

Considering a Services Approach for Data Quality

Standardize Data Quality Capabilities for Increased Efficiency
and Lower Overall Cost

WHITE PAPER:

DATA QUALITY & DATA INTEGRATION

David Loshin • President
Knowledge Integrity, Inc.



Considering a Services Approach for Data Quality

Standardize Data Quality Capabilities for Increased Efficiency and Lower Overall Costs

ABSTRACT

WHEN THERE ARE NEEDS FOR ACCURATE, COMPLETE, AND CONSISTENT DATA ACROSS MANY ORGANIZATIONAL USES, A SERVICES APPROACH TO IMPLEMENTING DATA QUALITY SHOULD BE CONSIDERED. USING A SINGLE TOOL SET AND COMMON IMPLEMENTATIONS REDUCES REDUNDANCY, DUPLICATED EFFORT, AND DUPLICATED RULES. SELECTING A SINGLE VENDOR SIMPLIFIES THE LICENSING, TRAINING, AND MAINTENANCE. IN ADDITION THERE ARE BENEFITS FOR STANDARDIZING DATA QUALITY CAPABILITIES ESPECIALLY WHEN ENCAPSULATED WITHIN A SET OF WELL-DEFINED SERVICES THAT REDUCE EFFORT, FUNCTIONAL REPLICATION, AND INCONSISTENCY.

A SINGLE TOOL SET AND COMMON IMPLEMENTATION REDUCES REDUNDANCY, DUPLICATED EFFORT AND DUPLICATED RULES

Between enterprise initiatives such as customer relationship management, enterprise resource planning, and a variety of other horizontal business applications, the demand for data sharing and reuse has never been higher. Yet even though data sets may be suitable for their original purposes, those same data sets may not necessarily meet the needs of the evolving secondary data usage scenarios.

Typically, the business information requirements of these alternate data consumers are not considered when the “origin” applications are being designed and built. The application of origin is often engineered to suit the functional needs of a specific, often isolated business process, and the expectations are guided by the immediate beneficiaries of the application. As a result, each downstream consumer is relegated to extracting and copying the original source data and is forced to define their own expectations and rules for ensuring that the data meets their business application needs.

Not only that, with an increasing number of business applications relying on transactional information and reference data from across different operational silos to compete their operations, there is growing pressure for immediate access to data that exhibits a predictable level of quality. There is little consideration for integration and interoperability between business applications, and the interdependencies between business processes are obscured from any enterprise-wide vision for a business information architecture.

Shortened durations for expectation of timeliness of delivery, an explosion in availability of information from a variety of web channels and cloud implementations, the need for integrating predictive results within operational environments, and reliance on a network of interoperating business environments all lead to the same conclusion: The traditional approaches to data quality are inefficient. Organizations must consider ways to streamline the rapid

provision of data while satisfying the higher expectations for data quality.

Replicated Data Quality Functionality is Inefficient

Because there is a need for accurate, complete, and consistent data across many organizational uses, and typically, each downstream consumer ends up implementing their own data integration and cleansing tactics. And while many of the downstream consumer data integration and quality tools and rules are similar, if not identical, the absence of governance over the deployment of data integration and data quality capabilities leads to significant inefficiencies, including:

- Duplication of effort,
- Multiple tool licenses,
- Duplicated reference data sets,
- Replicated implementations and functionality, and
- Inconsistently-applied transformation and cleansing rules.

Considering Data Quality Services

There are ways that these high-cost inconsistencies can be addressed. Instead of looking at data correction and cleansing as a downstream responsibility, data transformation and data quality techniques can be encapsulated as services with various levels of parameterization as a way of standardizing data services. This article considers these issues, and then proposes that defining standards for technology development and selection, implementations, reference data integration, and data quality rules encapsulated as data quality services will reduce overall operating costs associated with data quality technology while improving consistency and predictability resulting from encapsulation and publication using a services-oriented approach.

