TDWI

MONOGRAPH SERIES

NOVEMBER 2007

Complex Data: A New Challenge for Data Integration

By Philip Russom

Senior Manager, TDWI Research
The Data Warehousing Institute

tdwi
THE DATA WAREHOUSING INSTITUTE

Table of Contents

About the Author

PHILIP RUSSOM is the senior manager of TDWI Research at The Data Warehousing Institute, where he oversees many of TDWI's research-oriented publications, services, awards, and events. Prior to joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research, Giga Information Group, and Hurwitz Group. He's also run his own business as an independent industry analyst and BI consultant, and was contributing editor with *Intelligent Enterprise* and *DM Review* magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org.

About Our Sponsor

Introduction to Complex Data

Data integration solutions have grown significantly in number and depth this decade. Today, data integration solutions are commonplace in corporate and governmental organizations, where they continually acquire, merge, and transport data. Integrated data is loaded into target databases and applications inside the organization or packaged into files and documents for exchange with other organizations. Growth has expanded both analytic data integration (which is largely about extract, transform, and load [ETL] for data warehousing) and operational data integration (which mostly involves the consolidation, migration, or synchronization of operational databases).

But there's a catch. Despite the dissemination of data integration practices in recent years, data integration solutions continue to follow older design paradigms that ignore or short-cut key issues.

- **Data quality.** Data integration solutions invariably reveal data quality issues, whether these are problems that require fixing or opportunities for improvement that merit leveraging. Although data integration and data quality are inherently linked, the old bad habit is to implement the two solutions separately—or to ignore data quality entirely. Data integration solutions need to incorporate more data quality functions to assure that integrated data has the highest value and most impact possible.

Complex data and its quality is a new challenge for data integration implementations.

- **Complex data.** Most of the data processed by data integration solutions today comes from lightly structured data sources—typically tabular data from relational database management systems (RDBMSs) and flat files in record format—whereas at least half of an organization's data is not structured in a tabular way. In other words, the benefits that organizations get from integrated data are limited to tabular data from structured sources, whereas similar value and benefits could also be achieved by embracing the complex data found in unstructured, semi-structured, and hierarchical data sources.

- **Data quality for complex data.** When data integration operates on complex data, it reveals data quality problems and opportunities, just as it does with structured data. As data integration practices and technologies expand to embrace complex data, data integration solutions must grapple with two tasks that are new to most data integration specialists: integrating data from complex and nontraditional sources and assuring the quality of data drawn from those sources. This paper explores the combination of these two tasks.

Defining Complex Data

Complex data is any source that does not have a tabular structure.

Before we dive into the problems and opportunities of complex data, let's define some terms:

- **Structured data.** Simple tabular data structures are the best examples, as seen in relational database management systems (RDBMSs), table dumps, and flat files in record format.

- **Unstructured data.** The least amount of structure is found in documents of mostly natural-language text, like word-processing files, e-mail, and text fields from databases and applications. Some documents may have light metadata, such as spreadsheets and RSS feeds.

- **Semi-structured data.** This data has minimal metadata. Data documents exchanged between organizations often combine unstructured and structured data or (when expressed in XML) text that's structured with metadata tags. Semi-structured data documents of this type usually comply with open standards for data exchange, like SWIFT, NACHA, HIPAA, HL7, RosettaNet, and EDI.

- **Complex-structured data.** This includes deeply hierarchical XML structures, as in the XML-based standards MISMO and ACORD. In this TDWI Monograph, the term *complex data* includes unstructured, semi-structured, and complex-structured data.

Complex Data and Analytic Data Integration

Structured data has long dominated the information content of data warehouses. That's because the source data that comes from operational applications and databases is almost exclusively structured. The warehouse itself is ruthlessly structured, with its tables, star schema, and other structured data models. Most reporting and analysis tools demand structured data that is accessible via structured query language (SQL). And data integration routines transform source data into the data structures required by the warehouse and reporting tools.

The average data warehouse is fed mostly from structured data sources.

Alas, little unstructured or semi-structured data makes its way into data warehouses today. To quantify the situation, TDWI ran a survey that asked survey respondents to estimate "the approximate percentages for structured, semi-structured, and unstructured data across your entire organization." (See the top bar in Figure 1.) The survey also asked for "the approximate percentages for structured, semi-structured, and unstructured data feeding into your organization's data warehouse or BI processes." (See the bottom bar in Figure 1.) Although semi- and unstructured data together account for over half of data across an entire organization, less than a quarter of the information in the average data warehouse comes from complex data sources.

