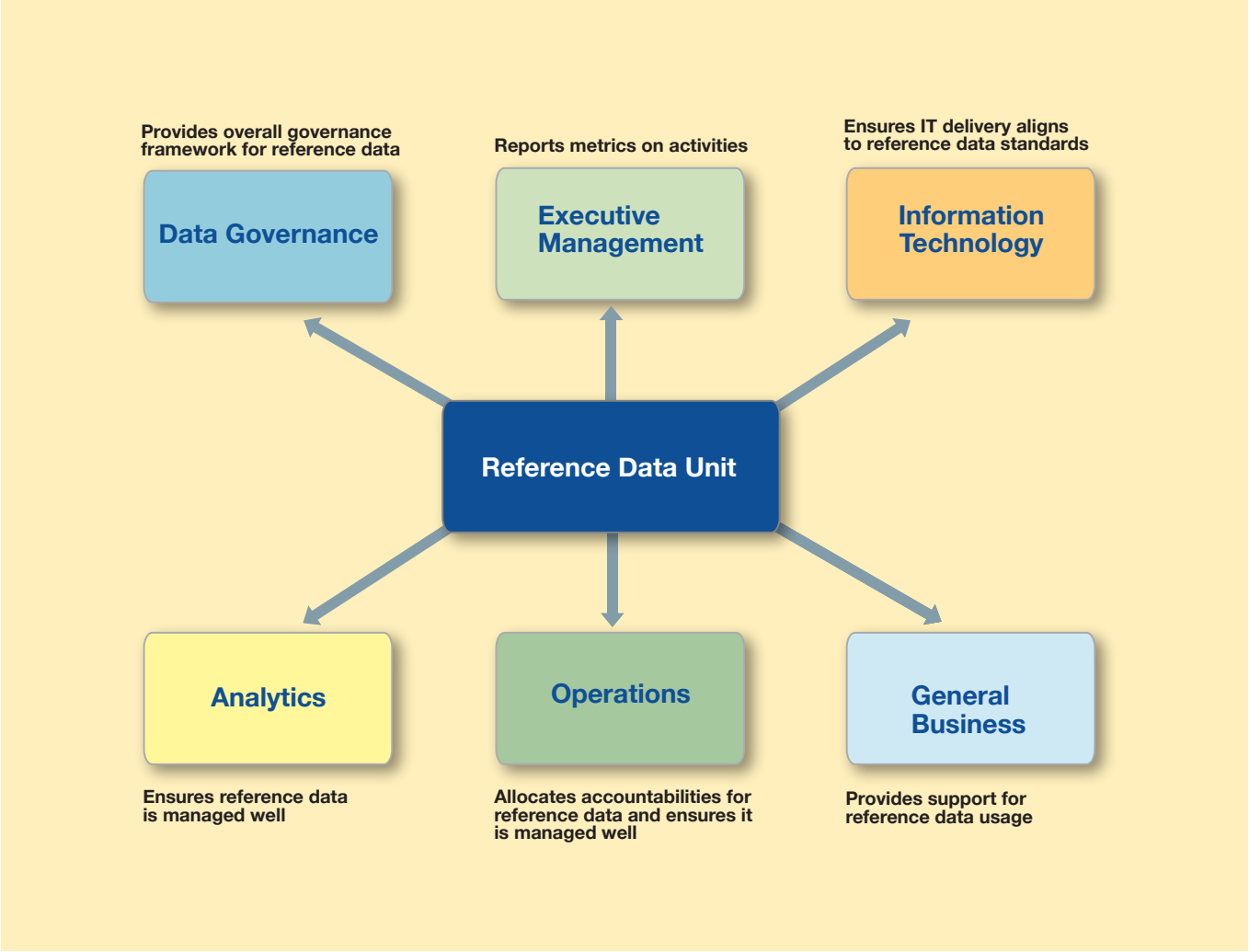

Figure 4: High-Level View of How a Central Reference Data Unit Can Interface with the Enterprise

External Reference Data Management

Once accountability for reference data governance has been assigned, the focus must shift to management practices. A good place to start is with external reference data: reference data that is created and maintained by an authority outside the enterprise such as ISO 3166-1 Country Codes, NACE Codes, and SIC Codes.