Considering a Services Approach for Data Quality

Standardize Data Quality Capabilities for Increased Efficiency and Lower Overall Costs

4

Standardizing Data Quality Techniques

In many environments, because of the constant reinterpretation of data semantics, the downstream data consumers take on the responsibility for data extraction, transformation, and cleansing. Yet the fact that the same (or extremely similar) data quality activities are performed across a wide swath of business applications suggests that resources are wasted in designing and implementing data transformation and cleansing – there is bound to be significant duplication of effort, leading to redundancy and inefficiency. Standardizing the methods for transformation and cleaning reduces that replicated effort.

In larger organizations, there may not even be an awareness of the diffuse use of similar tools and technologies, whether they come from the same tools vendor, or from a variety of vendors. This complicates technology management, especially when attempting to manage and optimize vendor relationships. Simplifying the relationship with tools vendors and standardizing the use of vetted tools allows the organization to negotiate more appealing licensing terms, while decreasing the variety of implementations, hardware requirements, the need for training, and maintenance arrangements associated with multiple tools.

Organically-developed applications often rely on their own defined reference data sets. In the absence of enterprise standards, those reference data sets might be defined internally, copied from any of a multitude of external sources, or acquired directly from data aggregators. This diverse use of data sets that should map to the same conceptual domains not only confuses reporting, but also leads to continuous reconciliations and reviews. When common conceptual data domains and their mapped value domains are consolidated, access can be exposed through a reference/metadata service. This reduces storage demands, as well as the need to reconcile reports with different representations for the same reference data values.

Lastly, while the types of data quality and cleansing activities performed are often the same, as are the types of rules that are applied. However, in many environments, different rules are applied, and even in environments where the same rules are applied, they may be applied in different orders, leading to different results.

Clearly, these issues suggest that there are benefits for standardizing data quality capabilities especially when encapsulated within a set of well-defined services that reduce effort, functional replication, and inconsistency. In order to standardize these capabilities as services, though, one must first take inventory of the data integration, transformation, and cleansing activities being performed across the application infrastructure and mapping those activities to recognized services:

1. Within each business application, document any use of a data source intended for an alternate primary purpose.
2. What data extraction is performed?
3. What data transformations? Are any transformations related to differences in data structures?
4. Are any of the data sets reference data sets?
5. Are data corrections and modifications performed to improve the data sets' usability?
6. Are multiple data sets merged together? What linkage is performed?
7. Are the data sets enhanced in any way?

By performing a survey of data quality techniques used by the collection of secondary data consumers will not only demonstrate the common uses, it will also present opportunities for consolidating and encapsulating capabilities to improve consistency while reducing complexity and duplicate effort.