Little unstructured or semi-structured data makes its way into data warehouses today.

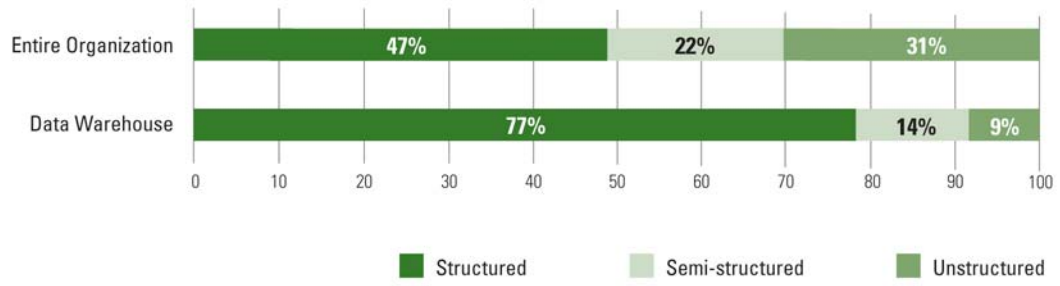

Figure 1. Based on 370 responses to an Internet-based survey that TDWI ran in late 2006.

A data warehouse isn't the whole truth without data from complex sources.

The problem with a narrow focus on structured data—typical of the average data warehouse—is that it excludes valuable information that resides in unstructured, semi-structured, and complex-structured data sources, which in turn hobbles decision making and organizational performance supported by the data warehouse. Conventional wisdom says that a data warehouse should be "a *single* version of the truth," so that all decision makers work from the same information. But a data warehouse is not the *whole* truth without representation from various types of complex data.

Complex Data and Operational Data Integration

Most implementations of operational data integration involve the consolidation, migration, upgrade, or synchronization of operational databases. However, a growing area within operational data integration is inter-organizational data integration, sometimes simply called *data exchange*. This usually takes the form of documents or files containing data that's exchanged between two or more organizations. Depending on how the organizations are related, data exchange may occur between business units of the same company or between companies that are partners in a business-to-business (B2B) relationship.

Centralizing data exchange via a hub reduces complexity and cost.

Many companies find that data exchange costs too much time and money when they create and maintain custom, point-to-point interfaces between:

- Customers and their electronic data interchange (EDI) gateway to take in new orders
- Order validation systems and their customers to send order acknowledgment
- Customers' stores and internal data warehouses to track point-of-sale data (see Figure 2)

Deploying point-to-point interfaces for data exchange results in complexity and cost.

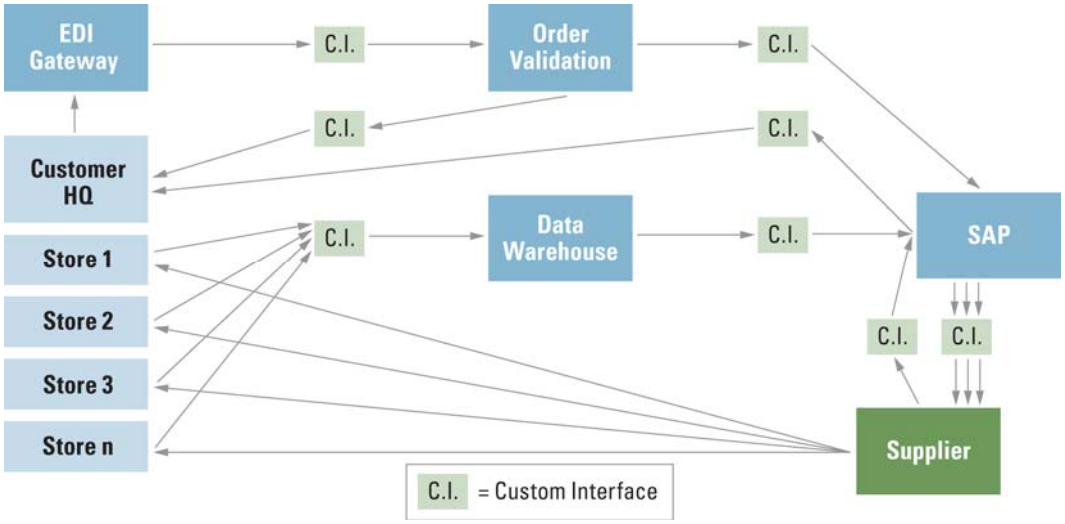

Figure 2. Source: Informatica Corporation.