A sample of the tasks to carry out includes:

- Discovery of external standards that exist for a given business concept, such as Country names.
- Creating a profile of the external standards authority for use in interacting with the external authority and assessing its reliability.
- Creating a profile of the reference data set maintained by the authority.
- Deciding which of the external reference data sets for a given business concept *(for example, Country names)* will be adopted by the enterprise to represent the business concept.
- Semantic analysis of the chosen reference dataset.
- Documentation of the semantics of the chosen reference dataset *(especially in some kind of tool or repository)*.

Best Practices for Managing Reference Data

1. ***Establishing a Central Reference Data Unit (RDU).*** A Central RDU oversees reference data management across the enterprise to achieve overall goals — especially standardization, quality, and operational efficiency.
2. ***Locating the Central Reference Data Unit.*** It needs to be decided where in the organization a central RDU is located. Ideally it will be close to similar functions such as Data Governance and Master Data Management, perhaps in the office of the Chief Data Officer (CDO). It is less desirable to locate it in IT or in areas responsible for Business Intelligence.
3. ***Managing External Reference Data.*** External reference data is maintained by authorities outside the enterprise. It needs to be discovered, selected, understood and ingested. Standard practices are a major help in doing this.
4. ***Managing Subscriptions for External Reference Data.*** Once external reference data has been set up, it needs to be kept current. Subscription management does this by ensuring that changes are detected and assimilated as rapidly as possible.
5. ***Governing Internal Reference Data.*** Internal reference data is for business concepts that are completely specific to the enterprise. It requires a federated approach, because it is created and managed by many different subject matter experts (SMEs). The central RDU must ensure that groups accountable for internal reference data use a standardized approach.
6. ***Governing Reference Data in Operational Environments.*** Operational units are challenged by changes to the business that often require rapid changes to reference data in application systems. This can create discrepancies and inconsistencies, and the central RDU must find ways to deal with local needs for change without creating difficulties at the enterprise level.
7. ***Distribution of Reference Data.*** Reference data is used widely throughout the enterprise. It is vital that all applications have synchronized copies, so distribution must be addressed. This requires a variety of approaches ranging from the fully automated to the fully manual. However, these approaches must be chosen carefully to maintain operational efficiency.

- Assigning responsibility for the ingest of the chosen reference dataset.
- Executing the ingest of the chosen reference dataset.
- Checking the ingest of the chosen reference dataset.
- Establishing a means by which the rest of the enterprise can access the chosen reference dataset.
- Communicating the availability of the chosen reference dataset to the rest of the enterprise.

Onboarding the chosen reference dataset, which includes:

- Setting up the environment to house the chosen reference dataset.
- Setting up the mechanism to ingest the chosen reference dataset.
- Deciding what to filter out of the chosen reference dataset.
- Deciding what transformations to apply to the chosen reference dataset.
- Deciding how to enrich the chosen reference data set.
- Establishing criteria for testing the success of the ingest of the chosen reference dataset.

This list is not exhaustive, but does illustrate the care needed for successfully ingesting an external reference dataset. It also gives a sense of the extensive need for reference data metadata to manage all of these tasks for maintenance, auditing, and other purposes.

In the past, many enterprises did not govern this work, with the result that individual application development teams simply did it for themselves on an as-needed basis, often using generalized tools like Excel.

Such efforts typically involved just enough effort for the reference data to be "good enough" for the particular application involved and did not include an enterprise perspective.

Small wonder, then, that multiple standards are implemented in many enterprises for concepts such as Country or Industrial Sector, leading to misalignments and errors when data must be shared or integrated.

External Reference Data Subscription Management

After the initial ingest of an external reference dataset, it needs ongoing maintenance. External reference data may change slowly, but it does change. For instance, country codes change an average of 3-5 times per year, and many more in some years. Currency codes change at an average rate of 5 to 10 times per year, again with some years seeing significantly more.

However, this maintenance is rarely dealt with in the absence of a central RDU. If there is no central RDU, an application development team may well be forced to perform an initial ingest of a reference dataset just to be able to test an application they are building. However, once the application is in production the IT team rarely worries about updating the reference data — they see that as a task for the users.

Business users — usually Operations staff — typically lack the time and understanding to track changes in an external reference data standard. Even where they can, it is often isolated teams that make their own decisions about what changes should be included in an application, when this should be done, and how it should be done.