THERE ARE A NUMBER OF DATA INTEGRATION AND DATA QUALITY CAPABILITIES THAT ARE SUITABLE FOR DEPLOYMENT AS SERVICES

SERVICE	DESCRIPTION
Data Access and Integration	Data integration services enable the perception of seamless connectivity and accessibility to a variety of data platforms and representations across the environment. Coupled with data transformation services, integration services enable data access, extraction, and normalization into formats that are suitable for data sharing. Data federation and virtualization can provide an abstraction layer with standardized data accessibility across a variety of platforms.
Reference Data and Metadata	Commonly-used reference data sets can be consolidated into a single data asset, managed via metadata management tools, to provide a unified and consistent reference values.
Parsing and Standardization	Parsing is a process of defining patterns and using those patterns to identify key tokens within data strings. Standardization uses those patterns to distinguish valid from invalid data values. Valid values are parsed and their component tokens can be recognized and then rearranged into a standard form. Invalid values appearing in common error patterns can be automatically corrected, then standardized.
Identity Resolution	This service provides similarity analysis between sets of records and determination of matching entity records (most often based on weighted approximate matching between a set of attribute values). Identity resolution is used to recognize when only slight variations suggest that different records are connected and where values may be cleansed, or where enough differences between the data suggest that the two records truly represent distinct entities.
Matching and Linkage	Identity resolution finds similar records in different sets. The matching and linkage services establish connectivity between similar records, and may even merge a pair of records into a single surviving record in preparation for data cleansing.
Data Enhancement	Data enhancement builds on parsing, standardization, and record linkage to append additional information to existing records from third-party data sets (such as name standardization, demographic data imports, psychographic data imports, and household list appends). A typical data enhancement activity, address standardization and cleansing, relies on parsing, standardization, and the availability of third-party address information.
Data Cleansing	The data cleansing service builds on the parsing, standardization, and enhancement services, as well as identity resolution and matching & linkage. By parsing the values and triggering off of known error patterns, the data cleansing service can apply transformation rules to impute data values, correct names or addresses, eliminate extraneous and/or meaningless data, and even merge duplicate records.
Geocoding	Geocoding is a location intelligence process for translating a conceptual location to specific coordinates on a map of the earth's surface. A geocoded is a pair of coordinates for the latitude and longitude of the location, and the geographic location (or "geolocation") service supplies a "latlong" coordinate pair when provided with an address.
Data Validation	A data validation service builds on the parsing and data profiling capabilities to proactively verify the validity of data against a set of defined (or discovered) data rules. This data validation service can generate notifications when data instances do not conform to defined data quality rules.

Table 1: Common Data Integration and Data Quality Services

Considering a Services Approach for Data Quality

Standardize Data Quality Capabilities for Increased Efficiency and Lower Overall Costs

6

Data Quality Capabilities

There are a number of data integration and data quality capabilities that are suitable for deployment as services, especially when there are opportunities to consolidate both resources (such as software licenses and hardware) and effort in implementation. Some of those key capabilities are described in Table 1 (see previous page).

Increase Consistency through the Services Approach

The survey of data quality techniques and applications will reveal that the same transformation and cleansing techniques are used, and similar rules are applied in similar contexts, to the same types of data, by different applications. Abstracting, encapsulating, and implementing those capabilities as shared services addresses many of our issues. Using a single tool set and common implementations reduces redundancy, duplicated effort, and duplicated rules. Selecting a single vendor simplifies the licensing, training, and maintenance. A standard reference data service lends clarity to reports and aggregated values as well. These benefits are facilitated through a service-oriented architecture that standardizes data integration and data quality capabilities and exposes them consistently to all data consumers.

However, perhaps the biggest benefits are associated with standardization of application functionality is consistency and predictability. Common services using shared data transformation and cleansing rules will reduce variation in secondary data use. The same transformations can be applied in the same order, with similar, if not identical results. This enables a consistent view of the data no matter where the values are used, which also means reduced need for review, timely reconciliations, and costly rework. When considering approaches to implementing data quality techniques across the organization, a reasonable and beneficial option is a services approach!

David Loshin is the President of Knowledge Integrity, Inc., a consulting and development company focusing on customized information management solutions including information quality solutions consulting, information quality training and business rules solutions. Loshin is the author of *Master Data Management*, *Enterprise Knowledge Management – The Data Quality Approach* and *Business Intelligence – The Savvy Manager’s Guide* and is a frequent speaker on maximizing the value of information. David can be reached at loshin@knowledge-integrity.com or at (301) 754-6350.

UNITED STATES

One Global View
Troy, NY 12180

1.800.327.8627

pbbi.sales@pb.com
www.pbinsight.com

CANADA

26 Wellington Street East
Suite 500
Toronto, ON M5E 1S2

1.800.268.3282

pbbi.canada.sales@pb.com
www.pbinsight.ca

EUROPE/UNITED KINGDOM

Minton Place
Victoria Street
Windsor, Berkshire SL4 1EG

+44.800.840.0001

pbbi.europe@pb.com
www.pbinsight.co.uk

ASIA PACIFIC/AUSTRALIA

Level 7, 1 Elizabeth Plaza
North Sydney NSW 2060

+61.2.9437.6255

pbbi.australia@pb.com
pbbi.singapore@pb.com
pbbi.china@pb.com
www.pbinsight.com.au