Companies need a centralized way of linking different organizations and systems without building and maintaining multiple, separate interfaces. (See Figure 3.)

Exchanging data via a central hub reduces complexity and cost.

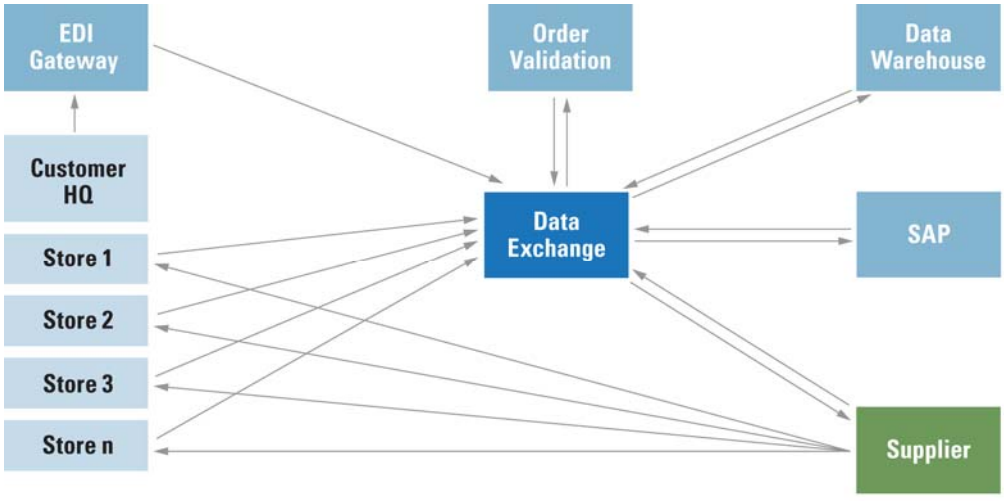

Figure 3. Source: Informatica Corporation.

With centralized data exchange, companies can achieve the following goals:

- Reduce high development and maintenance costs of point-to-point interfaces
- Increase supply chain and business process efficiency
- Avoid regulatory penalties and loss of data via compliance with standards and their versions

Data exchange is XML's killer app.

Data exchange very often transports complex data tagged with extensible markup language (XML). Although the *X* in XML stands for extensible, it might as well stand for exchange, because the majority of uses of XML involve inter-organizational data exchange. In many ways, data exchange is XML's killer app.

One of the interesting architectural features of data exchange is that it links together organizations that are incapable of communicating directly. That's because data is flowing from the operational and transactional IT systems of one organization to those of another, and the systems of the two are very different in terms of their inherent interfaces and data models. Hence, data exchange almost always requires a third, neutral data model in the middle of the architecture that exists purely for the sake of interorganizational communication and collaboration via data. (See Figure 4.) As an analogy, consider the French language. In the seventeenth, eighteenth, and nineteenth centuries, French was the language of diplomacy—literally a *lingua franca*—through which people from many nations communicated, regardless of their first languages.

Data exchange architecture requires a third, neutral data model in the middle.

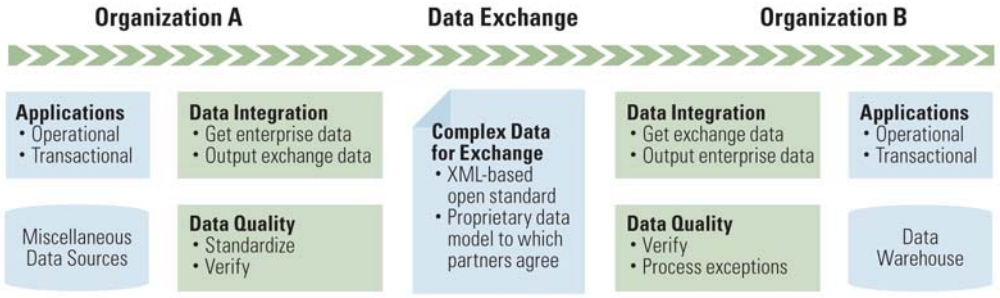

Figure 4. Basic data flow (left to right) for inter-organizational exchange of complex data.