Such lack of governance means that individual applications inevitably drift away from synchronization with the external standard, and from each other. As a result, no matter how well an individual application functions, the data it produces will become increasingly difficult to share, integrate, and understand outside of the context of the application itself.

A central RDU can ensure that subscriptions are established with external reference data authorities. Very often these subscriptions are free, such as a free newsletter of changes. In other cases, the subscriptions cost money, and some may be quite expensive. Sometimes, no subscription is available, and no one can determine if a change in an external reference data standard has occurred except by periodically examining the actual data maintained by the external authority.

It makes sense to centralize subscriptions, for both operational efficiency and reduction of overall

External Reference Data Subscription Management
(continued from page 9)

subscription costs. If you adopt a more federated model, the central RDU can oversee the management of subscriptions by other units. However, a single technical environment that houses subscription information is highly desirable. Modern reference data management tools are beginning to support this requirement, especially the reference data metadata requirements involved.

A more significant element of the work to be done is what happens when a change is announced in a subscription. This must be carefully processed in order to be ready for adoption by the relevant applications in the enterprise. There will be metadata and documentation that needs to be updated to ensure that business users have the correct and up-to-date information on how to properly use the reference data.

However, besides the actual detection of the change in a timely manner, the greatest challenge is to distribute the reference data to the rest of the enterprise, which we will discuss in more detail below.

Governing Internal Reference Data

Internal reference data is reference data for which no external authority exists and which must be managed entirely within the enterprise. Typically, enterprises have their own Customer Types, Product Lines, and so on. Certain groups produce more of this type of reference data than others. For instance, Marketing is always looking at new ways of "segmenting" customers, products, and markets. Such classifications may only apply to short-lived marketing campaigns, but they should still be well-managed.

A central RDU is much less likely to manage internal reference data, because it is generated by SMEs within the enterprise who understand the business concepts involved. The central RDU must concentrate on good governance for the development of reference data content, which is not the case for external reference data, where the RDU has no influence over the external authority.

Examples of the tasks involved in good governance of internal reference data include:

- Ensuring that each business concept represented in reference data has formal assigned accountabilities, and that these are not duplicated.
- Ensuring that the accountabilities for internal reference data management are known across the enterprise.
- Ensuring that each internal reference dataset is semantically analyzed.
- Ensuring that the semantics of each internal reference dataset are documented in a standardized manner that is accessible across the enterprise.
- Ensuring that the content of each internal reference dataset is of the highest quality and does not contain defects.
- Ensuring that internal reference datasets that have relationships or dependencies are harmonized, especially in terms of update cycles.
- Ensuring the effective distribution of updates to internal reference datasets across the enterprise.
- Ensuring that each internal reference dataset is periodically reviewed for relevance and kept up to date with business changes.
- Ensuring that information on material changes to the enterprise are channeled to the accountable parties for internal reference datasets so that these can be updated in a proactive manner.
- Ensuring that accountable parties for internal reference datasets provide adequate support for these datasets.

As we can see, governance of internal reference data inevitably requires a federated model, whereas governance of external reference data can, if desired and practical, be more centralized.

There is no doubt that federated environments require greater governance skills than centralized ones, because many parties must work in harmony and adopt standardized ways of working. This is why the option for Data Governance to take on these tasks, rather than a central RDU, is a real one. However, this must be balanced with the need for specialized domain

Combining Methodology with Adequate Infrastructure is Key to Effective Management

Having a functioning central reference data unit (RDU) means that a single overall methodology can be developed for all aspects of reference data management, thereby improving efficiency and raising reference data quality.

Accountabilities can be distributed in a formal manner so that everyone knows just what they have to do and whom they must communicate with. Most importantly, enterprise-level infrastructure can be built to support this aspect of reference data management and integrated with the methodology.

This is particularly important today as more and more reference data tools are appearing that provide functionality that is difficult for enterprises to develop by themselves. By doing this, the enterprise can achieve greatly increased efficiency gains.

A sound methodology coupled with adequate tooling will also ensure that the enterprise can be agile and adapt to new demands, because all applications use the same set of external reference data, rather than a cacophony of different standards that is wildly expensive and difficult to integrate.

knowledge of what is required for good governance of internal reference data. Take, for instance, the following point in the list on page 10:

Ensuring that the content of each internal reference dataset is of the highest quality and does not contain defects.