XML-based standards and other complex data provide a lingua franca for data exchange.

For the data model in the middle of a data-exchange architecture to be effective as a lingua franca, all parties must be able to read and write it. This is why open standards are so important to data exchange, including industry-specific standards like ACORD for insurance, HIPAA and HL7 for healthcare, MISMO for mortgages, Rosettanet for manufacturing, SWIFT and NACHA for financial services, and EDI for procurement across industries. From this list, you can see that data exchange is inherently linked to standards that model complex data, in the sense of semi-structured data (as with all XML) and complex structured data (as in ACORD and MISMO). Even so, note that some organizations forgo XML-based open standards in favor of unique, proprietary data models that all parties agree to comply with.

Generating and processing complex data for exchange requires data integration and quality technology.

Implicit in the data exchange architecture shown in Figure 4 is the fact that data integration technology is required to populate XML documents and other datasets for exchange, as well as to parse, convert, cleanse, and process them in the receiving organization. In a recent TDWI survey, 41% of respondents reported using ETL as the main technology for operational data integration, followed by hand coding (28%).[1] Likewise, data quality tools aid in standardizing and verifying outbound complex data, as well as verifying data and handling exceptions on the inbound side. And, finally, real-time operation for data integration and data quality is required when the inter-organizational exchange of complex data defines a transaction or event that requires immediate processing or monitoring.

[1] TDWI ran the survey in February 2006 and February 2007. Statistics cited here combine responses from the two runs.

Why You Should Care about Complex Data

Improving the integration and quality of complex data has several advantages:

Complex data makes analytic applications more complete and accurate.

- **Complex data completes your view of corporate performance.** Users that TDWI has interviewed report having an epiphany after adding information drawn from complex data sources to their data warehouse environment, which was previously fed exclusively from structured data sources. For example, metrics and other aggregated data related to supply chain entities (products, suppliers, distributors) tend to tell one story based on structured data, and another based on complex data. Hence, both are required for an accurate view. Achieving a complete and accurate view of corporate performance is the most compelling reason for expanding data warehouse sources to address the full range of structured, unstructured, semi-structured, and complex-structured data.

- **Centralizing data exchange via a hub reduces complexity and cost.** Each industry has a variety of document formats and industry standards to accomplish its data exchange. Companies must comply with the latest formats and standards in order to maintain their competitive edge. The standards are typically defined by industry groups (or in some cases government organizations) that maintain and evolve the standards. Staying current with such standards forces users with point-to-point interfaces to invest in constant coding and maintenance to avoid regulatory penalties and loss of data. A centralized data exchange solution—built atop a vendor tool that is automatically updating as standards change—helps companies avoid constant recoding and maintenance.

- **Complying with industry standard formats reduces risk.** If your IT systems don't generate and process data in full compliance with industry and regulatory standards, you may violate service-level agreements with partners or breach governmental regulations. Furthermore, applying data exchange standards inconsistently erodes customer confidence and incurs internal costs. Data quality tools help you support mission-critical standards.

As with structured data, integrating complex data surfaces quality issues.

- **Data has quality issues, whether structured or complex.** Decisions based on data-driven facts are supported or tainted by the quality and completeness of data. Operational processes are likewise affected. Conventional wisdom says that data quality measures should be applied to structured data. We all now need to stretch that assumption to encompass data coming from semi-structured, unstructured, and complex-structured sources.

- **Data integration reveals data quality issues, even with complex data.** One of the best practices established this decade is to consistently apply data integration and data quality techniques in the same solution. That's because data integration invariably ferrets out problems and opportunities that data quality tools or functions can address. The tandem application of integration and quality techniques is firmly established for projects involving structured data. It now needs to extend to projects involving complex data.

Data quality for complex data requires special stewardship, which helps reduce business risk.

- **Data quality measures for complex data require special stewardship.** A data steward is someone who understands the business and its processes, and also has enough technical savvy to identify data that requires or would benefit from improvement. Relevant to complex data (especially in a data exchange scenario), a steward may need to handle exceptions that a tool cannot process automatically. For instance, a data steward may need to approve and direct individual transactions or review and correct mappings between ambiguous product descriptions. For this to work, data integration and data quality solutions must support special interactive functions for data stewards or integrate with applications with similar functions.

- **Embracing complex data enables new applications and extends operational integration.** This includes analytic applications based on new views of corporate performance as seen from complex data. And there's a host of operational applications that aren't possible (or aren't as successful) without full integration and quality measures for complex data, including B2B gateways (e.g., applications for online payment or trade reconciliation), internal integration hubs (to assure compliance with HIPAA or to consolidate shared data about customers or products), and legacy renewals (which enable mainframe applications to support open standards).

Applications Involving Complex Data

Complex data—especially semi-structured data exchanged between organizations—drives and enables several types of prominent applications. All these have special requirements for the integration and quality of complex data.

Large applications like B2B gateways, process hubs, and BPO involve many complex data standards, plus versions and variants.

Business-to-Business (B2B) Gateway. Most of these are trading portals of some kind that enable B2B communications with trading partners. Even when partners are using the same open standard, they will use different versions and create slight variations to accommodate unique data requirements. Sometimes a partner demands the use of its proprietary data format as a requirement for doing business. Furthermore, a single organization may have multiple back-end applications tied into the gateway (possibly due to mergers and acquisitions), which generates a lot of internal, cross-application integration of complex data involving multiple standards. Consequently, data integration infrastructure must support open standard data models, versions and variants of the standards, and ad hoc or proprietary data models.

Integration Hub. A hub of this sort seeks to centralize and control data and its usage. Technical reasons for this include making data consistent, clean, and broadly accessible. Business reasons include complying with data usage standards (like HIPAA or internal data governance policies) or to reducing risk (by monitoring transactions or sensitive activities). These hubs usually involve a messaging infrastructure for enterprise application integration (EAI) and/or enterprise service bus (ESB). Some companies have messaging tools from multiple vendors (IBM, TIBCO, WebMethods), and SAP shops are aggressively moving to SAP NetWeaver. In addition, hubs that consolidate data or master data (about customers, products, financials, patients, claims, etc.) may involve large databases that are updated and accessed frequently. To transform, integrate, and standardize complex data broadly in support of an integration hub, data integration infrastructure must interoperate seamlessly with all these messaging tools and databases, increasingly through a service-oriented architecture (SOA).

To handle complex data, data integration tools must interface deeply and in real time with messaging systems.

Business Process Outsourcing (BPO). When complex data exchange is far removed from the core competency of an organization—or simply beyond its technical prowess—it makes sense to outsource the exchange and processing of complex data. For example, many healthcare organizations (physicians, clinics, and hospitals) prefer to outsource payment processing to a third party that provides a transaction hub so they can stay focused on their mandate of serving patients. The challenge for the hub provider is to build a data integration infrastructure that supports healthcare standards (like HL7), plus nonstandard data models arriving in COBOL results sets, Word and PDF files, custom formats, and so on. In turn, the hub provider usually interfaces with healthcare insurance organizations, and each of these will have its own data exchange and processing requirements.

Legacy Modernization. Many organizations have legacy applications (often running on legacy platforms, like mainframes) that provide powerful processing of transactions, claims, payments, and other business entities that may be associated with complex data. The problem is that many of these legacy applications predate modern data exchange standards or interfaces for integration and messaging. The challenge for data integration tools is to modernize these legacies by providing access to the legacy platform or application natively, while also providing transformation services based on open standards and other data schema. When the legacy application involves a hierarchical database (as many mainframe applications do), this adds another form of complex-structured data that a data integration tool must handle.

Don't forget: real-time operation and data quality functions are required of complex data exchange and integration applications.

Real-Time Operation. Some applications that depend on complex data and/or its exchange need to operate in real time (or close to it) to process time-sensitive or content-sensitive information. For example, real-time operation is appropriate to the straight-through processing (STP) seen in trades, transactions, and payments based on complex data exchange. Likewise, monitoring these events in real time helps to detect risk, fraud, and opportunity as early as possible. Other applications where complex data may need real-time processing include mortgage or loan origination (especially when a price quote is presented), real-time access to data consolidated in a hub (as with fetching a patient record via a HIPAA-compliant service), and operational dashboards (largely fed by complex data) in transportation, logistics, and energy utilities.