This implies that reference data governance teams understand the quality criteria and have adequate tools to address reference data identification, defects, and maintenance. The individuals performing these tasks must have specific domain knowledge about reference data that cannot be substituted for by general knowledge of data management.

So, if a Data Governance unit will govern the reference data environment in an enterprise, it should contain one or more individuals who really understand reference data management and provide adequate tool support to the enterprise for managing reference data issues.

Governing Reference Data in Operational Environments

Staff in operational environments find from time to time that a particular reference data table is no longer adequate for business needs. This is because the enterprise changes in small ways every day, and very often these small changes manifest themselves as the content of a reference data table no longer being adequate for business needs.

Since most applications allow operational staff to update reference data tables, it is quite natural for operational staff to do so when the need arises. Unfortunately, they often do this in a tactical manner that does not consider the wider effects of the change, does not document the change, or does not communicate the change beyond the operational team that performs it. *Figure 5* provides an example.

In *Figure 5* the Credit Level Table originally contained records only for "GOLD", "SILVER", and "BRONZE". It has been overloaded by operations because they need to identify accounts that are suspended for non-payment and, for compliance reasons, accounts that belong to employees. Someone in Operations adds two new records to this table: "SUSPENDED" and

Governing Reference Data In Operational Environments

(continued from page 11)

"EMPLOYEE". John Smith is a non-employee customer who was "GOLD" but has not paid his bill on time, and his Credit Level has been set to "SUSPENDED". If he can clear the non-payment within 30 days he will be set back to his original Credit Level.

The problem is that the prior level cannot be seen in his current Customer record, because "GOLD" has been overwritten by "SUSPENDED". This means that Operations staff must search images of statements previously sent to John Smith to find what his Credit Level was prior to being set to "SUSPENDED". Rather discouragingly, many operations units would prefer to do this than get IT involved to add another column to the Customer record which can house a "SUSPENDED" flag.

This approach to managing reference data also causes problems downstream, where BI reporting will incorrectly classify the John Smith record as not being "GOLD" during the period when he was "SUSPENDED".

Again, however, it is difficult to blame Operations since they are always being pushed for efficiency — meaning reduced costs, greater quantity, and more timeliness. Furthermore, without any reference data governance in an enterprise, how else are Operations staff supposed to behave? To solve this problem, an enterprise needs a central RDU that can establish policies around changing production reference data, including documentation of it, and ensure that all stakeholders are informed of changes.

A central RDU will have an even more subtle challenge to deal with in regard to semantics of reference data in operational environments. This is where Operations staff use a code value in a way that cannot be inferred from its description and which is not documented. Effectively, Operations has changed the definition. This happens quite often. An example outside of reference