Data Quality for Complex Data. Let's not forget that data quality functions are inherently part of a complete data integration solution, and this also applies to applications that generate and process complex data. For example, pharmaceutical companies generate mountains of clinical trial data in unstructured formats that need standardization prior to processing. Banks consolidate customer data and merge it with third-party consumer data, typically in flat files; this practice is prone to redundancy, and so it benefits from deduplication in the form of match-and-merge and householding. And industries for which EDI was not designed (like biotech) are plagued by EDI variants; data profiling helps you get up to speed quickly when a new variant arrives.

Data Stewardship and Complex Data

Data stewardship is a best practice that's well-established in corporate initiatives for data quality, data warehousing, customer data integration, and product information management. Although traditional data stewardship is usually applied to structured data, there's a need for special forms of stewardship that address complex data and its unique issues.

Data stewardship can assist with several situations that commonly arise with complex data:

- **Solution design.** Data stewards have traditionally helped discover and prioritize data structures that would benefit from data quality measures, which deeply influences the design of a software-based solution. Stewards are well-positioned to make similar contributions to data quality and data integration projects involving complex data, because they typically know the standards and business processes that govern data exchange better than data integration specialists do.

Stewards are key to accurate mappings and transformations for complex data.

- **Data mappings.** Product catalog data is commonly exchanged among manufacturers, suppliers, retailers, and other partners via a supply chain. Since each partner has a unique way of naming a product and its specifications, accurately mapping a partner's data into your IT systems can be problematic. Given an appropriate tool, a data steward can apply his/her product expertise to reviewing and correcting the mappings. Since most mappings have one or more data transformations embedded, the steward can likewise review and correct transformations.

- **Match and merge.** Complex data exchange often reveals redundant data. For example, product catalog data coming from suppliers is rarely deduplicated, and data about customers and consumers invariably results in conflicting records. For data integration and quality tools to achieve high accuracy in their automation, a steward needs to review exceptions kicked out by the tools, then adjust business rules for match and merge accordingly.

- **Routing and approval.** In many cases, complex data exchange transports high-value transactions, as is the case with standards like SWIFT. In other cases, data transfer must comply with internal policies or regulatory standards like HIPAA, so the exchanged data should be routed through an integration hub for approval. Stewards and other business people who understand the standards, transactions, and policies involved can review these cases, then route or approve them accordingly.

Stewardship is key to effective exceptions processing for complex data.

- **Exceptions processing.** Note that most of these examples involve some kind of *exceptions processing*. In other words, a data integration or data quality solution automatically processes most of the data it handles, but it pushes data it can't process into an exceptions file or queue that requires human intervention—by a steward or other business person—to act on the exception. While it's possible that stewards could handle all exceptions manually, this is obviously not ideal, because it lacks the efficiency of software automation and may be subject to human error or inconsistency. Instead, the work stewards do with exceptions should be captured by a tool and folded into the solution as repeatable business rules, so that over time the solution's automation becomes more accurate and generates fewer exceptions.

Solution Requirements for Exception Processing

To be involved in design, mappings, or manual processing of exceptions, stewards need an appropriate tool, which may take various forms:

Data stewards need their own, specialized tools.

- **Tools or applications for business process management or integration.** When complex data exchange is tied into an integration hub, it may be best (for consistency and central control) for the hub's tools to present exceptions for processing, whether the hub is authorizing transactions and payments, consolidating customer and product data, or integrating master data. In these cases, data integration and quality tools must interface with the process hub.

- **Business-oriented area within a data integration or data quality tool.** Though this area is for data stewards and similar mildly technical users, it must integrate tightly with areas in the tool for technical personnel. As an extension of the development process, this area enables the steward to participate in the definition, testing, and debugging of data mappings and transformations. To accommodate a steward's level of expertise, the area should be an easy-to-use GUI, with simple drag-and-drop or point-and-click functionality, so the steward can mark up and map data items to and from XML-based standards and other data-exchange schema.

- **Exceptions presented for processing in Excel or a Web GUI.** Sometimes the stewards and business analysts involved in exceptions processing aren't open to learning another tool. As a result, it can be critical to give them a way to create and correct integration and quality design in a tool they are already familiar with. Microsoft Excel or a Web-based GUI are very comfortable and productive tool environments for most business users, so they're natural media for presenting exceptions that need to be processed by a business-savvy decision maker.

Reviewing exceptions in Microsoft Excel or a Web GUI are good options for some data stewards.