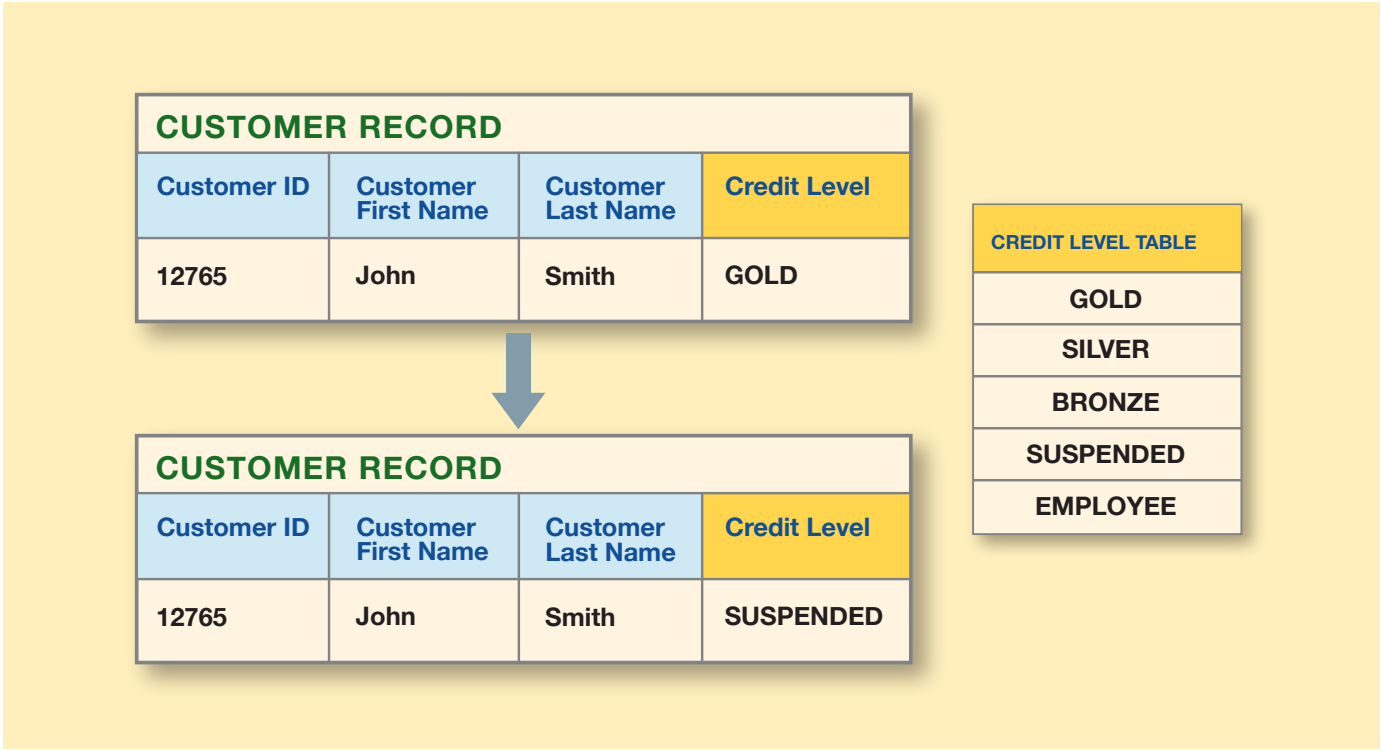

Figure 5: Overloading of a Reference Data Table

Reference Data and Semantics

Reference data resembles metadata in that both have meaning. Very often, however, this meaning is not adequately captured. Codes have descriptions or names associated with them, but they rarely have good definitions. The result is that stakeholders ranging from data entry operators to data scientists in analytic environments often have a poor understanding of what the reference data they work with really means. Since reference data exists to turn other data into business information, this can have a significant downside.

Today, more and more enterprises understand that while reference data may be structurally simple and low in volume, it is a rich and complex semantic environment. As a result, they are devoting more resources to ensuring that the semantics of reference data are properly managed. Reference data management tools are also picking up on this trend.

data is the acronym "ERP", which originally meant "Enterprise Resource Planning", but now refers to a set of integrated applications than can be used to run wide areas of a business — something quite different.

These challenges can be met by a central RDU, but they again illustrate the need for adequate tooling. The scattered Operations teams that can update reference data must take advantage of some kind of common environment if the federated model of reference data governance is to succeed. Only this type of infrastructure can efficiently connect the Operations team with all stakeholders and ensure that they make globally optimized decisions when changing reference data.

Distribution of Reference Data

As enterprises increasingly understand the need for reference data management, they are moving away from letting each siloed application manage its own reference data and instead toward architectures that facilitate the creation and distribution of reference data.

When it comes to the distribution of reference data, a hub model is the default. The hub provides a central location from which all other applications can source reference data.

How do other applications perform the sourcing? Ideally, there will be several methods, including:

- Direct read of reference data from the hub.
- Batch movement of reference data from the hub to the subscribing application.
- Pushing of messages through an Enterprise Service Bus (ESB) to subscribing applications.
- Database replication of tables from the hub to other applications.

Distribution of reference data must typically cope with applications that require reference data to be implemented in technologies ranging from mainframes to Big Data environments. Ultimately, this varied set of technologies will place limits on how much distribution can be automated. This means that the central RDU must also plan to assist with any application that cannot take advantage of automated distribution, both in terms of initially setting up reference data and keeping it up to date.

Distribution of Reference Data *(continued from page 13)*

The latter is the greater challenge. In the end, the central RDU may need to send information about updates to data stewards in each application environment who can then manually process the update.

A hub should also permit periodic reconciliation back to the "single version of the truth" for each reference data table. No matter how reliable the distribution of updates to reference data tables may appear, it must be periodically proven by reconciling each reference data table in each application to the master copy in the reference data hub.