As an example of an exceptions processing application appropriate to data stewards and business analysts, consider the GUI in Figure 5. This straightforward application, which requires no training, presents in Excel customer records that are matches. The records are exceptions from a data integration or quality process. A tool generated the Excel file and can process it later after the user has marked it up. The user simply clicks on combo boxes in an Excel document to select the most complete, accurate, or standard fields for name, address, social security number, date of birth, and phone number. The application then assembles the selections into a merged record.

A GUI for stewards enables them to perform design or quality tasks quickly and accurately.

Company Name	Address	City	State	Zip	Country
Abbot Laboratories	100 Abbot Park Road	Abbotsville	NC	22383	USA
Abbot Laboratories	100 Abbot Park Road		NC	22383	
Abbot Labs	100 Abot Road	Abbotsville		22383	United States
Abbot Labs	100 Abbot Park Road	Abbotsville	NC	22382	USA
Abot Labs			NC	22383	
AbLabs	Abbot Park Road			22382	United States

Figure 5. Source: Informatica Corporation.

Similar applications would enable a data steward to process exceptions by mapping product catalog records to internal product names, by selecting an appropriate master data definition, or by assigning an appropriate workflow process for the data to be routed through.

Recommendations

Complex data gives analytic applications greater accuracy and enables new operational applications.

- **Represent more complex data in your data warehouse.** Three-quarters of data in the average data warehouse comes from structured sources, although only half of data across the rest of the enterprise is structured. Most organizations need to close this gap; otherwise, the data warehouse will remain a single version of the truth, but not the whole truth.
- **Embrace complex data and its exchange to enable new applications.** Some of the examples mentioned in this monograph include various types of B2B gateways, integration hubs, business process outsourcing, legacy modernization, and real-time monitoring of time-sensitive or risky business events. To reap the full benefits of these, most organizations will need to substantially revamp their data integration infrastructure.
- **Revamp your data integration infrastructure and extend your operational integration.** You designed most (or all) of it to integrate structured data. You must maintain what you've built, but deploy new interfaces to unstructured, semi-structured, and complex-structured data, which data integration tools can provide. Furthermore, your data integration implementation is probably mostly or exclusively operating in batch mode, whereas you need to add capabilities for real-time operation. Furthermore, it may not support messaging and services. These additions to data integration infrastructure are prerequisites to a successful data integration or data exchange application.

Success with complex data demands additions to data integration infrastructure, especially support for data quality and data exchange standards.

- **Coordinate data quality with data integration and vice versa.** Data integration inexorably ferrets out data quality problems, and data quality is hamstrung without help accessing and joining diverse data. Expect to exercise data integration and quality practices in tandem, regardless of whether the data being handled is structured or complex.
- **Profile source data carefully, or suffer the consequences.** Data profiling gives data integration and data quality solutions higher-quality data deliverables, more accurate project scope, and a reduction of "gotchas" that pop up in testing and deployment. Ample profiling is critical to successful complex data exchange, because it helps you get up to speed quickly with new data schema (whether based on standards or not) and helps you spot variations in how your partners apply data standards.
- **Depend on ETL for complex data integration.** Most organizations do, because the T in ETL is very useful in transforming information from its original semi-structured and unstructured data formats to data models conducive to business intelligence or to complex data exchange. For an ETL tool to contribute this way, it needs to support a long list of XML-based standards and be able to transform among these in real time, as well as batch.

Choose vendor tools that support multiple data management functions for both technical and business people.

- **Select data management tools that interoperate deeply and easily.** A full-blown data integration infrastructure will include design environments and deployable servers for data integration, data profiling, data quality, legacy data access, message queues and brokers, open-standards transformation, both batch and real-time operation, and so on. To reduce the overhead of coordinating these efforts and making diverse tools interoperate, consider using a suite of data integration and quality tools from a single vendor, assuming they integrate and interoperate with each other in both development and production.
- **Provide appropriate tools for data stewards and business analysts.** These people need an area within an integration suite or an individual tool that's business-friendly, so they can apply their domain expertise to identifying and prioritizing data that should be integrated and/or cleansed, as well as monitor how data exchange standards are used. Furthermore, most stewards and analysts are involved with exceptions processing, so they need an appropriate tool for reviewing, correcting, and routing exceptions; presenting exceptions in Microsoft Excel or a Web-based GUI might be preferred, since most business users are very productive with these.