Once again, modern reference data tools support this requirement, although there are limits when it comes to very old, very new, or very exotic technologies. However, where there are exceptions, manual intervention can be used to do the reconciliations.

WHAT CAPABILITIES TO LOOK FOR IN A REFERENCE DATA SOLUTION

We have seen that reference data management and governance have many needs, and it is natural to ask what capabilities any solution that would support these needs should include.

One broad requirement that encompasses all others is the need to manage the wide array of reference data metadata that exists for all the different governance and management needs.

Up to now enterprises have had no choice but to use spreadsheets, or in a few cases, to try to develop their own specialized reference data solutions. As discussed above, these approaches will rarely meet the needs of a modern enterprise.

If an enterprise seriously considers the use of a modern reference data tool, what capabilities should such a tool possess? Some of the most important are as follows:

- ***Ability to create a profile of an external reference data standard.*** This is used to track all interactions with the external authority and assess its reliability.
- ***Ability to create a profile of a reference data set maintained by an external authority.*** Metadata is required to describe the dataset and each element in it.
- ***Ability to perform semantic analysis of each element in the dataset and identify the business concepts that it maps to.*** This requires much metadata and may involve managing decisions by appropriate SMEs and stakeholders about what these mapping decisions are, all of which generates even more reference data metadata that needs to be captured.
- ***Ability to properly document the semantic analysis after it has been performed.*** This may include facts about the reference dataset or individual codes. Such facts help users of the reference data understand how to interpret and use it.
- ***Ability to import external or internal reference data into a central repository.*** Such import must include capabilities for extraction, filtering, transformation, and enrichment. As much as possible, this should be metadata driven.
- ***Ability to assign accountabilities for all aspects of reference data management per reference dataset, particularly for internal reference datasets.*** This achieves the federated governance model needed for internal reference data. Obviously, this capability requires a rich set of reference data metadata elements.
- ***Ability to track changes to reference data***—for example, if an external reference dataset changes.
- ***Ability to distribute reference data.*** A variety of distribution mechanisms such as exports, web services and ESB integration should be provided.

This list only includes the most common and most important capabilities. As enterprises differ in their particular reference data needs, they will require additional, more specific capabilities. However, the above list is fundamental for enterprises planning to stand up a reference data solution.

CONCLUSION

The different nature and unique challenges of reference data are increasingly being understood by enterprises, many of which are moving to better govern and manage this vital resource. There is no doubt that this movement still has some way to go, but the appearance of a new class of enterprise tools specifically targeted to reference data means that there is now a solid foundation on which to build. Enterprises are at last in an environment where they can quickly mature their reference data practices in the next few years.

About the Author

Malcolm Chisholm has over 25 years experience in data management, and has worked in a variety of sectors, including finance, insurance, manufacturing, government, defense and intelligence, pharmaceuticals, and retail. He is an independent consultant specializing in data governance, master/reference data management, metadata engineering, business rules management/execution, data architecture and design, and the organization of Enterprise Information Management. Malcolm is a well-known presenter at conferences in the US and Europe, writes columns in trade journals, and has authored the books: Managing Reference Data in Enterprise Databases; How to Build a Business Rules Engine; and Definitions in Information Management. In 2011, Malcolm was presented with the prestigious DAMA International Professional Achievement Award for contributions to Master Data Management. He holds an M.A. from the University of Oxford and a Ph.D. from the University of Bristol, and can be contacted at mchisholm@refdataportal.com.

About TopQuadrant

TopQuadrant's standards-based solutions enable organizations to evolve their information infrastructure into a semantic ecosystem, the foundation for intelligent business capabilities and integrated big data. As a result, data can be organized, shared and exchanged regardless of its structure, origin or location. TopBraid Reference Data Manager™ supports the governance and provisioning of reference data, including the enrichment of reference datasets (code lists) with comprehensive metadata. TopBraid Enterprise Vocabulary Net™ supports collaborative management of enterprise metadata, models, business glossaries and taxonomies used in search, content navigation and data integration. TopBraid Insight™ is a semantic virtual data warehouse that enables federated querying of data across diverse data sources as if they were in one place. TopQuadrant customers include many government agencies and Fortune 1000 companies in numerous industries including pharmaceutical, financial services, energy and digital media. For more information, visit *topquadrant.com*, or contact us at rdm-info@topquadrant.com, or 919-300-7945.